Towards a Scalable and Energy-Efficient Framework for Industrial Cloud-Edge-IoT Continuum

Ons Aouedi[+*] and Kandaraj Piamrat[*]
[*]*Nantes Université, École Centrale Nantes, CNRS, INRIA, LS2N, UMR 6004, F-44000 Nantes, France*
firstname.lastname@ls2n.fr
[+]*SnT, SIGCOM, University of Luxembourg, Luxembourg*
aouedions9@gmail.com

***Abstract*—The Cloud-Edge-IoT (CEI) continuum integrates edge computing, cloud computing, and the Internet of Things (IoT), fostering rapid Industrial Internet of Things (IIoT) development. Despite its potential, it faces significant challenges, including robustness issues, communication-induced latency, and inconsistent model convergence due to system and data heterogeneity. Machine Learning (ML), a vital technology in this domain, further complicates privacy and overhead concerns. To mitigate these issues, Federated Learning (FL) appeared as a promising solution where the FL setting allows the devices to collaboratively train a model while keeping training data local. However, in practice, it suffers from several issues such as robustness (due to a single point of failure), latency (it still requires a significant amount of communication resources), and model convergence (due to the heterogeneity of system and statistics). To cope with these issues, we propose to integrate Hierarchical FL (HFL) and Spiking Neural Networks (SNN) into the framework for building a scalable and energy-efficient solution for the industrial CEI continuum. We present an in-depth overview, discussions on emerging applications, and a performance evaluation via a case study in IoT image classification. We also identify and explore open research topics crucial for the future realization of such a continuum.**

***Index Terms*—Federated Learning, Hierarchical Federated Learning, Spiking Neural Network, Internet of Things, Industrial Internet of Things**

I. Introduction

Huge advances have been made in recent years concerning both end devices and edge/cloud computing, which build up a new paradigm called the Cloud-Edge-IoT (CEI) continuum. It represents a conceptual architecture that encompasses various layers of computing and data processing. It reflects the evolving landscape of how data is generated, processed, and analyzed, with each layer serving distinct purposes and catering to specific requirements. At one end of the continuum is *cloud computing*, which represents centralized servers typically. It offers vast computational power, storage capabilities, and scalability. Moving towards the middle, we encounter *edge computing*. This layer involves processing data closer to its source, often edge servers located at the network's edge. It is essential for reducing latency, enhancing real-time processing, and improving the responsiveness of applications. At the far end, the *industrial IoT (IIoT)* is represented, where countless connected devices generate and transmit data to the edge and the cloud servers. IIoT devices include sensors, wearables, appliances, and more, which collect data from the physical world. IIoT systems rely on edge computing to preprocess and filter data locally before transmitting information to the cloud for further analysis. Therefore, the CEI continuum represents a shift from centralized to hybrid computing, emphasizing adaptability based on factors like latency and scalability. This brings up new challenges regarding the hybrid approach, balancing cloud, edge, and IIoT resources.

At the same time, the accelerating rise of Deep Learning (DL) models has resulted from an exponential increase in the amount of data generated by different IoT devices. Despite the features of the DL-based models, they are bringing out crucial concerns regarding the heterogeneity and privacy of users. The need to protect user data is accentuated by several international regulatory policies. For example, the General Data Protection Regulation (GDPR)[1] has completely redefined the data management policy. Moreover, the increase in complexity and heterogeneity of data could degrade the performance of the model, bottleneck the whole network, and cause an extra computational cost for both storage and processing. To overcome the shortcomings of centralized learning and local training, Federated Learning (FL) has been proposed as an alternative learning solution to minimize data communication [1]. FL enables the devices to learn collaboratively without the need for data sharing with a cloud or central server.

