Enhancing Communication-Efficient Federated Learning for Human Activity Recognition Through Knowledge Distillation

1st Hafiz Muhammad Bilal
Department of Information Engineering and Computer Science
University of Trento, Italy
hafiz.liaqat@studenti.unitn.it

2nd Mir Hassan
Department of Information Engineering and Computer Science
University of Trento, Italy
mir.hassan@unitn.it

3rd Giovanni Iacca
Department of Information Engineering and Computer Science
University of Trento, Italy
giovanni.iacca@unitn.it

***Abstract*—Human Activity Recognition (HAR) has profound applications in domains such as healthcare, wearable devices, and smart environments, where continuous monitoring is essential. However, traditional centralized learning approaches raise privacy concerns and are infeasible for resource-constrained devices due to high communication costs. Federated Learning (FL) offers a privacy-preserving solution by training models locally and aggregating updates, but it still encounters considerable communication overhead. This paper proposes an optimized FL framework that integrates model compression through Knowledge Distillation (KD) to reduce communication costs while preserving model performance. Using a Teacher-Student model architecture, the framework enables the deployment of highly compressed Student models without compromising accuracy or other performance metrics. Experimental evaluations on a smartphone-based HAR dataset show that the proposed framework achieves comparable accuracy, precision, and recall w.r.t. the original (uncompressed) models, with up to 75% reduction in model size (and, consequently, communication cost). This approach demonstrates scalability and feasibility for real-world HAR applications on resource-constrained devices, also laying the groundwork for efficient, privacy-preserving distributed learning in heterogeneous computational environments.**

***Index Terms*—Human Activity Recognition, Federated Learning, Distillation, Compression.**

I. Introduction

Human Activity Recognition (HAR) has gained significant attention due to its applications in healthcare, wearable technology, smart environments, and surveillance. By leveraging sensor data from smartphones, smartwatches, and IoT devices, HAR models infer activities such as walking, running, and sitting, enabling real-time monitoring for personalized health and fitness applications [1]. Traditional HAR approaches rely on centralized learning, where raw sensor data are transmitted to a central server for model training. While effective, this paradigm raises critical concerns regarding privacy, scalability, and resource constraints. Continuous data transmission increases privacy risks and imposes substantial computational and storage demands, making large-scale deployments impractical [2]. Furthermore, resource-constrained devices often face bandwidth and energy limitations, underscoring the need for distributed learning methodologies that preserve privacy and ensure computational efficiency.

Federated Learning (FL) addresses these challenges by enabling devices to collaboratively train models without sharing raw data [3]. While FL enhances data security, it introduces significant communication overhead, especially for deep learning models with large parameter counts [4]. Techniques such as gradient sparsification and quantization have been explored to reduce update sizes, but often at the cost of degraded model accuracy [5]. Client selection and asynchronous updates improve communication efficiency but can introduce bias in non-IID settings common in HAR [6]. Hierarchical FL further reduces communication through intermediate aggregation but suffers from synchronization complexities and increased latency [7]. To address communication bottlenecks, FL combined with Knowledge Distillation (FL-KD) has emerged as a promising alternative. Instead of transmitting full model updates, clients exchange distilled knowledge, significantly reducing communication load. However, most existing FL-KD methods primarily focus on handling model heterogeneity rather than optimizing communication efficiency. For example, FedAKD [8] facilitates heterogeneous model training but does not explicitly target communication reduction. Similarly, contrastive KD methods such as CapMatch [9] improve HAR accuracy but operate under centralized settings, limiting their use in privacy-preserving FL. Other approaches like FedBKD [10] integrate bidirectional KD but require auxiliary datasets, which may not be available in real-world HAR scenarios. Lightweight KD-based HAR models, such as LHAR [11], primarily focus on model complexity rather than communication overhead in federated environments. Despite these efforts, achieving an optimal trade-off between communication efficiency, privacy preservation, and model performance for HAR under resource constraints remains an open challenge.

In this work, we propose a novel FL-KD framework specifically tailored for HAR. Our approach introduces a Teacher-Student compression mechanism within FL to minimize communication overhead while maintaining high model performance. Clients train lightweight Student models instead of transmitting large updates, significantly reducing communication costs. Unlike prior FL-KD methods that emphasize model heterogeneity or complexity reduction, our framework directly targets communication efficiency, enabling practical deployment of FL for resource-constrained edge devices.

The key contributions of this work are as follows:

- We proposed a Teacher-Student FL-KD framework that integrates FL and KD to minimize communication costs while preserving model accuracy. Unlike previous works that treat FL and KD separately, our approach compresses model updates before transmission, significantly reducing bandwidth usage.
- We demonstrate that our approach reduces communication costs by up to 75% compared to traditional FL methods, without compromising performance.
- We benchmark a 1D Convolutional Neural Network (1D-CNN), a Long Short-Term Memory (LSTM) network, and a Multi-Layer Perceptron (MLP) within FL-KD to determine the most efficient HAR model for FL settings.
- We conduct a comprehensive evaluation of FL with and without KD, validating its impact on reducing training and inference times as well as the model size and resulting communication overhead.

