Multi-Frequency Federated Learning for Human Activity Recognition Using Head-Worn Sensors

Dario Fenoglio[†*], Mohan Li[†*], Davide Casnici[†], Matias Laporte[†],
Shkurta Gashi[‡], Silvia Santini[†], Martin Gjoreski[†], Marc Langheinrich[†]
[†] Università della Svizzera italiana, Switzerland
{dario.fenoglio, mohan.li}@usi.ch
[‡] ETH Zurich, Switzerland

***Abstract*—Human Activity Recognition (HAR) benefits various application domains, including health and elderly care. Traditional HAR involves constructing pipelines reliant on centralized user data, which can pose privacy concerns as they necessitate the uploading of user data to a centralized server. This work proposes multi-frequency Federated Learning (FL) to enable: (1) privacy-aware ML; (2) joint ML model learning across devices with varying sampling frequency. We focus on head-worn devices (e.g., earbuds and smart glasses), a relatively unexplored domain compared to traditional smartwatch- or smartphone-based HAR. Results have shown improvements on two datasets against frequency-specific approaches, indicating a promising future in the multi-frequency FL-HAR task. The proposed network's implementation is publicly available for further research and development.[**]**

***Index Terms*—Federated Learning, Human Activity Recognition (HAR), Head-worn sensors, Earables, Glasses**

I. Introduction

Human Activity Recognition (HAR) refers to the process of identifying and categorizing the specific activities performed by an individual through the analysis of various sensor data [1]. This technology has rapidly emerged as an essential tool with wide-reaching applications across numerous domains in recent decades. The importance of HAR lies in its ability to enable the provision of helpful context information that can be applied in various fields. Among these applications is the management of chronic diseases, where HAR can track and monitor patients' physical activities to provide tailored healthcare solutions and interventions [2]. Similarly, in healthcare settings, detecting abnormal patient behavior can be automated through HAR, thus enhancing the efficiency of care and possibly even saving lives [3]. Furthermore, HAR offers insights into individual habits and routines, from personalized fitness tracking to occupational health and safety, allowing for the design of personalized programs to enhance overall well-being [4]. In summary, HAR not only adds a technological edge to many industries but also brings a human-centered approach to monitoring and understanding behavior. HAR applications are expanding and their impact on daily life is profound, shaping a new era of personalized, context-aware services and care.

[*]Equal contribution
[**]https://github.com/dariofenoglio98/Multi_frequency_STResNet.git

Different approaches have been used in HAR, namely vision-based and wearable-based. Vision-based approaches utilize external sensors like cameras that provide a powerful way to recognize and analyze human activities by capturing visual data. However, they suffer from significant drawbacks. The efficacy of these tools is compromised when users are out of the sensing field. Privacy invasion is an even more critical concern, making them potentially unsuitable in personal or sensitive environments. These challenges underscore the limitations of vision-based HAR, particularly in the context of ubiquitous computing, where seamless and unobtrusive integration is a key consideration.

In contrast, wearable-based HAR offers a more flexible and privacy-respecting solution. Magnetometers, gyroscopes, and accelerometers — essential components of inertial measurement units — wearable sensors have become prominent in HAR for their ability to overcome some of the limitations of external sensors, such as their poor portability. Being compact and easily integrated into everyday devices like earbuds or glasses, these sensors align seamlessly with the principles of ubiquitous computing. Their unobtrusive nature allows for continuous monitoring and data collection, permitting a more intuitive and user-centered approach to activity recognition [5].

Unfortunately, the introduction of machine learning, specifically deep learning (DL), still complicates privacy preservation as these algorithms require data for training. Even when data are collected in a more privacy-preserving manner, transmitting it to a centralized server for model training cancels users' exclusive ownership of the latter, opening up possibilities for data misuse and inadvertent exposure.

In response to these challenges, Google introduced Federated Learning (FL) [6] in 2016, an innovative ML paradigm that enables neural network models to be trained across multiple decentralized devices or servers, each possessing its data samples locally. This approach significantly enhances data privacy and security by ensuring user data remains on their device, presenting a promising equilibrium between advancing HAR capabilities and maintaining robust user privacy protections.

An interesting aspect of FL is device heterogeneity, i.e., different devices can collaboratively train a model with a common goal (e.g., HAR). Recent HAR studies (such as FLAME [7]) have explored multi-device FL in synchronized

setups, e.g., if a user wears earbuds and smart glasses simultaneously, FLAME could be used to train joint models across the synchronized devices.

Differently from the existing work, this study explores multi-frequency FL in asynchronous setups, i.e., users can have diverse active sensors (e.g., combinations of magnetometer, gyroscope, and accelerometer); these sensors can sample at various frequencies (e.g., some devices at 5Hz, others at 40Hz); and, these devices — despite utilizing varying sensors and recording frequencies — can collaboratively build a joint HAR model. Furthermore, we focus on FL for head-worn wearable devices, a relatively unexplored domain compared to traditional smartwatch or smartphone-based HAR.

To this end, this work makes the following contribution: *a novel multi-frequency FL method for HAR*. The method is based on an existing end-to-end learning approach, Spectro-Temporal Residual Network (STResNet) [8], that we adapted to work in a federated and multi-frequency setup. We compared the novel method on two datasets against centralized and frequency-specific models. The results show that our multi-frequency model allows the exploitation of all available data (i.e., all clients and sensors), thus outperforming frequency-specific models. In addition, our model demonstrated high flexibility and robustness, maintaining high performance while accepting a variable number of input sensors.

The rest of the paper is organized as follows. Section II introduces related works in FL and HAR. Section III provides details of the two datasets used in this work. Sections IV and V present, respectively, the used methods and the experiments with the corresponding results. Section VI discusses these results, while Section VII provides concluding remarks from them, as well as potential directions for future research.