Although FL implicitly offers a certain degree of privacy, some limitations exist. One of these limitations is the problem of a heterogeneous environment including system, statistical, and communication heterogeneity. For example, the convergence of the FL-based models in the IIoT system is not always guaranteed due to the limited computational capabilities of the IIoT devices. Similarly, the data (statistical) heterogeneity in the IIoT system leads to the non-independent and identical distribution (non-IID) of data (i.e., data at each client are different in size and distribution). It can cause the divergence of the FL model. In addition, the classical FL assumes the FL server resides in the cloud for model aggregation. Using the cloud server as an FL server faces several challenges such as latency and communication costs. Even though cloud computing provides huge computing and storage capacity, it can not satisfy today's delay-sensitive applications (e.g., healthcare services, autonomous driving) since it is usually located far from the

[1]https://gdpr-info.eu/issues/data-protection-officer/

TABLE I: Comparison among different levels of aggregation algorithm

Metric	Description	Cloud-based FL	Edge-based FL	HFL	HFedSNN
Scalability	This metric means that a new node will not degrade the system performance.	Fair	Worst	Good	Best
Latency	This refers to the latency in training the global model	Best	Worst	Good	Fair
Communication overhead	This refers to the communication cost for model training	Best	Worst	Good	Fair
Robustness	This refers to the successful operation of the FL system during edge/cloud server failure.	Worst	Fair	Good	Good

Note: Worst $<$ Fair $<$ Good $<$ Best

end-users. To solve these issues, a Hierarchical FL (HFL) has been proposed to integrate several sub-aggregations of local models taking place at the edge servers [2]. However, although the performance of HFL reduces the impact of non-IID data on the model performance, its training process occurs on limited computing and low-energy IIoT devices. This makes the participation of the constrained devices in the learning process almost impossible. Moreover, communication cost is often a bottleneck. Therefore, the main contribution of this paper is a novel framework that integrates HFL and Spiking Neural Network (SNN) for the industrial CEI continuum. The SNN is a new generation of neural networks. It is an event-driven learning process and in turn, significantly reduces energy consumption. Within the CEI continuum, the IIoT devices train locally an energy-efficient SNN-based model in a private way. Then, after some local iterations, each device sends its local SNN model updates to the corresponding edge server for sub-global model aggregation. Finally, the sub-global models are aggregated at the cloud server to yield a global model. As shown in Table I, the integration of HFL and SNN in the continuum helps to achieve efficient communication overhead and enhance the robustness and flexibility of a large-scale continuum. It also reduces the latency to meet the delay-sensitive application requirements. In particular, HFL helps to reduce the burden on network bandwidth and central servers, allowing the system to scale efficiently as the number of devices grows. Additionally, the event-driven nature of SNNs ensures that energy consumption is minimized, making this integration exceptionally well-suited for IIoT devices that operate under strict power limitations. As a result, together, HFL and SNN match perfectly to the learning paradigm in the Industrial CEI continuum where energy consumption and the ability to scale are pivotal for sustainable operations.

To the best of our knowledge, no existing work explores the HFL and SNN benefits for the Industrial CEI continuum. To fill this gap, this paper presents and details the integration of HFL and SNN within this continuum. In brief, our main contributions are as follows.

- We present the benefits of using hierarchical FL and SNN for the industrial Cloud-Edge-IoT continuum, i.e., guaranteeing convergence with a heterogeneous environment (statistical and system), optimizing communication overhead, and reducing energy consumption.
- We evaluate its performance against baselines using the MNIST reference dataset for image classification tasks, which has shown promising results.
- Finally, we discuss related future research directions that need to be addressed for the full realization of the CEI continuum in future industries.

The rest of the paper is organized as follows. Section II presents how to tackle the future of the industrial CEI continuum using HFL and SNN including the key principles and the different applications in the industrial CEI continuum. Section III details the key benefits of our framework. Experimental settings and results are presented in Section IV. Finally, the conclusion and future directions are given in Section V.

II. Future Industrial CEI continuum

This section presents how to tackle the future of learning in the industrial CEI continuum.

A. Key Principles

In our framework, *HFL* acts as the structural backbone that allows Spiking neural networks to be distributed across the IIoT layer. Indeed, *SNN* is a biologically inspired neural network, in which the neurons process spike signals over time, rather than real numbers [3]. The sparsity of the synaptic spiking inputs and its event-driven nature, offer significant energy reduction compared to conventional artificial neural networks (ANNs). In particular, the energy consumed by the SNNs-based model during the learning and inference is essentially proportional to the number of spikes processed and communicated by the neurons. The spikes are emitted when the membrane potential exceeds the pre-defined threshold. The accumulated spike value will continue to increase each time the neuron fires and gradually decay toward a resting value when the neuron is not firing due to a leak factor. After the neuron fires, the membrane potential is lowered by the amount of the threshold.