The rest of the paper is organized as follows. In Section II, we provide a brief overview of state-of-the-art methods. Our proposed method is introduced in Section III. Section IV describes the experimental setup, datasets, and baseline methods used for our framework. The results and discussion are presented in Section V. Finally, Section VI draws the conclusions.

II. Related Work

HAR systems often rely on wearable device data, necessitating methods that minimize communication costs without compromising model performance. Several studies have addressed high communication overheads in FL applied to resource-constrained applications like HAR. The foundational FedAvg approach [3] reduces communication rounds by allowing multiple local updates before aggregation. Client selection strategies [6] further optimize communication by involving only a subset of devices based on network or data quality. However, selective participation can introduce bias, particularly under non-IID conditions common in HAR. Sparse updates and gradient quantization techniques [5] transmit only significant model changes, reducing communication loads but often degrading accuracy, especially in tasks requiring precise activity classification. Moreover, sparse updates struggle to converge when data are unevenly distributed.

Local staleness approaches [12] enable asynchronous updates to reduce communication frequency but risk poor convergence under highly dynamic data distributions. Hierarchical FL [7] aggregates updates through intermediary gateways to reduce device-server communication but shifts the burden to gateway nodes, which must handle high-dimensional sensor data. Asynchronous FL [13] improves efficiency by removing synchronization requirements, but can suffer from inconsistencies due to unrelated datasets across devices. Allowing multiple local epochs before communication [14] works well when activities are predictable, but can impair model accuracy in dynamic settings. Decentralized FL [15] eliminates the need for a central server, relying on peer-to-peer communication, but increases overall communication overhead and complicates model consistency. Despite these innovations, the fundamental challenge remains: the volume of model updates. Techniques like FedAvg and sparse updates reduce the frequency or size of transmissions but do not sufficiently address the problem of large model payloads, especially for deep learning-based HAR systems requiring high-capacity models. Model compression techniques, particularly Knowledge Distillation (KD), offer a promising solution. By transferring knowledge from larger Teacher models to smaller Student models, KD significantly reduces communication costs without sacrificing performance. However, KD-based compression remains underexplored in the context of communication-efficient FL for HAR. Addressing this gap forms the core focus of this work.

III. Methodology

A. Proposed Framework

In this section, we describe the proposed approach to enhance FL using model compression (based on KD) in order to achieve effective HAR while minimizing communication overhead as well as maintaining high accuracy, scalability, and privacy. Specifically, we integrate FL with KD to address the challenges of communication and resource constraints in edge environments. Our framework consists of the following components:

- **Edge Devices:** Perform local training on private datasets without sharing raw data. This preserves privacy and ensures compliance with data sovereignty requirements.
- **Central Server:** Aggregates model updates sent by edge devices and applies the FedAvg algorithm to construct a global model.
- **Knowledge Distillation:** Compresses the model by employing a Teacher-Student paradigm. The global model (Teacher) guides smaller, resource-efficient models (Students) deployed on edge devices, significantly reducing the communication overhead.

The workflow of the proposed framework is detailed as shown in Figure 1. Leveraging this framework, we compare the performance of the baseline, uncompressed FL (Algorithm 1) vs. that of the proposed FL-KD scheme (Algorithm 2).

B. Knowledge Distillation

KD is a model compression technique that facilitates the transfer of knowledge from a large, complex model, referred to as the Teacher, to a smaller, resource-efficient model known as the Student [16]. In this framework, the global model trained on aggregated data serves as the Teacher, capturing rich and nuanced data representations. The Student model, which is computationally lightweight, is trained using the soft predictions (logits) generated by the Teacher model. This