II. Related Work

This section provides an overview of the related research areas. We highlight two main differences between previous works and this study: (1) we focus on FL for head-worn wearable devices, a relatively unexplored domain compared to traditional smartwatch- or smartphone-based HAR; (2) to the best of our knowledge, this is the first study that explores a multi-frequency FL method for head-worn HAR in an asynchronous setup. The closest method to ours is FLAME, which utilizes synchronization (time alignment) across devices from the same user. Thus, FLAME is useful for scenarios where users simultaneously use multiple devices. On the other hand, in our proposed method, users need only one of the multiple devices (or sensors) to participate in the FL process.

A. Federated Learning

The FL community has been growing fast since it was introduced. As concluded from recent surveys [9], [10], most of the research advances focus on core challenges, including reducing computing costs, tackling system or statistic heterogeneity, and enhancing privacy protection. Optimized communication and aggregation strategies [11], [12] have been proposed to relieve the computational burden without hurting the overall performance. Considering the participation of heterogeneous hardware, efforts have been made with adaptive task-assigning based on device capability, dropout of incapable devices, or tolerance as a more friendly approach [10]. To handle the typical non-independent non-identically distributed (non-IID) and unbalanced local data, one may resort to personalization through clustering [13] or model adaptation [14], [15]. Even though the data are locally preserved, the vanilla model sharing and aggregation are exposed to malicious attacks such as data poisoning. Many related advanced works have followed Secure Aggregation [16] and Differential Privacy [17] as reliable solutions. Besides HAR, FL has been proven successful in many other fields such as the Internet of Things [18], healthcare [15], vehicular systems [19], and recommender systems [20].

B. Human Activity Recognition

A recent survey [21] has well captured recent advances in the HAR task with various sensing modalities. Frequently used signals include inertia, electrocardiogram (ECG), and vital signs such as respiration and temperature. In addition to the general HAR task [22], other targets such as hand gesture recognition [23] are also highly related. The unexploited unlabeled data have also gained much recent attention [24]. A large amount of data remains at the edges and is not applied for model training because labeling them is an overwhelming and knowledge-demanding task. More researchers now try to incorporate them in a semi-supervised or unsupervised manner to achieve better performance [25].

C. Federated Learning for Human Activity Recognition

Even though current smart devices can collect billions of sensor samples every day with great potential to improve HAR performance, the cost of data transmission and the invasion of individual privacy are difficult to comprehend. Konstantin et al. [26] have made one of the earliest contributions that deploy the HAR learning task with the FL framework to tackle privacy issues. Tu et al. [27] proposed FedDL, where the center HAR model merges local updates based on a dynamic sharing scheme to speed up the convergence while maintaining high accuracy. Ouyang et al. [28] introduced ClusterFL, a similarity-aware FL approach to cluster different clients in a multitasking manner to achieve high model accuracy and low communication overhead for HAR applications. Xiao et al. [29] developed advanced feature extraction approaches from sensor data to improve the overall performance. Unsupervised learning and personalization have also been proven powerful as future directions [30].

Despite the great success of FL on HAR, few of these stuides have explored wearable device data collected with earbuds and glasses, or in a multi-frequency setup. A recent-to-date work [31] practiced leveraging wearable smart glasses data to achieve personalized treatments and interventions for enhanced healthcare outcomes. Following their promising results, we explore multi-frequency FL for HAR.

III. DATASETS

Our head-worn dataset consists of the *USI-HEAR Dataset* [32] and the *OCOsense Smart Glasses HAR Dataset* [31]. An overview is summarized in Table I.

TABLE I
SUMMARY OF DATASETS.

Dataset	*USI-HEAR*	*OCOsense*
Participants	30	24
Device	eSense earbuds	OCOsense Smart Glasses
Sensors	3-axis accelerometer 3-axis gyroscope	3-axis accelerometer 3-axis gyroscope 3-axis magnetometer pressure sensor 3-axis Euler virtual sensor
Activities	Speak and Walk Head Shaking Speaking Nodding Eating Walking Staying	Sitting Standing Laying Walking Transition Jogging Stair Climbing

A. USI-HEAR Dataset

The *USI-HEAR* Dataset was collected with the eSense earbuds developed by Nokia Bell Labs [33]. These earbuds consist of two Bluetooth-enabled units, each equipped with one microphone, while the left unit further houses one 6-axis Inertial Measurement Unit (IMU) sensor, comprising one 3-axis accelerometer and one 3-axis gyroscope.

Participants were provided with one left earbud (containing the IMU). They performed seven scripted activities, each lasting 3 minutes, with the data subsequently transferred to the experimenter's laptop for verification. The experiment involved seven distinct activities, each carefully selected to represent a range of non-interacting and interacting behaviors. These activities were:

- ***Speak and Walk***: Participants combined walking and speaking, illustrating the complexity of simultaneous activities.
- ***Head Shaking***: Participants moved their heads horizontally, with different intensities and intervals, representing a gesture of disagreement or denial.
- ***Speaking***: Participants engaged in verbal communication with the experimenters, reflecting natural speech patterns and intonations.
- ***Nodding***: Participants were instructed to nod their heads with different intensities and intervals, simulating a common gesture of agreement or acknowledgment.
- ***Eating***: Participants consumed food, allowing for the observation of jaw movements and related motions.
- ***Walking***: Participants walked at a comfortable pace, capturing the dynamics of regular locomotion.
- ***Staying***: Participants remained still or seated, providing a baseline for motion detection.

Overall, the dataset comprises more than 10 hours of streaming data for each channel of both gyroscope and accelerometer, with a universal downsampled frequency of 50Hz.

