

FedFitTech: A Baseline in Federated Learning for Fitness Tracking

Zeyneddin Oz*

Ubiquitous Computing, University of Siegen
Siegen, NRW, Germany
zeyneddin.oez@uni-siegen.de

Marius Bock

Ubiquitous Computing, University of Siegen
Siegen, NRW, Germany
marius.bock@uni-siegen.de

Shreyas Korde

Ubiquitous Computing, University of Siegen
Siegen, NRW, Germany
shreyas.korde@student.uni-siegen.de

Kristof Van Laerhoven

Ubiquitous Computing, University of Siegen
Siegen, NRW, Germany
kvl@eti.uni-siegen.de

Abstract

The rapid evolution of sensors and resource-efficient machine learning models has spurred the widespread adoption of wearable fitness tracking devices. Equipped with inertial sensors, such devices can continuously capture physical movements for fitness technology (FitTech), enabling applications from sports optimization to preventive healthcare. Traditional Centralized Learning approaches to detect fitness activities struggle with data privacy concerns, regulatory restrictions, and communication inefficiencies. In contrast, Federated Learning enables a decentralized model training by communicating model updates rather than potentially private wearable sensor data. Applying Federated Learning to FitTech presents unique challenges, such as data imbalance, lack of labeled data, heterogeneous user activities, and trade-offs between personalization and generalization. To simplify research on FitTech in Federated Learning, we present the FedFitTech baseline, under the Flower framework, which is publicly available and widely used by both industry and academic researchers. Additionally, to illustrate its usage, this paper presents a case study that implements a system based on the FedFitTech baseline, incorporating a client-side early stopping strategy and comparing the results. For instance, this system allows wearable devices to optimize the trade-off between capturing common fitness activities and preserving individuals' nuances, thereby enhancing both the scalability and efficiency of privacy-aware fitness tracking applications. The results show that this reduces the overall redundant communications by 13%, while maintaining the overall recognition performance at a negligible recognition cost by 1%. Thus, the FedFitTech baseline creates a foundation for a wide range of new research and development opportunities in FitTech, and it is available as open source at: <https://github.com/shreyaskorde16/FedFitTech>

CCS Concepts

- Computing methodologies → Distributed artificial intelligence; Machine learning;
- Human-centered computing → Ubiquitous and mobile computing systems and tools.

Keywords

Federated learning, Human Activity Recognition, Fitness Tracking

ACM Reference Format:

Zeyneddin Oz, Shreyas Korde, Marius Bock, and Kristof Van Laerhoven. 2025. FedFitTech: A Baseline in Federated Learning for Fitness Tracking. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '25), October 12–16, 2025, Espoo, Finland*. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3714394.3756332>

1 Introduction

With recent advances in sensor technologies and resource-efficient machine learning models, wearable fitness trackers in various forms, such as wrist watches, glasses, earbuds, and rings, have gained widespread use. These devices generate a vast amount of valuable sensor data for Human Activity Recognition (HAR) applications, in which body movements can be analyzed in detail with minimal intrusion. For example, inertial sensors can continuously monitor motion and gestures at specific body locations, thus providing a detailed representation of the user's motion patterns. This makes such devices particularly useful for an extensive range of applications in fitness technology (FitTech), including medical assistance and the optimization of complex work processes [1].

FitTech is a crossover term that combines fitness and technology, encompassing innovations that enhance fitness, wellness, and health experiences. Using wearable devices, mobile apps, and virtual personal training platforms, FitTech can monitor physical activity, provide real-time feedback to users, and tailor exercise programs to their needs. Through machine learning, FitTech promises to transform the fitness industry by making workouts more personalized, efficient, and data-driven, helping users achieve better training results.

Current research tends to focus largely on Centralized Learning of fitness activities, which can be used to develop powerful machine learning models by being able to process large datasets and leverage substantial computational power. However, Centralized Learning requires sensor data to be shared and stored in a central location, which can deter data sharing due to regulatory restrictions (e.g., the European General Data Protection Regulation (GDPR) [2] and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

UbiComp Companion '25, Espoo, Finland

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1477-1/2025/10
<https://doi.org/10.1145/3714394.3756332>

the California Consumer Privacy Act (CCPA) [3]), making users reluctant to grant access for training. Moreover, Centralized Learning struggles to meet the needs for scalability and communication efficiency.