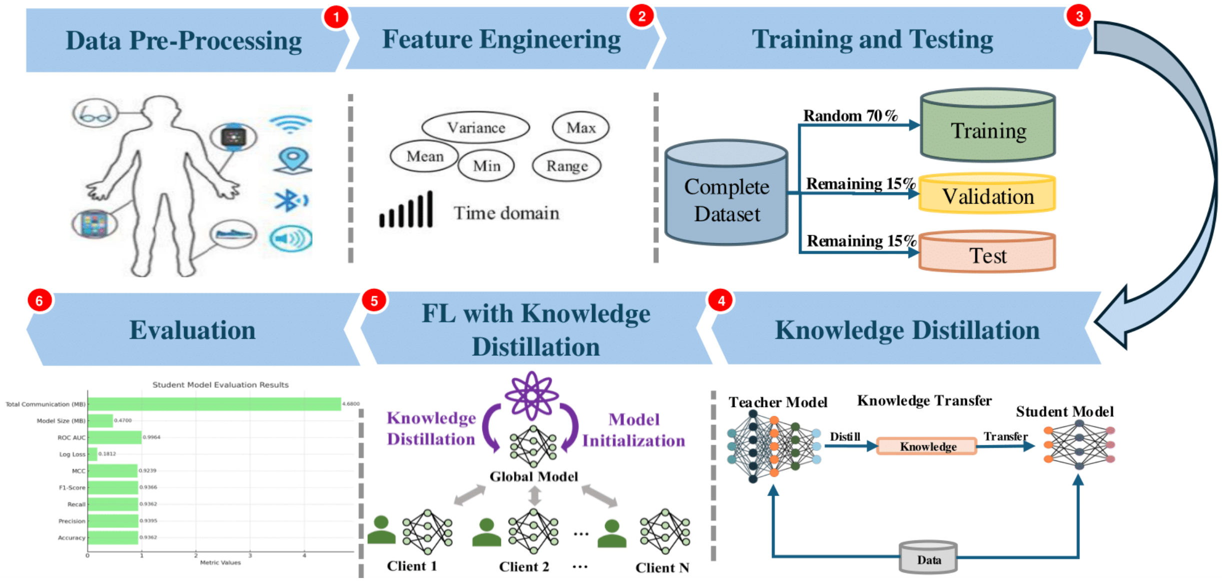

Fig. 1: Workflow of the proposed FL-KD approach. ① Data Pre-Processing: Clean and normalize HAR data, extracting key activity features. ② Feature Engineering: Select relevant time and frequency domain features to enhance model performance. ③ Training and Testing: Train Deep Learning models with a structured train-test split. ④ Knowledge Distillation: Apply model compression. ⑤ FL with Knowledge Distillation: Integrate KD to optimize communication and performance. ⑥ Evaluation: Balance communication cost and accuracy for an optimized system.

Algorithm 1 Federated Learning Without Compression (Baseline)

```
Initialize global model G on the central server.
for each communication round t do
    Broadcast G to all participating edge devices.
    for each device i in parallel do
        Perform local training on the private dataset
        to update Δ_i.
        Send (uncompressed) updates Δ_i to the server.
    end for
    Aggregate updates Δ_i from all devices using FedAvg
    to update G.
end for
Return the final global model G.
```

approach enables the Student model to mimic the behavior of the Teacher while requiring significantly fewer resources.

IV. Experimental Setup

To evaluate the effectiveness of the proposed FL-KD framework for HAR, extensive experiments were conducted on the UCI Machine Learning Repository HAR dataset [17]. The framework was implemented using PyTorch within the Flower FL simulation environment [1], utilizing Python 3.11.10 [2] and TensorFlow. The experiments simulated a heterogeneous, non-IID environment with 10 clients over 5 communication rounds. Three different Deep Learning architectures, namely

[1]https://flower.ai

[2]https://docs.python.org/release/3.11.10/

Algorithm 2 Federated Learning with Knowledge Distillation (Proposed)

```
Initialize global model G on the central server.
for each communication round t do
    Broadcast G to all participating edge devices.
    for each device i in parallel do
        Perform local training on the private dataset
        to update Δ_i.
        Apply Knowledge Distillation to compress Δ_i.
        Send (compressed) updates Δ_i to the server.
    end for
    Aggregate updates Δ_i from all devices using FedAvg
    to update G.
end for
Return the final global model G.
```

a 1D-CNN, an LSTM network, and an MLP were used. For each architecture, a Teacher model was trained centrally, and knowledge was distilled into lightweight Student models on client devices to optimize communication efficiency. The 1D-CNN Teacher model included two Conv1D layers (64 and 128 filters, kernel size 5, `tanh` activation), followed by MaxPooling1D, a dense layer (128 units), dropout 0.5, and a softmax output layer. The Student model was a simplified version with reduced filters (32 and 64) and a smaller dense layer (64 units). The LSTM Teacher model contained 100 units with L2 regularization ($\lambda = 0.001$) and dropout 0.3, while the Student model used 50 units with the same regularization. The

MLP Teacher model comprised three dense layers (512 units, `ReLU` activation) with dropout 0.5, while the Student model had a single dense layer (64 units) with the same activation function. For all models, the learning rate was 0.00005.

A. Dataset and Preprocessing

The selected dataset [17] is a benchmark for HAR research. It comprises time-series data from smartphone accelerometers and gyroscopes for six activities: `Walking`, `Walking Upstairs`, `Walking Downstairs`, `Sitting`, `Standing`, and `Lying`. In the preprocessing phase, the data were segmented into fixed-length windows with overlapping segments to capture activity transitions. Following segmentation, raw signals were transformed into 561 features using time-domain and frequency-domain methods to facilitate model learning. The dataset was divided as 70% for training, 15% for validation, and 15% for testing.

B. Baseline Models

For baseline, the three models mentioned above (1D-CNN, LSTM network, and MLP) were trained and evaluated also in a centralized environment, to establish benchmarks for accuracy and efficiency.