As illustrated in Figure 1, the framework consists of three layers: *IIoT*, *edge*, and *cloud*; matching with its integration into

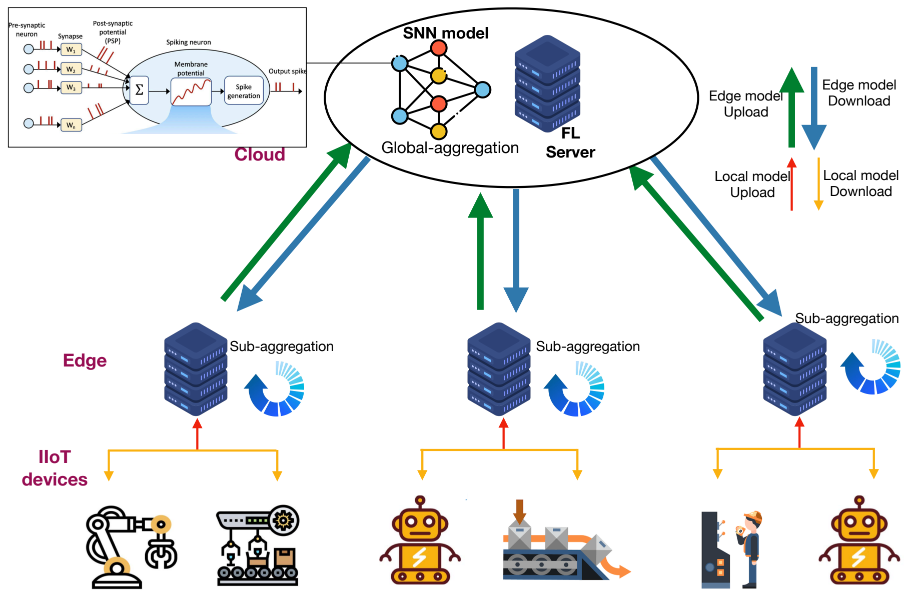


Fig. 1: Envisioned framework for industrial CEI Continuum. In this framework, the IIoT devices train the SNN model locally, the edge servers perform sub-aggregation, and the cloud performs the global aggregation.

the industrial CEI continuum. The IIoT devices represent the lowest layer of the framework. The edge servers are located in the middle layer, which is used for sub-global aggregations of IIoT device models. The cloud servers located at the top layer are used for global model aggregation. Similar to FL, HFL allows IIoT devices to train a shared global model while the raw data are kept local. The learning process in the continuum includes the following key steps:

1) *Distributed local training and updates*: Once the subset of the IIoT devices that participate in the learning process is selected, the cloud server sends an initial SNN model (similar to [4], we used VGG9 model) to them to trigger the distributed training (*Global SNN model downloading*). Then, after some local iterations, each IIoT device sends its local SNN model updates to the corresponding edge servers for sub-global model aggregation (*Local model uploading*).
2) *Sub-global model aggregations and uploads*: Upon receiving all the updates from the IIoT devices, the edge servers perform the sub-global SNN model aggregation and transfer it back (*Edge model downloading*) to their assigned IIoT devices to update their local SNN models accordingly. Then, after a specific number of iterations, the edge servers send their sub-global models (*Edge model uploading*) to the cloud server.
3) *Global model aggregation*: After receiving the sub-global models, a combined global model is created by averaging the parameters of the edge models. Finally, the global model parameters are transmitted along the hierarchy downwards to the IIoT devices.
4) *Iterated Training*: The HFL training is iterated until the desired performance is achieved. Similar to the FL, the Federated Average (FedAvg) algorithm is used for model aggregation in HFL.

B. Application in industrial CEI continuum

In this section, applications of the framework in different IIoT continuums are discussed including intelligent healthcare, intelligent manufacturing, and intelligent transportation.