B. OCOsense Smart Glasses HAR Dataset

The dataset was collected in 2022 by Emteq Labs using their *OCOsense Smart Glasses* [34]. The device is equipped with one 3-axis accelerometer, one 3-axis gyroscope, one 3-axis magnetometer, one pressure sensor (barometer), and one 3-axis Euler virtual sensor to combine data from the accelerometer and gyroscope to provide the orientation of the glasses in three dimensions (yaw, pitch, roll).

24 participants were asked to perform the following activities while wearing the smart glasses:

- ***Sitting*** (39.3%) includes *Sitting*, *Sitting Still*, *Sitting-looking around*, *Sitting-using a PC*, and *Sitting-using a phone*.
- ***Standing*** (27.3%) includes *Standing*, *Standing Still*, *Standing-looking around*, and *Standing-using a phone*.
- ***Laying*** (18.3%) includes *On the back*, *On the left side*, *On the right side*, and *On the stomach*.
- ***Walking*** (9.1%) includes *Walking*, *Walking-looking around*, and *Walking-using a phone*.
- ***Transition*** (2.2%) includes *Sitting down* and *Standing up*.
- ***Jogging*** (1.7%) includes *Jogging*.
- ***Stair climbing*** (1.7%) includes *Stair climbing*.

The dataset contains 1.7M samples (9.5 hours duration in total) approximately equally distributed among all participants. The sampling frequency matches the earbuds dataset at 50Hz.

IV. METHODS

In this section, we describe our comparisons among different centralized training models, to choose the best-performing model for FL. In addition, we describe the FL setup and the multi-frequency network.

A. Centralized Machine Learning

The DL pipeline employed in this study encompasses five distinct neural network architectures. Among these, four are 1D convolutional neural networks (ConvNets), and the remaining model is a deep multimodal spectro-temporal residual neural network (STResNet) [35].

TABLE II
COMPARISON OF STRUCTURES AMONG FOUR 1D CONVNETS

Model	#Conv.	#Dense	Act. functions	#params.
ConvNet1	3	4	LeakyReLU	78,680,443
ConvNet2	3	3	ReLU	5,962,453
ConvNet3	2	2	ReLU	1,937,575
ConvNet4	2	2	PReLU/ReLU	1,909,031

Centralized model comparison among the four 1D convolutional neural networks (ConvNets) of structures, activation functions, and number of parameters. **#Conv.**: number of the convolutional layers. **#Dense**: number of the dense layers. **Act. functions**: activation functions. **#params.**: number of the parameters.

The four 1D-convolutional deep models are characterized by different configurations of convolutional and dense layers, along with specific activation functions. The architectures are summarized in Table II. They share common features such as the softmax function as the final activation function and sparse

categorical cross-entropy as the loss function. To mitigate overfitting, L2 regularization (rate = 0.0001) and dropout (rate = 0.5) were employed. Additionally, early stopping was implemented using the validation loss as the stopping criterion, further safeguarding against overfitting.

The STResNet model builds upon the concept of end-to-end unimodal time-series classification using residual networks. It incorporates multimodal and spectro-temporal information fusion, essential components of a successful HAR system. STResNet extracts channel-specific spectro-temporal information for each sensor channel. The spectral information is obtained by calculating the logarithm of the amplitude spectrogram in decibels for each input window. The temporal representation is extracted by residual blocks containing CNN layers with 1-dimensional (1D) filters. The shortcut connections in the residual blocks combat the gradient vanishing problem, making training more tractable. L2 regularization and dropout are applied to the dense layers, and the final output is provided by a softmax layer, representing class probability for each of the seven activities. Among the models finally employed in this study, STResNet stands out as the second most computationally demanding model, with a total of 14,005,415 trainable parameters. This substantial complexity is indicative of the model's capacity to capture intricate patterns and relationships within the data. However, it is second only to ConvNet1 in terms of computational demand, reflecting a careful balance between model complexity and computational efficiency within the overall DL pipeline.

B. Federated Learning

In our study, we implemented the Weighted Federated Averaging algorithm via the Flower library [36]. Each participating client performs one training epoch on their local dataset and then forwards their model weights and the count of their training samples to the server. To balance the contributions from all clients, the server applies a weighted average to these weights. We chose this approach due to the heterogeneous distribution of data across our datasets, resulting in varying numbers of samples per client. The model training was conducted using a sparse categorical cross-entropy loss function, combined with an Adam optimizer set at a learning rate of 0.0001. Considering the average number of samples per client, we standardized the batch size at 32 samples for both datasets. The final model was chosen based on its highest accuracy on the validation set during training.

For our FL environment, we exclusively employed the STResNet model, due to its superior performance in predicting activity on the *USI-HEAR* dataset (as shown in V-A). To ensure a precise comparison between centralized and federated learning, STResNet was implemented under both these settings across the *OCOsense* and *USI-HEAR* datasets. Additionally, to evaluate the robustness of FL under different conditions, we initially tested it with all available clients in the datasets, followed by a gradual reduction in the number of training clients. This allowed us to assess the impact of dataset size on the performance and effectiveness of the FL framework.

Specifically, to validate both centralized and federated approaches using the same test set clients, we employed a person-independent 5-fold cross-validation (ensuring non-overlapping test clients across different folds). This involved dividing the dataset into five distinct groups of clients, ensuring each client's data was excluded once from the training process once. In each fold, clients not involved in the training were evenly split into validation and test sets. This strategy ensured a comprehensive and unbiased evaluation of both centralized and federated learning approaches.