C. Evaluation Metrics

The proposed FL-KD framework was evaluated using three key metrics to ensure a comprehensive performance analysis. First, *Performance* was measured in terms of the models' classification ability across six human activities, using evaluation metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Accuracy, Precision, Recall, and F1 Score. Second, *Communication Efficiency* was assessed by quantifying the amount of data transmitted during model updates in the FL process, which is directly proportional to the number of model parameters. Third, *Training and Inference Times* were evaluated to reflect the framework's suitability for real-time applications, capturing the time required for both model training and prediction.

V. Results

This section presents the experimental results comparing the performance of Teacher and Student models under both federated and non-federated settings, with and without model compression via KD. The experiments were designed to validate the effectiveness of the proposed approach in reducing communication costs while maintaining high model accuracy for HAR tasks.

A. Performance

We evaluated three models (1D-CNN, LSTM, and MLP) as Teacher models alongside their compressed Student counterparts. Table I summarizes performance metrics including Accuracy, Precision, Recall, F1 Score, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) across federated and non-federated environments.

1) Non-Federated Setting: In the non-federated setting, the 1D-CNN Student model achieved 95.15% accuracy, closely matching the Teacher's 95.22%. It also slightly outperformed the Teacher in RMSE (0.2794 vs. 0.2818) while maintaining identical MAE (0.0563), demonstrating effective KD. The LSTM Student model showed a minor performance drop, with 93.25% accuracy compared to the Teacher's 94.88%, and a higher MAPE (6.00% vs. 4.13%), indicating some degradation in sequential feature learning. For the MLP model, the Student slightly outperformed the Teacher, achieving higher accuracy (91.75% vs. 91.48%) and improved error metrics (lower MAE and RMSE).

2) Federated Learning Setting: In the federated setting, the 1D-CNN Student model surpassed its Teacher, achieving 96.00% accuracy compared to 95.50%. It also exhibited improved MAE (0.0460 vs. 0.0640) and RMSE (0.2450 vs. 0.2960), confirming robustness in distributed environments. For the LSTM, the Student model achieved 93.80% accuracy, slightly lower than the Teacher's 95.20%, along with a noticeable decline in recall and precision, reflecting challenges in preserving temporal dependencies under FL. The MLP Student model improved upon the Teacher, achieving 92.00% accuracy compared to 91.40%, with better MAE (0.0760 vs. 0.0800), demonstrating efficient compression with minimal performance loss.

3) Analysis of Confusion Matrices: The confusion matrices for the 1D-CNN model (Figure 2) show high classification accuracy for activities such as *Laying* (L), *Walking* (W), and *Standing* (St), with minimal confusion. Minor errors between *Walking Downstairs* (WD) and *Walking Upstairs* (WU) appeared in the federated Student model but did not significantly affect overall performance. FL enhanced model robustness, with the Student outperforming the Teacher. For the LSTM model (Figure 3), recurrent misclassifications occurred between *Sitting* (S) and *Standing* (St), likely due to overlapping temporal features. Although non-federated Student models showed higher misclassification rates, FL reduced these errors and improved sequence learning. In the MLP model (Figure 4), challenges persisted in distinguishing WU and WD activities due to limited temporal feature extraction. While non-federated models showed consistent misclassifications, FL modestly improved accuracy and robustness by exposing the models to decentralized, diverse data.

Overall, these results affirm that Student models, especially when integrated with FL, maintain competitive accuracy while significantly reducing resource demands. Confusion matrix analysis highlights FL's positive impact on model generalization and error reduction across all architectures.

B. Communication Efficiency

A key objective of our framework is to enhance communication efficiency in FL through model compression. As illustrated in Figure 5 for the 1D-CNN model (selected based on the best results in Table I), the Student model significantly reduces communication costs compared to its Teacher counterpart. Specifically, the Student model size decreased