To address these challenges, Federated Learning [4] has emerged as an alternative to Centralized Learning, offering a model communication mechanism instead of raw data sharing. In Federated Learning, a global model is shared with devices (also known as clients, nodes, or workers) for local training using local data. These tuned local models are then sent back to the server, which aggregates them to update the global model. The updated global model is redistributed to the devices for further training rounds. At the end of the process, all devices obtain a global model trained on all local data, while data privacy is maintained, as sensitive raw data are not transmitted to external sources (see Fig. 1). Additionally, Federated Learning offers benefits such as improved data availability, scalability, fault recovery, and communication efficiency [5].

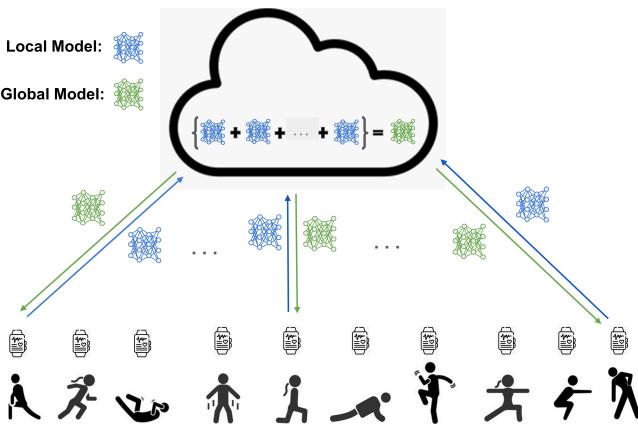


Figure 1: In FedFitTech, wearable fitness devices locally train a global classification model for fitness activities (green model) shared by a server. Tuned local models (blue models) are shared with the server for aggregation to update a global model across all participants. Then, this global model is redistributed for further training rounds, which leads to a private system while profiting from outer patterns.

Although Federated Learning is widely recognized as a viable approach for training machine learning models on decentralized devices, its potential impact on FitTech remains underexplored. Wearables such as smartwatches typically share similar sensors, particularly MEMS inertial sensors, as well as processing and wireless networking components, making FitTech applications relatively homogeneous compared to other environments. This similarity results in consistent computation and communication capabilities across smartwatches, enabling training of the same global model with synchronized communication.

Additionally, researchers of [4] identified communication as the primary challenge in Federated Learning systems. However, unlike other domains, fitness tracking does not require daily communication. Since people do not exercise frequently throughout the day, data is not generated continuously, and wearable devices accumulate small datasets. This allows for less frequent and more

convenient communication times, such as during charging or when connected to unmetered Wi-Fi. Moreover, modern processors make the computation of tiny machine learning models more efficient, reducing costs. Furthermore, authors of [6] compared various Federated Learning algorithms and found that Federated Learning can achieve comparable or even superior performance to Centralized Learning techniques in the HAR domain. Thus, the FitTech environment is naturally well suited to Federated Learning.

On the other hand, each person exhibits unique movement patterns and the performing of the same exercise can vary depending on individual performance. The position of the sensors on a person can affect the recorded motion patterns, even when repeating the same actions [7]. Factors such as gender, age, weight, height, and exercise frequency further contribute to these unique characteristics [8]. For instance, one person might jog once a week, while another person might jog every other day, eventually leading to data imbalances. The literature shows that Federated Learning underperforms when device data is imbalanced, as excessive external patterns can reduce performance on local data due to the personal nature of movements [9]. Moreover, annotated local data is often limited, requiring patterns from other people to accurately detect a person's exercise. Given the advantages of Federated Learning in FitTech, we thus present in this paper FedFitTech, a Federated Learning baseline designed for research in the FitTech domain.

Furthermore, we give an example using the FedFitTech baseline by designing a client-side early stopping implementation and analyzing its benefits. In FitTech, due to the varied motion patterns of the user, the global model should not override individual variations. Thus, early stopping may be helpful when the global model underperforms on local datasets. The attractiveness of implementing client-side early stopping in FedFitTech is that, in the standard Federated Learning setting [4], all devices participate in the training of the global model for a fixed number of communication rounds. This can lead to unnecessary resource consumption for devices that do not need to attend all training rounds. In addition, the FitTech domain requires incorporating only common patterns from others without losing the influence of personal data in the global model. Although Federated Learning provides a private way to benefit from external patterns, users need the global model to be tailored to local data. Therefore, it is essential to design a Federated Learning system that balances personalization and generalization in FitTech. Thus, our case study of the FedFitTech baseline involves an early stopping method to mitigate these challenges.

Our contributions are threefold:

- **Federated Fitness Activity Recognition:** We highlight the domain of fitness activity recognition (or HAR) as interesting and suitable for the field of Federated Learning.
- **Flower Baseline for Fitness Tracking:** We designed FedFitTech as an easy-to-use baseline built under the Flower framework [10], a widely adopted and openly available platform that enables the implementation and benchmarking of reproducible experiments.
- **Use Case Experiment of Early Stopping:** We present an example usage of this baseline, by allowing client-side early stopping to reduce energy consumption and thus improving the overall efficiency of the system.