C. Multi-Frequency STResNet

To address the challenge of clients equipped with sensors operating at varying frequencies, we developed a novel Multi-Frequency STResNet model. This model processes input signals from sensors with different sampling frequencies. For instance, we simulated a scenario in which half of the clients' sensors operated in a low-battery mode at 5Hz (or 3Hz), while the other half functioned at 40Hz. Instead of employing separate models for each battery mode — which would reduce the dataset size — we propose a versatile model that expands the potential data and client pool. Our approach involves creating distinct spectral-temporal encoders for each sensor (or channel), tailoring both temporal and spectral feature extraction to the sensor's sampling frequency.

As depicted in Figure 1, our encoder receives a raw sensor signal as input and processes it through parallel pathways, which we call *temporal* and *spectral*. In the temporal pathway, after an initial batch normalization step, the signal undergoes four residual blocks. Each block consists of two 1D convolutions and two Leaky ReLU activations, followed by a single 1D max pooling operation to effectively capture the temporal dynamics. Concurrently, in the spectral pathway, the signal is transformed into a spectrogram and processed through three blocks, each containing a 2D convolution and a Leaky ReLU activation. To integrate the multidimensional spectral features, a dense layer and dropout regularization are employed.

Additionally, we introduced a context vector to mask activations from sensors that are not present (left-side of Figure 2). This vector assigns a value of 1 where sensor input is available and 0 otherwise. Likewise, input channels are set to 0 in the absence of sensor data. Prior to the fully connected layers, the context vector is utilized to zero out activations from absent sensors, thereby preventing consideration of non-zero values due to the bias values involved in the neural networks' training.

The rest of Figure 2 illustrates how our Multi-Frequency STResNet accommodates users with sensors in both low- and full-battery modes. We evaluated our model in a centralized environment using person-independent 10-fold Monte Carlo cross-validation, ensuring consistent training and testing on the same client groups for each iteration. We benchmarked our model against those trained exclusively on clients with either 5Hz or 40Hz signals, a model trained on all clients at 5Hz (including downsampling those at 40Hz), and an ideal scenario where all clients operate at 40Hz (i.e., no low-battery mode). Furthermore, as our model is capable of being evaluated on

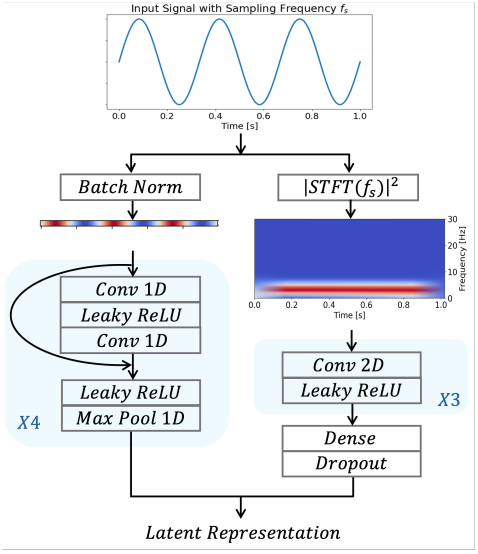

Fig. 1. Encoder architecture for a single input channel. The input undergoes both a temporal and a spectral encoding to generate the latent representation.

test clients at both 5Hz and 40Hz, assessments were conducted under each condition to facilitate a fair comparison with frequency-specific models.

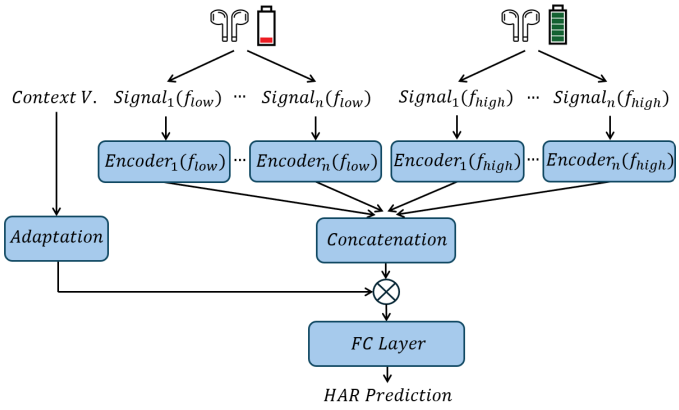

Fig. 2. Multi-frequency model architecture designed for devices in low- and full-battery modes. A context vector is combined with the concatenation of the encoders' outputs to zero-out any absent sensors. Finally, fully connected layers produce HAR predictions from the combined latent representation.

V. Experiments and Results

This section outlines the experiments conducted and the corresponding results. In Subsection V-A, we evaluate five DL models across various sensor streams, identifying the STResNet model as the most accurate. Subsection V-B illustrates the comparable accuracy of federated and centralized learning, emphasizing FL's scalability and adaptability, even with varying numbers of training clients, while still ensuring user privacy. Finally, Subsection V-C presents the results for the novel approach, demonstrating the Multi-Frequency STResNet model's efficiency in handling different sensor frequencies, outperforming frequency-specific setups. It should be noted that all models in our experiments were person-independent, ensuring non-overlapping training and testing client groups, which is crucial for the generalizability and applicability of our findings to real-world scenarios.

A. Model Selection with Centralized Training

We first carry out centralized training with all five DL models to select the one with the best performance. This model will then be used for the FL setup. We use different sensor streams for comprehensive validation from the *USI-HEAR* dataset: accelerometer (ACC), gyroscope (GYR), magnitude (MAG), first-order derivatives (DER), and all combined (ALL). The results are shown in Table III with accuracy scores expressed as percentages, and standard deviations as percentage points.