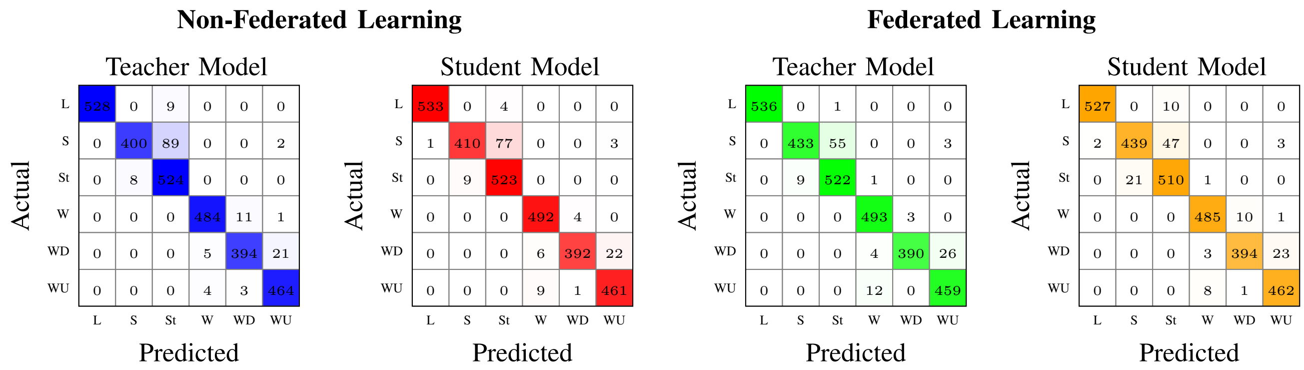

Fig. 2: Performance comparison of 1D-CNN Model. Activity classes: W - Walking, U - Upstairs, D - Downstairs, S - Sitting, St - Standing, L - Laying.

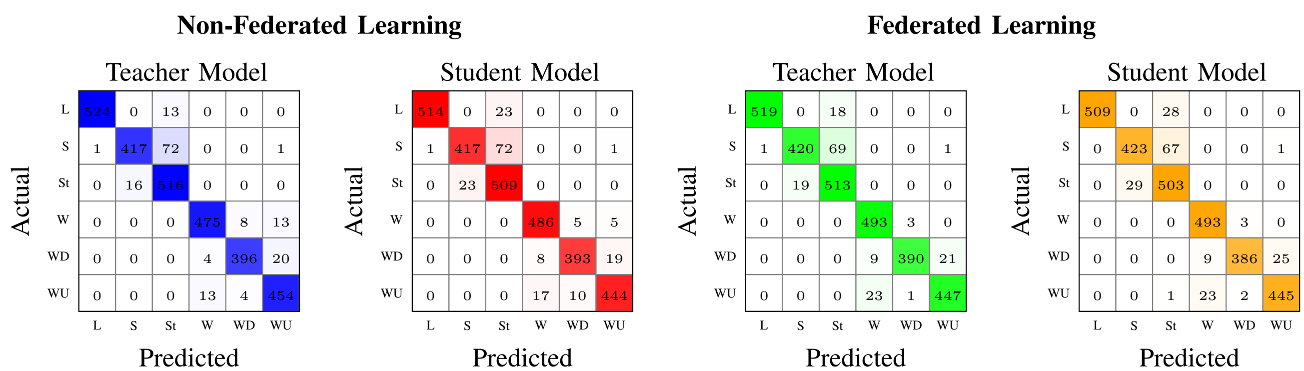

Fig. 3: Performance comparison of LSTM models. Activity classes: W - Walking, U - Upstairs, D - Downstairs, S - Sitting, St - Standing, L - Laying.

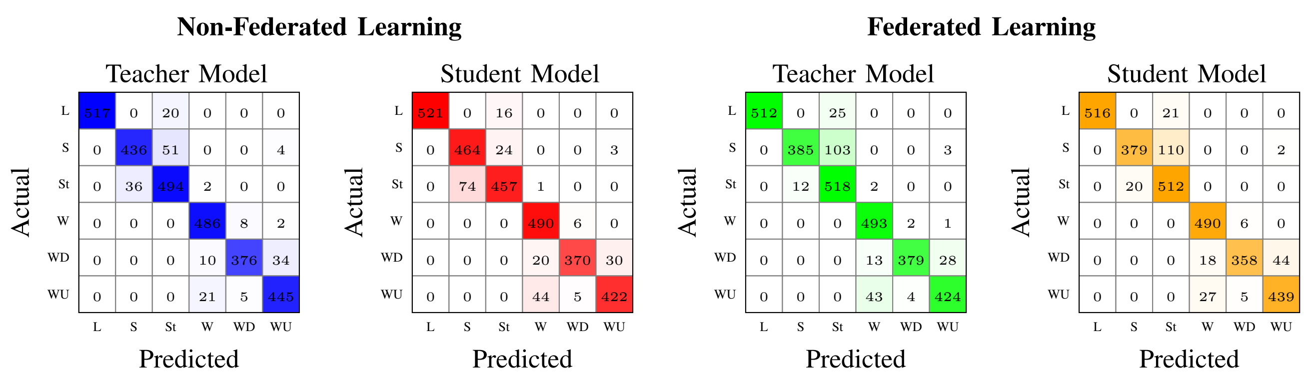

Fig. 4: Performance comparison of MLP models. Activity classes: W - Walking, U - Upstairs, D - Downstairs, S - Sitting, St - Standing, L - Laying.

from 26.21 MB to 6.58 MB, achieving a 75% reduction. This substantial decrease in model size directly lowers the communication overhead per round, making the framework more suitable for bandwidth-constrained environments such as wearable HAR systems. By minimizing transmitted data, our approach ensures efficient resource utilization and scalability in large-scale federated networks.