- **Intelligent healthcare:** In contrast to other applications, personal healthcare data are extremely sensitive for patients and hospitals. In this context, both HFL and SNN bring their own set of advantages to intelligent healthcare systems. They can effectively address challenges such as data privacy, resource constraints, and real-time analytics that are often present in healthcare scenarios. In particular, it allows data to stay on local devices (e.g., personal wearables, and hospital equipment). This is essential for healthcare applications where patient confidentiality is crucial. Also, it is an energy-efficient framework, making it well-suited for resource-constrained devices like wearables and portable health monitors [5]. Furthermore, it can process data with very low latency, providing real-time analytics that is often necessary for medical emergencies or continuous monitoring.
- **Intelligent manufacturing:** Integration HFL and SNN can address several challenges inherent to the manufacturing domain, including real-time decision-making, data security, and resource utilization. It can be designed to adapt to anomalies or failures, through automated diagnostic and self-healing routines that minimize downtime. For example, SNN can process data with very low latency due to its event-driven architecture. This makes such integration ideal for real-time control systems in manufacturing processes where millisecond-level responses may be required. HFL is a decentralized

data processing, making it possible to perform complex analytics efficiently. It also allows new sensors to be added seamlessly. By combining the real-time capabilities of SNNs with the decentralized, scalable architecture of HFL, intelligent manufacturing systems can be elevated to new levels of efficiency, flexibility, and resilience. This fusion has the potential to significantly improve various aspects of manufacturing such as supply chain management, setting the stage for more agile, secure, and efficient manufacturing ecosystems.

- **Intelligent transportation:** Intelligent transportation systems (ITS) and autonomous driving are IoT-assisted applications for IIoT systems. SNNs are well-suited for processing time-series data and useful for capturing traffic patterns, and vehicle speeds for optimized traffic management. In particular, its event-based nature rather than clock-based processes information based on the occurrence of events (like spikes). This makes them inherently good at capturing the temporal dynamics of a system, such as varying traffic patterns during different times of the day. It uses sparse coding, which allows the network to be highly responsive to relevant features in the data stream while ignoring redundant or irrelevant information. This is valuable for analyzing real-time traffic data where only specific events may require immediate attention. In addition, the spiking nature of SNN makes it good at identifying outliers or anomalies in data. Such anomalies can be sudden spikes in traffic, uncharacteristic slowdowns, or unexpected pedestrian movement, which may be indicative of an accident that requires immediate attention. Last but not least, HFL processing and decision-making are done at the edge, closer to where the data is generated. This results in quicker decisions without the latency introduced by sending data to a central server. Using their combination for ITS can achieve unprecedented levels of efficiency, flexibility, and safety.

III. Key Benefits

This section describes how the aforementioned key principles can be integrated into a promising framework transforming the landscape of IIoT. The combination of HFL and SNN is an excellent candidate for tackling the challenges of the industrial CEI continuum, from energy constraints to data privacy and real-time analytics as well as supporting different applications with seamless connectivity. This also includes reducing the communication cost between the central server and IIoT devices and the impact of the heterogeneous environment on the model performance.

- **For communication optimization:** Given the large number of IIoT devices participating in the training process and the increasing size of the DL models, the communication cost often dominates the total cost of the system. For example, Google reports that obtaining a recurrent neural network model for next-word prediction through FL converges in 3000 rounds over 5 days. This demonstrates that traditional FL suffers from high communication overhead and latency.

  To solve this issue, the proposed framework can bring significant advantages in managing communication overhead, which is one of the critical costs in IIoT systems. First, since SNN is active only when a neuron fires spikes, only the spikes are communicated between hierarchical layers. This sparse representation significantly reduces the amount of data that needs to be transferred, thereby lowering the communication overhead. Second, in a hierarchical setup, raw data does not travel up the entire hierarchy. Instead, local computations are aggregated at various levels before reaching the central server. Hence, it requires end-to-end communications less often than with traditional FL (between cloud and end devices). This in turn improves the latency as an intermediate aggregate model can also be used by end devices. Therefore, integrating HFL and SNN minimizes the data sent over the network and reduces consequently the overheads.

- **For non-IID data:** In the industrial CEI continuum, sensors are spread geographically. Their deployments are often highly specialized, being tuned to specific tasks, locations, or conditions. This results in data that can vary significantly between sensors or over time, thus becoming non-IID. This type of data can degrade the global model performance with the conventional FL. For example, it can complicate predictive maintenance models. In particular, wear and tear on machinery could be very different depending on its usage patterns, resulting in non-IID data that is challenging to generalize across multiple machines.

  To solve these issues and enhance the learning performance, the hierarchical structure exploits the edge servers either to use their commonly shared data during the training or to use them as cluster heads. The edge servers can be used to group the IoT device models into smaller clusters based on the similarity of weight updates or other criteria (i.e., proximity). In particular, the updated local model from all clients is used to judge the similarity among the IoT devices, and the clustering algorithm (e.g., agglomerative clustering) is employed to iteratively merge the most similar devices into clusters. Such a method attempts to reduce the variance of the updated weights and in turn, preserve the uniformity in the cluster.