2 Related Work

Several studies have investigated various implementations of Federated Learning for HAR, addressing different challenges through novel strategies. FL-PMI [7] proposes a system leveraging unlabeled data by applying a cleaning process with auto-labeling propagation to support smart healthcare applications. Meta-HAR [8] employs a model-agnostic meta-learning approach that extracts adaptive signal embeddings from sensor data and integrates a task-specific classification network through fine-tuning. PMF [11] adopts a secure data handling approach by combining differential privacy with secure multi-party computation and introduces an adaptive weight adjustment mechanism to ensure fair contributions to a global model. Also, the work by [12] presents a perceptive extraction network system that enhances data representation while employing encryption and decryption to mitigate potential data leakage risks.

To address the challenges of inadequate data and secure data sharing, researchers in [13] propose a two-dimensional framework that employs both horizontal and vertical FL paradigms, while another study [14] utilizes specialized training algorithms aimed at increasing data diversity and fostering inclusivity for devices with inferior network conditions.

Focusing on model training improvements, several works explore client grouping and dynamic layer sharing. PS-PFL [15] introduces a personalized FL model based on profile similarity that aggregates local models through weighted computations. An attention-based clustering mechanism in [16] calculates an optimal weighted model combination to encourage similar clients to group together without relying on a single global model. ClusterFL [17] leverages a cluster indicator matrix and iterative optimization techniques to update client model weights, facilitating cluster-wise straggler dropout and correlation-based client selection. ProtoHAR [9] integrates a prototype-guided technique that separates representation learning from categorization, thereby achieving uniform activity feature representation and improved convergence. Also, dynamic layer-sharing is investigated by FedCLAR [18], which selectively shares portions of model weights based on a hierarchical clustering, and by FedDL [19], which merges lower layers of models to capture user-specific similarities while enhancing communication efficiency.

Furthermore, to enhance privacy, the researcher of PDP-FL (Personalized Differential Privacy-based Federated Learning) [20] specifically investigated the fitness tracking application using two stages. First, it allows users to customize their privacy levels, with corresponding noise added to their local model prior to sharing with a server. The second stage involves the server adding further noise to ensure global privacy protection.

The weaknesses noted across these studies include that FL-PMI [7] may require additional computational resources, which could impact battery life, and does not explicitly address sensor inaccuracies. Meta-HAR [8] may incur higher computational and communication overhead, posing challenges for very large-scale deployments. While promising, PMF [11] requires further empirical validation of its performance under large-scale data scenarios and the integration of scalable privacy-preserving measures. The two-dimensional framework by [13] could potentially expose finer-grained activity details under adversarial conditions, and the approach in [14] may face challenges with dynamic client participation and scaling

to real-world deployments. Additionally, client grouping strategies such as PS-PFL [15] incur extra computational overhead on resource-constrained devices, and ClusterFL [17] might expose user-cluster associations under adversarial conditions while necessitating an additional optimization step. Lastly, ProtoHAR [9] assumes a stationary data distribution, which might limit its adaptability in continuously evolving environments. FedCLAR [18] assumes the availability of ample labeled data — a condition that may not hold universally — and FedDL [19] requires further integration of robust privacy measures along with additional validation of its effectiveness beyond HAR. Besides, while PDP-FL [20] enhances privacy by allowing varied user privacy preferences, this diversity can impact model performance and effectiveness. In other words, differing noise levels, corresponding to individual privacy choices, influence global model aggregation. To add more, PDP-FL simulates the fitness tracking environment using image data, which does not fully capture the complexities of the domain. Fitness data often have temporal dependencies, varying sampling rates, and different types of noise characteristics compared to image pixels.

While other studies addressed HAR applications, this study especially focuses on the FitTech domain using FL, similar to PDP-FL. However, our FedFitTech baseline utilizes a real fitness tracking dataset and a tiny model, reflecting original FitTech characteristics and providing further exploration of FL potential in this area.