TABLE III
Comparison of Results Among Centralized Models

Model	ACC	GYR	MAG	DER	ALL
ConvNet1	38.95 ±17.29	50.77 ±15.53	52.89 ±13.97	61.74 ±17.52	57.12 ±13.50
ConvNet2	**42.91 ±14.94**	56.02 ±14.54	55.12 ±13.83	62.23 ±17.17	65.49 ±12.11
ConvNet3	34.86 ±14.05	45.64 ±17.74	**58.32 ±13.05**	56.03 ±15.32	61.78 ±11.43
ConvNet4	20.01 ±5.63	43.16 ±18.75	36.23 ±10.07	38.33 ±16.31	40.55 ±13.25
STResNet	38.86 ±16.31	**57.13 ±13.31**	57.12 ±14.74	**62.55 ±11.91**	**69.22 ±11.78**

Comparison of accuracy (with standard deviation) among five centralized models with different input settings, validated on the *USI-HEAR* dataset. STResNet outperforms with inputs of all sensor streams combined. **ACC**: accelerometer. **GYR**: gyroscope. **MAG**: magnitude. **DER**: first-order derivatives. **ALL**: all above combined.

The results show that among the original sensor streams, the gyroscope emerges as the most pertinent. Within virtual sensor streams, the derivatives stand out as a better input for most models compared to the magnitudes. The highest performance is attained with the STResNet model, taking advantage of all the sensor streams and exploiting the spectral information of the signals, underscoring the importance of a comprehensive approach in sensor data analysis. This consistent performance across different scenarios justifies our choice of STResNet for the rest of the experimental setup.

B. Comparison between Centralized and Federated Learning

Table IV presents the accuracy, F1-score, and cross-entropy loss for both centralized and federated learning techniques across the *USI-HEAR* and *OCOsense* datasets. These results underscore the efficacy of FL in handling diverse HAR datasets. Notably, the performance metrics in the FL setup were comparable to those in the centralized setup. This equivalence highlights FL's ability to effectively train a global model without the need for direct data sharing from clients.

Table V compares the performance of centralized and federated learning techniques with varying numbers of training clients. Following this comparison, it becomes evident that our FL approach demonstrates robustness under these conditions. The similarity in results between centralized and federated methods, even with different number of participants,

TABLE IV
COMPARISON BETWEEN CENTRALIZED AND FEDERATED LEARNING WITH ALL PARTICIPANTS

	Centralized		Federated	
	USI-HEAR	***OCOsense***	***USI-HEAR***	***OCOsense***
Accuracy	**70.0 ± 4.7**	84.9 ± 2.7	69.43 ± 4.14	**85.19 ± 1.99**
F1-Score	**70.2 ± 4.8**	**87.8 ± 1.7**	67.26 ± 4.02	87.72 ± 1.72
CE Loss	**1.052 ± 0.165**	**0.389 ± 0.083**	1.105 ± 0.242	0.444 ± 0.085

Federated learning shows competitive performance to centralized learning with both *USI-HEAR* and *OCOsense* datasets. **CE Loss**: cross-entropy loss.

indicates a high degree of scalability and adaptability in the FL approach. This suggests that FL can maintain consistent performance despite variations in the size of training data and the number of clients, crucial in real-world applications where client availability may vary.

TABLE V
COMPARISON BETWEEN CENTRALIZED AND FEDERATED LEARNING WITH VARIOUS PARTICIPANTS NUMBERS

#Part.	**Centralized**		**Federated**	
	USI-HEAR	***OCOsense***	***USI-HEAR***	***OCOsense***
2	55.55 ± 1.87	**77.14 ± 3.05**	**55.60 ± 2.43**	75.19 ± 1.97
3	**58.41 ± 4.85**	**80.77 ± 2.15**	56.36 ± 3.87	79.18 ± 3.21
4	**60.02 ± 4.22**	**81.92 ± 2.78**	58.64 ± 5.33	81.24 ± 2.47
6	59.73 ± 7.72	84.43 ± 2.37	**60.34 ± 5.39**	**84.76 ± 1.76**
8	**62.02 ± 5.13**	85.16 ± 2.01	61.53 ± 6.00	**86.60 ± 1.42**

Comparison of F1-Score (with standard deviation) between Centralized and Federated Learning trained under different numbers of participants, with both *USI-HEAR* and *OCOsense* datasets. **#Part.**: number of training participants.

C. Multi-Frequency Model

Table VI provides a comprehensive comparison of F1-scores for the *USI-HEAR* and *OCOsense* datasets, across different numbers of participants included in the training (*#Part.* and their respective frequency) and diverse frequency settings. This table compares the performance of our Multi-Frequency STResNet model against various configurations as outlined in the first column: exclusively 5Hz clients, all clients downsampled to 5Hz (*Down-5Hz*), exclusively 40Hz clients, and an ideal scenario with all clients at 40Hz (*Ideal 40Hz*). Our multi-frequency model was trained on all the clients (both 5Hz and 40Hz) at their original frequency. To ensure a fair comparison, our our multi-frequency model (*Multi-*), capable of processing both 5Hz and 40Hz frequencies, was tested under both these conditions (*Multi-5Hz* and *Multi-40Hz*) on the same test clients.

Notably, our multi-frequency model exhibited superior F1-scores in both *Multi-5Hz* and *Multi-40Hz* configurations across both datasets, demonstrating its effectiveness over single-frequency settings (5Hz and 40Hz clients). This improvement is mainly attributed to the model's ability to utilize all available original data to train a unified model. However, it should also be noted that the *Down-5Hz* model (i.e., an approach that first downsamples all the data to the lowest joint frequency (5Hz in this case) and then trains a model) slightly outperforms the multi-frequency approach.