  Furthermore, since SNN processes feature as events in time, where each neuron in an SNN maintains a state that evolves over time, allowing it to capture temporal dynamics naturally, it can handle the temporal dynamics often associated with non-IID data. This enables SNN to be more effective in real-time scenarios like those commonly encountered in IIoT systems. Thus, when integrated into an HFL, SNN can offer good local optimization through their adaptive learning and global optimization through federated aggregation. This is particularly beneficial for dealing with spatial and temporal variations in non-IID data, which might not be achieved by using the conventional HFL-based approaches.

- **For resource constraints devices:** IIoT often comprises resource-constrained devices, such as low-power sensors and embedded systems [6]. HFL and SNN can offer a synergistic solution to meet this challenge. SNN is inherently energy-efficient because it only activates neurons (compute) when an event of interest occurs. This event-driven computation significantly lowers energy consumption. As a result, it leads to efficient use of memory, which is a crucial advantage for limited memory and battery devices in IIoT. In HFL, the device's data stays local and the devices perform the bulk of computations and only share model updates or specific parameters. This reduces the need for continuous communication to the cloud server thereby conserving energy. Additionally, HFL helps deploy the model on multiple end devices and edges respectively. In this context, HFL combined with the exit method allows samples that can be accurately classified in the early stages of the network to exit the training process and hence cut down the computing resource overhead. Thanks to the HFL, a considerable amount of samples can directly exit from end devices, and only a small part will be sent to edges [7]. This enables scalability, which is crucial for IIoT systems to adapt to changing conditions, such as the introduction of new sensors, higher data volumes, or more complex tasks. In particular, new nodes can be added to existing edge servers without the need for reconfiguring the entire network. This is particularly useful in industrial settings where new sensors or devices may need to be introduced frequently.

IV. Performance Evaluation

We present numerical results to evaluate the performance of our framework for the industrial CEI continuum. We evaluate its performance for an image classification task using the MNIST and CIFAR-10 reference datasets. Our simulation, scenario consists of a cloud FL server, edge servers, 100 devices, and 10% participants in each round. In alignment with our main benchmark [4], we experimented with the VGG9 model in the SNN version and we also used the same parameters. Then, We compare the performance of our proposed framework (HFedSNN) with the baseline (FedSNN) and also with another baseline ANN with HFL (HFedANN) in terms of accuracy, communication cost, and energy consumption.

Energy consumption can be estimated based on the number of floating point operations (FLOPs) of ANNs or SNNs, which is approximately equivalent to the number of multiply-and-accumulate operations. As illustrated in Fig. 2, the total energy estimate for our model is approximately 53.24μJ. In contrast, the VGG9-based ANN model required approximately 227.99μJ, making the SNNs model **4.3**$\times$ more energy-efficient. This significant improvement is due to the binary propagation process in SNNs, which performs accumulation operations, thereby reducing energy consumption.

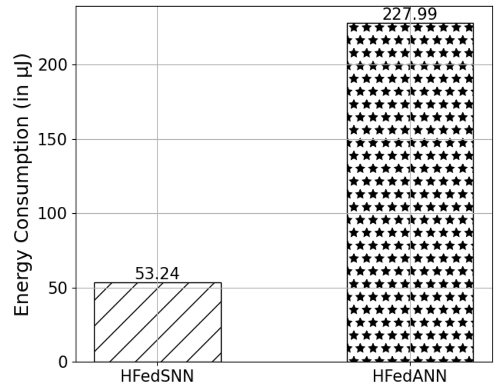
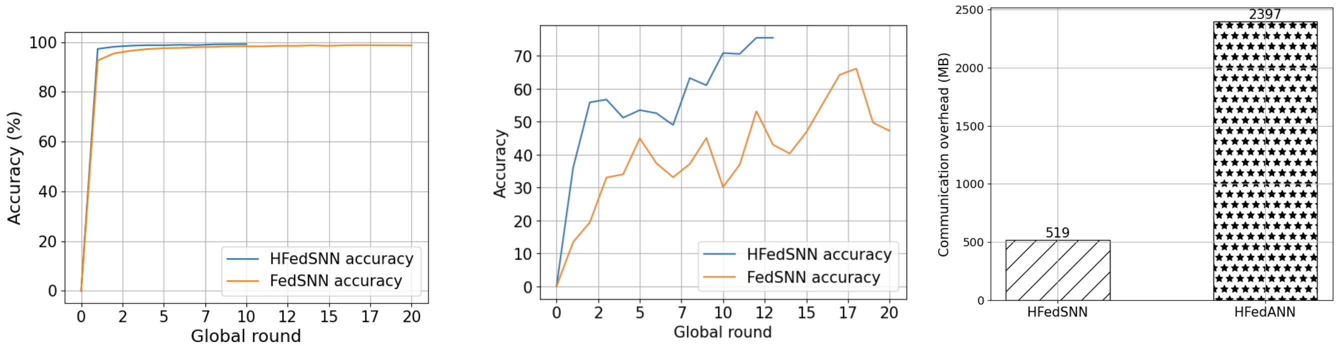