About designing a case study of FedFitTech, the implementation of early stopping in the FL setting has already been explored in FLrce [21] and FLASH [22]. FLrce employs a resource-efficient early stopping strategy tailored for FL. FLASH leverages concept drift detection through client-side early stopping and server-side adaptive optimization. However, their complex mechanisms increase computational overhead and risk premature termination. Also, these methods are primarily designed for image recognition in non-FitTech domains. Our case study especially focuses on the FitTech domain and employs an early stopping approach based on the stability of the validation F1-score over a sliding window [23]. While this method is simple to implement, it ensures that the global model retains valuable patterns learned from all clients, including those with less significant contributions, which is an essential factor in FL for FitTech. Moreover, the case study emphasizes communication efficiency and balanced generalization. Despite a negligible F1-score drop, our case study allows clients to benefit from external patterns and terminate training when global weights diverge from local patterns, preventing redundant computation and communication.

3 Experimental Setting

3.1 Model

Over the past decade, Deep Learning (DL) methods have become the dominant approach in inertial-based HAR, consistently outperforming traditional Machine Learning algorithms [24, 25]. For this reason, we used TinyHAR [26], which is a lightweight DL model specifically designed for HAR. TinyHAR leverages a combination of convolutional layers, a transformer block featuring self-attention and fully connected layers, and a Long Short-Term Memory (LSTM) to extract and model features from sensor data efficiently. Initially,

convolutional layers are applied to each sensor channel to capture local patterns. This is followed by a transformer block that facilitates effective cross-channel feature interaction. Next, a fully connected layer fuses the extracted features, which are then processed by an LSTM to capture long-term temporal dependencies. Finally, a self-attention layer dynamically recalibrates the importance of features across time. This architecture achieves competitive performance with a significantly reduced parameter count, making it highly suitable for deployment on resource-constrained devices.

3.2 Dataset

We used inertial-based data from the WEAR (Wearable and Ego-centric Activity Recognition) (<https://mariusbock.github.io/wear/>, [27]) dataset, which consists of labeled activities performed by 22 participants and categorized into three types: jogging, stretching, and strength exercises. The jogging category includes five labels: normal, rotating arms, skipping, sidesteps, and butt-kicks. The stretching category also comprises five labels: triceps, lunging, shoulders, hamstrings, and lumbar rotation. Lastly, the strength category encompasses eight labels: push-ups, push-ups (complex), sit-ups, sit-ups (complex), burpees, lunges, lunges (complex), and bench dips. In addition, the dataset includes a NULL label that represents unlabeled or undefined periods where no activity was explicitly carried out. It serves as a placeholder for missing annotations and can be useful for filtering, handling uncertainty, or distinguishing between known and unknown states. As a result, participants engaged in 18 different outdoor sports activities. Unlike previous egocentric datasets, WEAR features an inertial sensor placement aligned with recent real-world application trends. It is the first dataset collected across multiple outdoor locations, each presenting varying surface conditions without revealing cues about the performed activities. Moreover, in FL in HAR applications, the selected dataset must mirror real-world conditions, as it should accommodate a large number of clients and a diverse array of classes and subjects. Herewith, based on the recent work [28], the WEAR dataset remains the best option for fitness tracking, since other datasets cannot fulfill the needs due to having fewer participants, classes, and labeled data.

3.3 Hyperparameters and Framework

This section provides a short summary of the most important parameters and framework used in our experiment.

Client setting: In the dataset, subjects 1 and 19, and also subjects 15 and 20, are from the same participants. These two participants were re-recorded in a different season and environment, resulting in 24 subjects in total. These two additional subjects are treated as separate devices for the same individuals to reflect real-world FitTech scenarios, thus setting the number of clients to 24.

Data splitting: In the WEAR dataset, participants perform the same activity several times in different time periods randomly, to represent realistic scenarios. This means that the activity patterns of people can change over time. For instance, while a person jogs fast at the beginning of exercise, he/she can feel tired later, which causes body movements to be slower. Furthermore, considering factors such as weather conditions and changes on the ground, it

cannot be expected that a person will make the same body movements. Therefore, the method of splitting local data into training and testing requires special scrutiny. For this reason, we select the first 20 % of each label in the time series to add to the testing data, and the remaining 80 % is left for training. In this way, for example, when a person jogs both at the start and later in the session, the test set will include both jogging bouts, ensuring temporal variation within each class.

Window size: After conducting multiple experiments, we determined that a window size of 100 is the optimal option, corresponding to a duration of 2 seconds in the time series sampled at 50 Hz.

Batch size: Despite the impact of other hyperparameters differing based on the used model and the dataset, [29] reported that a batch size of 32 emerged as the optimal choice for image data. Nevertheless, we set it to 32, since the hyperparameter recommendation for time series remains underexplored, and this value is used in the literature as well [19].

Optimizer: We employed the Adam (Adaptive Moment Estimation) [30] optimizer because it has been shown to converge quickly and efficiently for complex models and large datasets. By adaptively tuning the learning rate of each parameter and using momentum to smooth the optimization, it requires less hyperparameter tuning, and these were found to be suitable for Federated Learning contexts [8, 18].