Precisely, in the *USI-HEAR* dataset, the multi-frequency model achieved F1-scores of 63.26% ± 3.08% and 65.53% ± 3.61% for 5Hz and 40Hz, respectively. These scores significantly surpass the single-frequency models' scores of 59.30% ± 3.33% (5Hz) and 62.57% ± 4.49% (40Hz), and they closely align with the ideal outcome where all clients operate at a sampling frequency of 40Hz.

In the case of the *OCOsense* dataset, a similar pattern emerges. Notably, as shown in Table VI, downsampling to 5Hz resulted in a higher F1-score (86.17% ± 2.29%) than the ideal scenario of all 40Hz clients (85.77% ± 2.17%), suggesting that a 5Hz sampling rate is sufficient for accurate HAR tasks. Our multi-frequency model also performed better on the same test clients when downsampled to 5Hz compared to 40Hz. Once again, the multi-frequency model outperformed single-frequency setups, underscoring the versatility of our model in effectively handling diverse sensor frequencies. For a more detailed analysis, refer to the extended table in the appendix (Section VIII), which includes both F1-scores and accuracy metrics. Additionally, the appendix presents analogous results for experiments conducted with a *critical-battery* mode (i.e., a lower frequency of 3Hz).

TABLE VI
COMPARISON BETWEEN CONFIGURATIONS WITH DIFFERENT FREQUENCY

Model		***USI-HEAR***		***OCOsense***	
		F1-Score	**#Part.**	**F1-Score**	**#Part.**
5Hz		59.30 ± 3.33	7_{5Hz}	72.81 ± 14.17	5_{5Hz}
Down-5Hz		65.38 ± 2.39	14_{5Hz}	86.17 ± 2.29	10_{5Hz}
40Hz		62.57 ± 4.49	7_{40Hz}	79.74 ± 2.66	5_{40Hz}
Multi- (Ours)	5Hz	63.26 ± 3.08	7_{5Hz}, 7_{40Hz}	85.38 ± 2.52	5_{5Hz}, 5_{40Hz}
	40Hz	65.53 ± 3.61		83.45 ± 1.99	
Ideal 40Hz		69.14 ± 2.96	14_{40Hz}	85.77 ± 2.17	10_{40Hz}

Comparison of mean F1-Scores (with standard deviation) for different frequency configurations, with both *USI-HEAR* and *OCOsense* datasets, with half of the participants sampled at 5Hz and the other half at 40Hz. Our Multi-Frequency model is the only one that allows training with all clients at their original sampling frequencies. **#Part.**: number of training participants.

VI. DISCUSSION

We discuss the experimental results and our findings in this section, respectively, on the innovations in HAR modalities (Subsection VI-A), insights into multi-frequency in HAR tasks (Subsection VI-B), and directions for future research (Subsection VI-C).

A. Novel Sensor Modalities in HAR

Compared to traditional vision-based HAR, which demands image or video data from cameras, wearable-based HAR improves device availability and reduces data size with sensor streams while preserving competitive performance. As the main raw data source, inertial information is collected with IMUs, which can be fused with different devices and accessories. In this paper, we looked into datasets from earbuds and glasses, two relatively underexplored domains compared to other works, focusing on mobile and wrist-worn devices.

It should be noted that the on-body location of sensors could potentially introduce divergence in results: activities with subtle changes on face or head movements, such as speaking or eating, could be more distinguishable with sensor data above the neck. Likewise, activities with limb motion,

such as running and walking, could be more clearly detected with sensors on the wrist or from the waist down. Predictably, a more systematic and comprehensive HAR framework should have multiple modalities to fully cover the range of human activities. Our work could be a good start to the concentration on more delicate differentiation of facial and head activities.

B. Multi-Frequency HAR in The Wild

Recording frequency has a direct impact on the size of the data stream, storage requirements, and computational load. From the perspective of an end-user or client, we hope the model has an acceptable performance — a generally faster response, higher device refresh rate, and less running-time memory taken — even with minimum data frequency. From the perspective of a server or aggregator, accommodating a wider range of data frequency signifies a larger size of the training dataset, participation from more diverse users, and thus better performance for all. Furthermore, multi-frequency tolerance in FL introduces great possibilities for communication cost reduction and heterogeneous systems collaboration. We hope this work may pave the way for further research on this track.

In terms of multi-frequency in HAR tasks, we noticed that the *Down-5Hz* model, i.e., an approach that first downsamples all the data to the lowest joint frequency (5Hz in this case) and then trains a model, has competitive results with the multi-frequency setup. This indicates a possible frequency threshold in the specific experimental datasets, where information introduced by a higher frequency is redundant. However, we expect that in real-life applications where the activities to be recognized are more dynamic, a 5Hz sampling rate would not be sufficient to achieve acceptable HAR performance. Thus, we hope our work inspires more flexible, lightweight, and energy-friendly frameworks.

C. Challenges and Future Research

Besides the future research directions above, there are more opportunities and challenges in the FL-HAR field. Related to this paper, we hope to address two important aspects in later works. First, we did not leverage the possibility of unlabeled data in this task. While millions of data streams appear in various sensor devices, end-users usually do not hold clean and well-labeled data. As our next step, we will explore both unsupervised and semi-supervised methods to further improve the model's quality and enlarge the training data size. Second, our work did not apply any personalization techniques, such as local re-training, which could be a solution to address the notorious non-IID problem in heterogeneous data. In the future, we may deploy a clustering or privacy-friendly knowledge-sharing approach to achieve better performance.