C. Training and Inference Times

Training and inference times are critical for real-time HAR applications, where rapid and accurate feedback is essential. Experimental results in Figure 5 show that Student models outperform Teacher models in both metrics. For the 1D-CNN configuration, inference time was reduced from 0.26 seconds (Teacher) to 0.15 seconds (Student), an improvement of approximately 42%. Training time decreased from 359.93

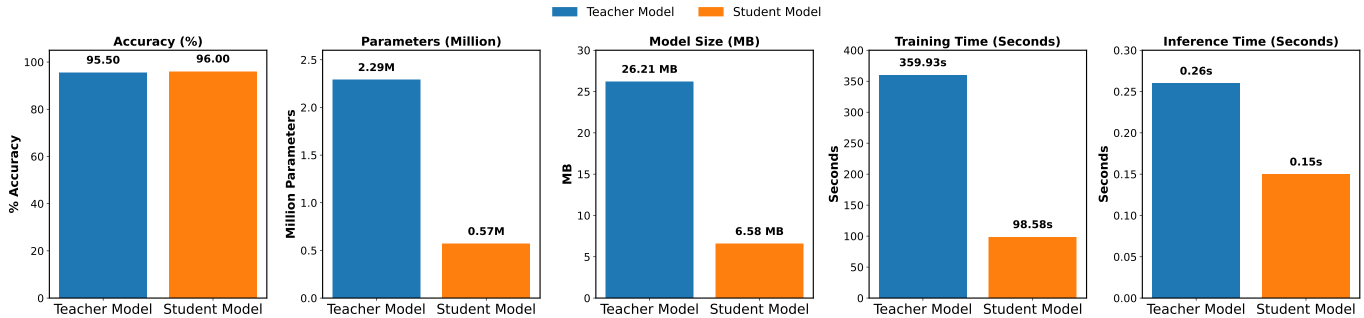

Fig. 5: Comparison of 1D-CNN Teacher and Student models for the different metrics of interest.

TABLE I: Performance comparison of 1D-CNN, LSTM, and MLP models with and without Federated Learning. Red values denote the best results across all models.

Model	Configuration	MAE	RMSE	MAPE	Accuracy %	Precision	Recall	F1 Score
1D-CNN	Non-federated (Teacher)	0.0563 ± 0.0001	0.2818 ± 0.0123	3.5539 ± 0.1242	95.22 ± 0.0017	0.9551 ± 0.0012	0.9522 ± 0.0018	0.9519 ± 0.0013
	Non-federated (Student)	0.0563 ± 0.0002	0.2794 ± 0.0104	3.7694 ± 0.1303	95.15 ± 0.0013	0.9543 ± 0.0011	0.9515 ± 0.0015	0.9512 ± 0.0011
	Federated (Teacher)	0.0640 ± 0.0041	0.2960 ± 0.0123	3.8420 ± 0.1529	95.50 ± 0.0023	0.9600 ± 0.0012	0.9580 ± 0.0021	0.9580 ± 0.0017
	Federated (Student)	0.0460 ± 0.0039	0.2450 ± 0.0145	2.6030 ± 0.1637	96.00 ± 0.0019	0.9600 ± 0.0015	0.9600 ± 0.0018	0.9600 ± 0.0016
LSTM	Non-federated (Teacher)	0.0631 ± 0.0032	0.2982 ± 0.0105	4.1341 ± 0.3210	94.88 ± 0.0021	0.9512 ± 0.0019	0.9488 ± 0.0023	0.9487 ± 0.0018
	Non-federated (Student)	0.0852 ± 0.0051	0.3529 ± 0.0121	6.0027 ± 0.5276	93.25 ± 0.0027	0.9368 ± 0.0024	0.9325 ± 0.0029	0.9328 ± 0.0021
	Federated (Teacher)	0.0860 ± 0.0067	0.3510 ± 0.0164	6.1890 ± 0.2471	95.20 ± 0.0027	0.9530 ± 0.0020	0.9520 ± 0.0023	0.9520 ± 0.0019
	Federated (Student)	0.0910 ± 0.0043	0.3720 ± 0.0181	6.5540 ± 0.2314	93.80 ± 0.0028	0.9380 ± 0.0023	0.9380 ± 0.0029	0.9380 ± 0.0024
MLP	Non-federated (Teacher)	0.1055 ± 0.0027	0.3903 ± 0.0154	8.1218 ± 0.4139	91.48 ± 0.0031	0.9287 ± 0.0029	0.9148 ± 0.0030	0.9143 ± 0.0028
	Non-federated (Student)	0.0916 ± 0.0034	0.3407 ± 0.0126	6.9772 ± 0.3055	91.75 ± 0.0026	0.9299 ± 0.0025	0.9175 ± 0.0027	0.9160 ± 0.0024
	Federated (Teacher)	0.0800 ± 0.0055	0.4410 ± 0.0203	10.759 ± 0.2819	91.40 ± 0.0032	0.9370 ± 0.0030	0.9140 ± 0.0032	0.9140 ± 0.0031
	Federated (Student)	0.0760 ± 0.0048	0.4140 ± 0.0195	9.4780 ± 0.2715	92.00 ± 0.0029	0.9370 ± 0.0027	0.9370 ± 0.0030	0.9370 ± 0.0028

seconds to 98.58 seconds, enabling faster model updates and more frequent FL rounds within the same communication constraints. These reductions in training and inference times enhance the feasibility of deploying our framework in dynamic, real-time environments, ensuring both efficiency and responsiveness.