Fig. 2: Energy consumption of our model against ANN

In Fig. 3(a)(b) and Fig. 4 (a)(b), we compare the classification accuracy of HFedSNN and FedSNN under IID/non-IID settings. It can be seen that the non-IID data decreases the performance of both HFedSNN and FedSNN. However, HFedSNN exhibits superior performance when dealing with non-IID data. This is mainly due to the use of edge servers for sub-model aggregation with SNNs. These figures confirm the advantage of HFL against FL in terms of convergence speed. In particular, HFedSNN requires fewer global rounds in comparison to FedSNN to converge. This is because the intermediate layer not only mitigates the impact of non-IID data as maximum as possible but also accelerates the convergence of the global model. We then investigate the communication cost of our framework. As shown in Fig. 3(c) and Fig. 4(c), HFedSNN significantly reduces communication overhead by **4.6**$\times$ and **3.7**$\times$ compared to HFedANN with MNIST and CIFAR-10 dataset, respectively. The reason behind these results is that the integration of SNNs into HFL is a communication-effective solution. From the above results, we believe that the improvements achieved over our approach will pave the way for new solutions to enable the next generation of IIoT systems. In summary, we conclude that the classical HFL solution is no longer efficient for the CEI continuum due to its high energy consumption and communication overhead in contrast to the HFedSNN scheme, which can be environmentally sustainable. Additionally, our HFedSNN model achieves superior scalability through its hierarchical architecture, which leverages edge servers for distributed processing. This structure not only facilitates the inclusion of more devices in the training process, owing to its lower energy consumption but also significantly minimizes communication overhead. As a result, the model efficiently operates even in environments with limited bandwidth, ensuring broad and effective participation across the network.

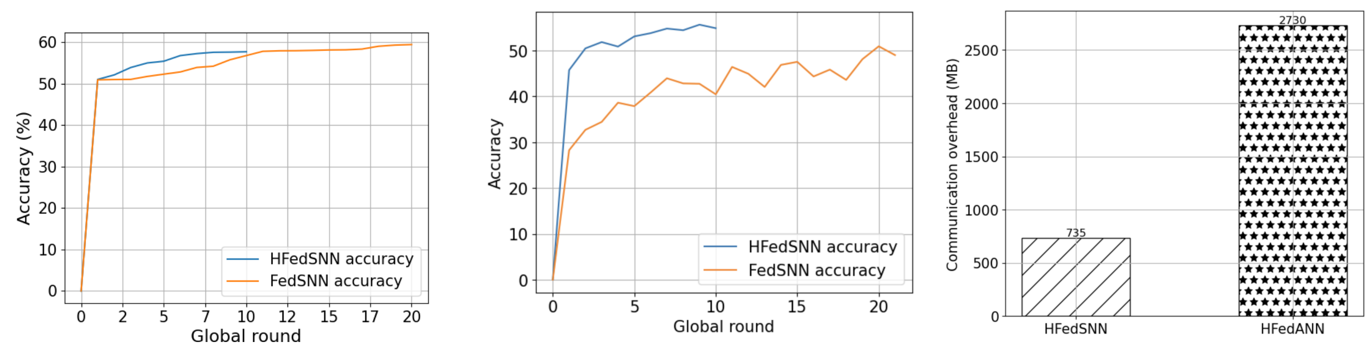
V. Conclusion and Future Directions

This article has provided a detailed overview of the integration of HFL and SNN into IIoT networks. Specifically, we