VII. Conclusion

This paper has introduced a multi-frequency FL framework for the HAR task. Our framework builds one unified model to accommodate sensor data from different frequencies and classify human activities under various device status, such as low-battery mode. We tested our framework with simulated 5Hz and 40Hz data streams from two datasets (one collected from earbuds, another from smart glasses), obtaining promising results in recognizing human activities with a frequency-fusion approach over heterogeneous sensors. We hope our work brings attention to the data frequency issue in HAR, inspiring more research on heterogeneous-friendly FL systems.

Acknowledgment

This study was funded by the projects TRUST-ME (205121L_214991), SmartCHANGE (GA No. 101080965), and XAI-PAC (PZ00P2_216405). Shkurta Gashi is supported by an ETH AI Center postdoctoral fellowship.

References

[1] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys*, vol. 46, no. 3, pp. 33:1–33:33, Jan. 2014.

[2] R. Liu, A. A. Ramli, H. Zhang, E. Henricson, and X. Liu, "An Overview of Human Activity Recognition Using Wearable Sensors: Healthcare and Artificial Intelligence," in *Internet of Things – ICIOT 2021*, B. Tekinerdogan, Y. Wang, and L.-J. Zhang, Eds. Cham: Springer International Publishing, 2022, pp. 1–14.

[3] F. Serpush, M. B. Menhaj, B. Masoumi, and B. Karasfi, "Wearable Sensor-Based Human Activity Recognition in the Smart Healthcare System," *Computational Intelligence and Neuroscience*, vol. 2022, p. e1391906, Feb. 2022.

[4] S. Consolvo *et al.*, "Activity sensing in the wild: A field trial of ubifit garden," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '08. New York, NY, USA: Association for Computing Machinery, Apr. 2008, pp. 1797–1806.

[5] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical Human Activity Recognition Using Wearable Sensors," *Sensors*, vol. 15, no. 12, pp. 31 314–31 338, Dec. 2015.

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2017, pp. 1273–1282.

[7] H. Cho, A. Mathur, and F. Kawsar, "FLAME: Federated Learning across Multi-device Environments," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–29, Sep. 2022.

[8] M. Gjoreski *et al.*, "Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors," *Information Fusion*, vol. 62, pp. 47–62, Oct. 2020.

[9] P. Kairouz *et al.*, "Advances and Open Problems in Federated Learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, Jun. 2021.

[10] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, May 2020.

[11] K. Bonawitz *et al.*, "Towards Federated Learning at Scale: System Design," *Proceedings of Machine Learning and Systems*, vol. 1, pp. 374–388, Apr. 2019.

[12] J. Hamer, M. Mohri, and A. T. Suresh, "FedBoost: A Communication-Efficient Algorithm for Federated Learning," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 3973–3983.

[13] F. Sattler, K.-R. Muller, and W. Samek, "Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3710–3722, Aug. 2021.

[14] F. Hanzely and P. Richtárik, "Federated Learning of a Mixture of Global and Local Models," Feb. 2021.

[15] D. Fenoglio *et al.*, "Federated Learning for Privacy-aware Cognitive Workload Estimation," in *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia*, ser. MUM '23. New York, NY, USA: Association for Computing Machinery, Dec. 2023, pp. 25–36.

[16] K. Bonawitz *et al.*, "Practical Secure Aggregation for Federated Learning on User-Held Data," Nov. 2016.
[17] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning Differentially Private Recurrent Language Models," Feb. 2018.
[18] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. Vincent Poor, "Federated Learning for Internet of Things: A Comprehensive Survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021.
[19] Z. Du, C. Wu, T. Yoshinaga, K.-L. A. Yau, Y. Ji, and J. Li, "Federated Learning for Vehicular Internet of Things: Recent Advances and Open Issues," *IEEE Open Journal of the Computer Society*, vol. 1, pp. 45–61, 2020.
[20] Z. Alamgir, F. K. Khan, and S. Karim, "Federated recommenders: Methods, challenges and future," *Cluster Computing*, vol. 25, no. 6, pp. 4075–4096, Dec. 2022.
[21] H. F. Nweke, Y. W. Teh, M. A. Al-garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Systems with Applications*, vol. 105, pp. 233–261, Sep. 2018.
[22] M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," *Future Generation Computer Systems*, vol. 81, pp. 307–313, Apr. 2018.
[23] G. Yuan, X. Liu, Q. Yan, S. Qiao, Z. Wang, and L. Yuan, "Hand Gesture Recognition Using Deep Feature Fusion Network Based on Wearable Sensors," *IEEE Sensors Journal*, vol. 21, no. 1, pp. 539–547, Jan. 2021.
[24] H. Haresamudram, I. Essa, and T. Plötz, "Assessing the State of Self-Supervised Human Activity Recognition Using Wearables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 116:1–116:47, Sep. 2022.
[25] Y. Jain, C. I. Tang, C. Min, F. Kawsar, and A. Mathur, "ColloSSL: Collaborative Self-Supervised Learning for Human Activity Recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 17:1–17:28, Mar. 2022.
[26] K. Sozinov, V. Vlassov, and S. Girdzijauskas, "Human Activity Recognition Using Federated Learning," in *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, Dec. 2018, pp. 1103–1111.
[27] L. Tu, X. Ouyang, J. Zhou, Y. He, and G. Xing, "FedDL: Federated Learning via Dynamic Layer Sharing for Human Activity Recognition," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '21. New York, NY, USA: Association for Computing Machinery, Nov. 2021, pp. 15–28.
[28] X. Ouyang, Z. Xie, J. Zhou, J. Huang, and G. Xing, "ClusterFL: A similarity-aware federated learning system for human activity recognition," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '21. New York, NY, USA: Association for Computing Machinery, Jun. 2021, pp. 54–66.
[29] Z. Xiao, X. Xu, H. Xing, F. Song, X. Wang, and B. Zhao, "A federated learning system with enhanced feature extraction for human activity recognition," *Knowledge-Based Systems*, vol. 229, p. 107338, Oct. 2021.
[30] Y. Li, X. Wang, and L. An, "Hierarchical Clustering-based Personalized Federated Learning for Robust and Fair Human Activity Recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 20:1–20:38, Mar. 2023.
[31] B. Sazdov *et al.*, "Privacy-aware Human Activity Recognition with Smart Glasses for Digital Therapeutics," in *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, ser. UbiComp/ISWC '23 Adjunct. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 592–596.
[32] M. Laporte, D. Casnici, M. Gjoreski, S. Gashi, S. Santini, and M. Langheinrich. (2024, Mar.) Usi-hear dataset. [Online]. Available: https://doi.org/10.5281/zenodo.10843791
[33] F. Kawsar, C. Min, A. Mathur, A. Montanari, U. G. Acer, and M. Van den Broeck, "eSense: Open Earable Platform for Human Sensing," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '18. New York, NY, USA: Association for Computing Machinery, Nov. 2018, pp. 371–372.
[34] J. Archer *et al.*, "OCOsenseTM Smart Glasses for Analyzing Facial Expressions Using Optomyographic Sensors," *IEEE Pervasive Computing*, vol. 22, no. 3, pp. 18–26, Jul. 2023.
[35] J. Zhang, Y. Zheng, and D. Qi, "Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017.
[36] D. J. Beutel *et al.*, "Flower: A Friendly Federated Learning Research Framework," Mar. 2022.