D. Discussion

This study evaluated the performance of three Deep Learning models 1D-CNN, LSTM, and MLP for Human Activity Recognition (HAR), emphasizing their deployment in resource-constrained environments via Federated Learning (FL) and Knowledge Distillation (KD). The comparative analysis revealed nuanced trade-offs among models, offering key insights for practical applications. The 1D-CNN achieved an optimal balance between accuracy and efficiency, reaching 96% accuracy with up to 40% reduction in inference time compared to baseline models on the HAR dataset. Its ability to efficiently capture local temporal features with minimal computational overhead makes it particularly suitable for real-time wearable applications. In contrast, the LSTM excelled in modeling long-term dependencies but incurred 30% higher communication overhead relative to 1D-CNN, limiting its feasibility on edge devices. The MLP, while achieving a lower accuracy of 90.2%, remains attractive for ultra-low-power deployments due to its simplicity and low resource demands. KD played a critical role in improving model efficiency. Student models distilled from Teacher models achieved over 70% reduction in communication overhead without sacrificing accuracy. Notably, the distilled 1D-CNN Student model achieved 96% accuracy, slightly outperforming its Teacher model (95.5%), confirming KD's regularization effect. These results surpass FedAKD [18], which achieved 60% overhead reduction but experienced minor accuracy drops. Similarly, distilled LSTM and MLP models maintained performance comparable to their Teachers while benefiting from reduced communication costs and faster inference. FL further amplified these benefits by enabling privacy-preserving, distributed model training while minimizing communication loads. Our approach aligns with recent studies [8], [19], highlighting the synergy of FL and KD for efficient HAR. Unlike FedAKD [18], which relied on auxiliary datasets, our method achieved over 70% communication reduction without auxiliary data and maintained higher model accuracy. Compared to Fed2KD [20] and FedMEKT [21], our approach offers competitive accuracy while demonstrating superior robustness in non-IID settings and further reducing communication overhead. Despite these advancements, limitations persist. The LSTM's high computational cost restricts its deployment on constrained devices.

Future work could apply model pruning [22] or dynamic resource allocation to optimize LSTM efficiency. Although 1D-CNNs are computationally efficient, they struggle with capturing long-term dependencies; hybrid architectures combining CNNs with lightweight recurrent units, such as Gated Recurrent Units (GRUs), may address this. MLPs, while lightweight, exhibit limited feature extraction capabilities; integrating attention mechanisms could enhance their performance without significantly increasing complexity. Domain-specific constraints, such as high accuracy requirements in healthcare or ultra-low latency needs in smart home applications, may affect the generalizability of the proposed models. Future work could explore domain adaptation and transfer learning techniques to enhance model portability across application domains. Overall, KD emerges as a particularly effective strategy for reducing model size while maintaining accuracy in energy-constrained IoT devices. Future research may investigate combining KD with quantization, pruning, and weight sharing [9], [23] to further optimize models. Additionally, adapting the FL-KD framework to support asynchronous updates and personalized models [24], [25] could enhance scalability and performance, particularly in highly heterogeneous environments such as personalized healthcare monitoring. Extending the proposed methods to other data types, including imaging, could broaden their applicability where communication and resource constraints are critical challenges.

VI. Conclusion

This study explored the use of FL combined with KD to address one of the key challenges in HAR applications, i.e., maximizing performance while reducing communication costs. FL enables distributed learning across devices, mitigating privacy concerns by keeping data localized, but its implementation is hindered by high communication costs. By integrating KD, i.e., a model compression technique that trains lightweight Student models from a more complex Teacher model, we were able to reduce communication overhead, making FL feasible for resource-constrained systems such as smartphones and wearable devices. Our experimentation demonstrated that Student models significantly reduced communication costs and model sizes by over 70% while maintaining comparable, and occasionally superior, performance. For instance, the compressed 1D-CNN Student model achieved 96% accuracy compared to the Teacher model's 95.50%. These findings highlight the practicality of deploying HAR systems in real-world scenarios with resource limitations and validate KD as a reliable method for enhancing FL's scalability and efficiency.

Acknowledgment

Funded by the European Union (project no. 101071179). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EISMEA. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] J. Wang, Y. Chen, and C. Hu, "Activity Recognition," in *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*. Springer, 2023, pp. 659–680.