(a) Accuracy on IID data (b) Accuracy on non-IID data (c) Communication overhead

Fig. 3: Performance evaluation on MNIST dataset



(a) Accuracy on IID data (b) Accuracy on non-IID data (c) Communication overhead

Fig. 4: Performance evaluation on CIFAR-10 dataset

shed light on its benefits for the industrial CEI continuum. These range from energy consumption, and communication optimization to convergence guaranteed under the heterogeneity of devices and non-IID data, as well as improvement of the privacy risks of IIoT data. Then, the feasibility of the framework has been demonstrated via a case study and experiment. To further investigate this topic in the future, in the following, we present several research directions.

A. Robustness

Although this framework has emerged as an efficient approach to optimize the traditional FL, there still exist several vulnerabilities such as single point of failure and increased risk of data leakage. The cloud server may be vulnerable and become a single point of failure/attack that in turn could compromise the integrity and quality of services (QoS). As alternatives, *decentralized FL* or *peer-to-peer FL* can effectively improve the resilience of the training process by removing the need for a central server. However, they may be slower and less efficient than centralized FL. Under this context, a *semi-hierarchical federated edge learning* in IIoT continuum can also mitigate the reliance on the cloud-based infrastructure [8] [9]. The semi-HFL uses several edge servers to coordinate many client nodes collectively. It leverages multiple edge servers for aggregating updates from IIoT devices and for fusing learned model weights without the need for a cloud or a central server. Existing simulation results demonstrate that semi-HFL outperforms peer-to-peer FL and classical FL in terms of accuracy and convergence speed, respectively [10]. On the other hand, the HFL process is potentially vulnerable to numerous forms of attacks, including but not limited to, data and model poisoning. Addressing the challenge of preventing poisoning attacks in HFL requires a multi-faceted approach, especially considering the potential vulnerability of intermediate layers at the edge. Conducting regular security audits and updating the system to patch known vulnerabilities without introducing prohibitive latency is crucial. Keeping up with the latest in cybersecurity and adapting the system accordingly can provide a strong defense against emerging threats.

B. Security and Privacy

Although our framework can provide privacy protection, it still creates security and privacy vulnerabilities for the end users. Specifically, during training, the communication between the clients and edge server may expose the model parameters and in turn be a target for several security threats such as membership inference attacks. In this context, many privacy mechanisms have been proposed to ensure privacy during parameter exchange such as *cryptographic and differential privacy methods*. However, such mechanisms may be expensive in terms of computation and their integration with SNN remains relatively unexplored. This presents an important area of exploration, as it necessitates adapting and evaluating these mechanisms within the unique context of SNN.

On the other hand, *meta-model* can be used to further enhance privacy [11]. It can be easier to apply with SNN. In other words, the federated meta-model can be trained on the meta-data instead of user-sensitive data and thus reduce the performance of reverse engineering attacks. Moreover, using *reputation management* (reward and punishment) based approaches can improve the security and privacy of the devices in the IIoT systems [12].

C. System Heterogeneity of IoT devices

While our framework can offer several advantages in terms of model training, real-time decision-making, and efficiency; it is important to note that it might struggle with heterogeneity among IIoT devices. Devices in a heterogeneous IIoT system may operate on different schedules or under different constraints, making it hard to coordinate learning and model updates effectively. Heterogeneity in this context can refer to different aspects like computational power, storage capacity, connectivity, or even the type and granularity of data that each device can collect and process. To address this challenge, performing local computations and communicating updates in an asynchronous manner [13], accommodating devices that operate under different conditions can be a promising solution. Moreover, performing computations locally on IIoT devices and only transmitting the essential updates to the aggregator helps in managing the different computational capabilities of the devices. This local processing approach reduces the need for continuous high-bandwidth connectivity and accommodates devices with limited computational power. Furthermore, SNN-based models can be designed to adapt to the capabilities of each device, enabling efficient distributed learning. Thus, more attention should be given to this direction and more studies should be conducted on its performance with heterogeneous model architectures [14].