VIII. Appendix

This appendix provides additional results and metrics, complementing those presented in Section V-C. These results offer a broader view of the Multi-Frequency STResNet model's performance under various conditions and different metrics.

A. Extended Performance Metrics

Table VII presents a comprehensive set of performance metrics for the Multi-Frequency STResNet model, introducing accuracy besides the F1-score. This table complements Table VI, reported in the main results.

TABLE VII
Comparison of Accuracy, F1-Score, and Number of Training Participants (5-40Hz)

Model		*USI-HEAR*			*OCOsense*		
		Accuracy	**F1-Score**	**#Part.**	**Accuracy**	**F1-Score**	**#Part.**
5Hz		60.17 ± 3.40	59.30 ± 3.33	7_{5Hz}	73.06 ± 6.43	72.81 ± 14.17	5_{5Hz}
Down-5Hz		65.74 ± 2.61	65.38 ± 2.39	14_{5Hz}	81.36 ± 2.44	86.17 ± 2.29	10_{5Hz}
40Hz		63.75 ± 4.05	62.57 ± 4.49	7_{40Hz}	76.90 ± 3.20	79.74 ± 2.66	5_{40Hz}
Multi-(Ours)	5Hz	63.69 ± 3.28	63.26 ± 3.08	7_{5Hz}, 7_{40Hz}	80.00 ± 1.87	85.38 ± 2.52	5_{5Hz}, 5_{40Hz}
	40Hz	66.34 ± 3.31	65.53 ± 3.61		78.78 ± 2.51	83.45 ± 1.99	
Ideal 40Hz		69.73 ± 3.02	69.14 ± 2.96	14_{40Hz}	80.77 ± 2.47	85.77 ± 2.17	10_{40Hz}

B. Additional Experiments

In addition to our primary experiments at a sampling frequency of 5Hz, we investigated the model's performance with signals sampled at a lower frequency of 3Hz. Table VIII presents the accuracy and F1-scores of our Multi-Frequency STResNet model, comparing its performance with configurations for exclusively 3Hz clients, clients downsampled to 3Hz (*Down-3Hz*), exclusively 40Hz clients, and an ideal scenario where all clients operate at 40Hz (*Ideal 40Hz*). Despite the reduced sampling rate, our multi-frequency model preserves its advantages over single-frequency models, registering an F1-score of 80.69% ± 3.64% and 80.49% ± 2.51% for 40Hz and 3Hz respectively, as compared to 78.90% ± 3.01% and 66.26% ± 18.15% for their single-frequency counterparts. Overall, it is observed that lowering the sampling frequency to 3Hz generally leads to a decline in HAR prediction accuracy. Nevertheless, particularly within the *OCOsense* dataset, downsampling all clients to 3Hz yields results that are still comparable to using the full 40Hz frequency.

TABLE VIII
Comparison of Accuracy, F1-Score, and Number of Training Participants (3-40Hz)

Model		*USI-HEAR*			*OCOsense*		
		Accuracy	**F1-Score**	**#Part.**	**Accuracy**	**F1-Score**	**#Part.**
3Hz		57.48 ± 4.45	57.06 ± 4.24	7_{3Hz}	70.48 ± 6.34	67.05 ± 15.89	5_{3Hz}
Down-3Hz		63.68 ± 2.43	63.09 ± 2.46	14_{3Hz}	80.20 ± 3.33	82.95 ± 5.28	10_{3Hz}
40Hz		62.94 ± 3.65	62.05 ± 3.95	7_{40Hz}	76.40 ± 3.66	78.96 ± 2.22	5_{40Hz}
Multi-(Ours)	3Hz	61.48 ± 3.45	61.13 ± 3.52	7_{3Hz}, 7_{40Hz}	78.48 ± 1.95	80.49 ± 2.51	5_{3Hz}, 5_{40Hz}
	40Hz	64.85 ± 2.20	64.10 ± 2.03		77.44 ± 2.56	80.69 ± 3.64	
Ideal 40Hz		68.58 ± 3.01	68.08 ± 3.09	14_{40Hz}	79.63 ± 2.20	82.86 ± 1.81	10_{40Hz}