[2] D. Nahavandi, R. Alizadehsani, A. Khosravi, and U. R. Acharya, "Application of artificial intelligence in wearable devices: Opportunities and challenges," *Computer Methods and Programs in Biomedicine*, vol. 213, p. 106541, 2022.

[3] 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.

[4] K. Yue, R. Jin, C.-W. Wong, and H. Dai, "Communication-efficient federated learning via predictive coding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 3, pp. 369–380, 2022.

[5] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3400–3413, 2019.

[6] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *IEEE International Conference on Communications*. IEEE, 2019, pp. 1–7.

[7] D. Li, Y. Guo, D. Liu, Y. Ren, R. Hu, and Z. Guan, "Client-Edge-Cloud Hierarchical Federated Learning Based on Generative Adversarial Networks," in *IEEE International Conference on Knowledge Graph*. IEEE, 2023, pp. 160–167.

[8] G. Gad, Z. M. Fadlullah, K. Rabie, and M. M. Fouda, "Communication-efficient privacy-preserving federated learning via knowledge distillation for human activity recognition systems," in *IEEE International Conference on Communications*. IEEE, 2023, pp. 1572–1578.

[9] Q. Xu, M. Wu, X. Li, K. Mao, and Z. Chen, "Contrastive distillation with regularized knowledge for deep model compression on sensor-based human activity recognition," *IEEE Transactions on Industrial Cyber-Physical Systems*, 2023.

[10] P. Qi, X. Zhou, Y. Ding, Z. Zhang, S. Zheng, and Z. Li, "FedBKD: Heterogenous Federated Learning via Bidirectional Knowledge Distillation for Modulation Classification in IoT-Edge System," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 1, pp. 189–204, 2022.

[11] S. Deng, J. Chen, D. Teng, C. Yang, D. Chen, T. Jia, and H. Wang, "LHAR: Lightweight Human Activity Recognition on Knowledge Distillation," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[12] S. Wang., T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[13] C. Xie, S. Koyejo, and I. Gupta, "Asynchronous federated optimization," *arXiv preprint arXiv:1903.03934*, 2019.

[14] A. Herzog, R. Southam, O. Belarbi, S. Anwar, M. Bullo, P. Carnelli, and A. Khan, "Selective updates and adaptive masking for communication-efficient federated learning," *IEEE Transactions on Green Communications and Networking*, vol. 8, no. 2, pp. 852–864, 2024.

[15] E. Hallaji, R. Razavi-Far, M. Saif, B. Wang, and Q. Yang, "Decentralized federated learning: A survey on security and privacy," *IEEE Transactions on Big Data*, 2024.

[16] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[17] J. Reyes-Ortiz, D. Anguita, A. Ghio, L. Oneto, and X. Parra, "Human Activity Recognition Using Smartphones," UCI Machine Learning Repository, 2013. [Online]. Available: https://doi.org/10.24432/C54S4K

[18] G. Gad and Z. Fadlullah, "Federated learning via augmented knowledge distillation for heterogeneous deep human activity recognition systems," *Sensors*, vol. 23, no. 1, p. 6, 2022.

[19] J. Liu, Q. Zeng, H. Xu, Y. Xu, Z. Wang, and H. Huang, "Adaptive block-wise regularization and knowledge distillation for enhancing federated learning," *IEEE/ACM Transactions on Networking*, 2023.

[20] H. Wen, Y. Wu, J. Hu, Z. Wang, H. Duan, and G. Min, "Communication-efficient federated learning on non-IID data using two-step knowledge distillation," *IEEE Internet of Things Journal*, vol. 10, no. 19, pp. 17 307–17 322, 2023.

[21] H. Q. Le, M. N. Nguyen, C. M. Thwal, Y. Qiao, C. Zhang, and C. S. Hong, "FedMEKT: Distillation-based Embedding Knowledge Transfer for Multimodal Federated Learning," *Neural Networks*, vol. 183, p. 107017, 2025.

[22] E. Cunegatti, M. Farina, D. Bucur, and G. Iacca, "Understanding sparse neural networks from their topology via multipartite graph representations," *arXiv preprint arXiv:2305.16886*, 2023.

[23] H. A. Desai, A. Hilal, and H. Eldardiry, "Resource-Efficient Federated Learning for Heterogenous and Resource-Constrained Environments," *arXiv preprint arXiv:2308.13662*, 2023.

[24] M. Sandhu, D. Silvera-Tawil, W. Lu, P. Borges, and B. Kusy, "Exploring Activity Recognition in Multi-device Environments using Hierarchical Federated Learning," in *International Conference on Pervasive Computing and Communications - Workshops and other Affiliated Events*. IEEE, 2024, pp. 720–726.

[25] S. Agac and O. D. Incel, "Boosting the Performance of Lightweight HAR Models with Attention and Knowledge Distillation," in *International Conference on Intelligent Environments*. IEEE, 2024, pp. 1–8.