D. Mobility

The IIoT network is characterized by the mobility of IIoT devices (e.g., drones or robots), it is hence challenging to ensure continuous communications between them and their associated edge servers. Moreover, its learning performance deteriorates with highly mobile users [15]. For example, the robots may move among multiple edge servers during the local training procedures, leading to incomplete training. Also, synchronizing the system's global clock becomes increasingly difficult, affecting the timing and sequence of FL rounds and model updates. Thus, the mobility of devices can cause frequent service interruptions, leading to challenges in collecting model parameters and executing FL tasks effectively. This can degrade the overall quality of service and the reliability of the learning process. Additionally, high mobility increases the complexity of maintaining synchronized operations, affecting the timing and sequencing of FL tasks. In this context, several solutions can be proposed such as Reinforcement Learning (RL). It can be used to build an *intelligent device selection*. In particular, it can allow the edge servers to select the devices with no or less mobility instead of other high-mobility devices. Furthermore, to reduce the impact of IIoT devices' mobility in HFL systems, DL-enabled *mobility prediction* can be another interesting direction. The mobility prediction helps the edge server prevent selecting or eliminating the participation of IIoT devices that might leave its coverage area in the near future. With such solutions, many mobility challenges can be mitigated to release the full potential of this framework in mobile IIoT environments.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[2] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[3] J. L. Lobo, J. Del Ser, A. Bifet, and N. Kasabov, "Spiking neural networks and online learning: An overview and perspectives," *Neural Networks*, vol. 121, pp. 88–100, 2020.

[4] Y. Venkatesha, Y. Kim, L. Tassiulas, and P. Panda, "Federated learning with spiking neural networks," *IEEE Transactions on Signal Processing*, vol. 69, pp. 6183–6194, 2021.

[5] N. Skatchkovsky, H. Jang, and O. Simeone, "Federated neuromorphic learning of spiking neural networks for low-power edge intelligence," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8524–8528.

[6] R. A. Khalil, N. Saeed, M. Masood, Y. M. Fard, M.-S. Alouini, and T. Y. Al-Naffouri, "Deep learning in the industrial internet of things: Potentials, challenges, and emerging applications," *IEEE Internet of Things Journal*, vol. 8, no. 14, pp. 11 016–11 040, 2021.

[7] Z. Zhong, W. Bao, J. Wang, X. Zhu, and X. Zhang, "Flee: A hierarchical federated learning framework for distributed deep neural network over cloud, edge and end device," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2022.

[8] Y. Sun, J. Shao, Y. Mao, J. H. Wang, and J. Zhang, "Semi-decentralized federated edge learning for fast convergence on non-iid data," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2022, pp. 1898–1903.

[9] L. Zhao, M. Valero, S. Pouriyeh, L. Li, and Q. Z. Sheng, "Communication-efficient semi-hierarchical federated analytics in iot networks," *IEEE Internet of Things Journal*, 2021.

[10] Y. Sun, J. Shao, Y. Mao, J. H. Wang, and J. Zhang, "Semi-decentralized federated edge learning with data and device heterogeneity," *IEEE Transactions on Network and Service Management*, 2023.

[11] N. A. A.-A. Al-Marri, B. S. Ciftler, and M. M. Abdallah, "Federated mimic learning for privacy preserving intrusion detection," in *2020 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*. IEEE, 2020, pp. 1–6.

[12] P. K. R. Maddikunta, Q.-V. Pham, D. C. Nguyen, T. Huynh-The, O. Aouedi, G. Yenduri, S. Bhattacharya, and T. R. Gadekallu, "Incentive techniques for the internet of things: a survey," *Journal of Network and Computer Applications*, vol. 206, p. 103464, 2022.

[13] Y. Sun, J. Shao, Y. Mao, and J. Zhang, "Asynchronous semi-decentralized federated edge learning for heterogeneous clients," in *ICC 2022-IEEE International Conference on Communications*. IEEE, 2022, pp. 5196–5201.

[14] J. Jang, H. Ha, D. Jung, and S. Yoon, "Fedclassavg: Local representation learning for personalized federated learning on heterogeneous neural networks," in *Proceedings of the 51st International Conference on Parallel Processing*, 2022, pp. 1–10.

[15] C. Feng, H. H. Yang, D. Hu, T. Q. Quek, Z. Zhao, and G. Min, "Federated learning with user mobility in hierarchical wireless networks," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 01–06.