Learning rate: Set to 0.001, following the literature [12].

Communication round: Based on the existing work in Federated Learning in HAR [6], we have set the communication round to 100, as it is also a standard setting for other domains in FL in general.

Local epoch: Set to 1, following the literature [9, 18].

Model Aggregation: We used FedAvg [4] strategy since it is commonly employed as the main aggregation algorithm or a baseline in Federated Learning in HAR applications [6, 8, 9, 12, 14–19].

Early stopping: Following the method in [23], we implemented early stopping based on the stability of the validation F1-score over a sliding window. The hyperparameters of this method are set based on our several experiments to determine the optimum values for this paper's case study. As a result of experimentation, the patience value is set to 5, and the threshold is set to 0.01 for the stopping criterion.

Framework: We use Flower (<https://flower.ai/>, [10]), which is a widely adopted and openly available framework suitable for both industrial and academic research in Federated Learning. It is designed to simplify the implementation of Federated Learning by managing communication, orchestration, and model aggregation. Flower's support for large-scale experiments, including millions of simulated clients, demonstrates its remarkable scalability for real-world federated deployments. In addition, it offers broad compatibility with existing and emerging machine learning frameworks, diverse operating systems, and a wide range of hardware platforms, including servers and mobile devices. This high level of interoperability makes it a highly versatile tool for a variety of research applications.

4 Evaluation Results

The evaluation of this work compares FedFitTech with its client-side early stopping case study, considering both communication loads and global model performance on local data.

4.1 Communication Cost

Fig. 2 illustrates that 9 out of 24 clients stopped early from training the global model, highlighting that prolonged participation does not guarantee continued benefit. The earliest stop was observed at the 40th communication round (see client ID 3 in Fig. 2), and some clients decided to stop early, even though they had participated in training the global model in more than 80 communication rounds.

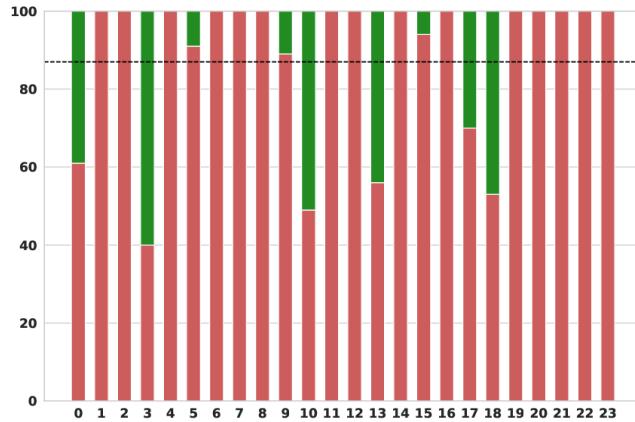


Figure 2: Amount of training rounds each client attended in the case study. The Y-axis shows the communication round, and the X-axis shows the client IDs. Red bars depict the communication cost, while green bars show the saved communication rounds, and the dashed black line is their mean.

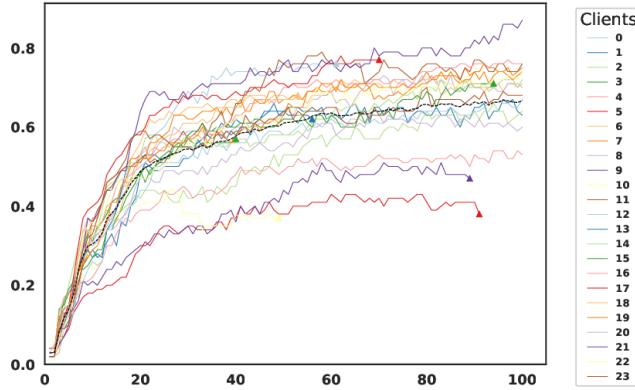


Figure 3: Communication rounds (X-axis) versus F1-scores (Y-axis). The dashed black line shows the mean values of all local performance, and the triangle markers depict rounds of early stopped clients.

Moreover, Fig. 3 shows the performance of the F1-score of clients who are dropped early (see triangle markers within the plot). Some

clients were unable to increase their local F1-score even after 30 rounds of training. Also, as can be seen from the black dashed line, the overall global model F1-score continues to increase, despite some clients being dropped.

4.2 Model Performance

The client-based F1-scores are depicted in Fig. 4 for both the baseline and our case study of early stopping. The FedFitTech baseline resulted in a 68% mean F1-score across all clients, with the early stopping case study performing closely with a mean F1-score of 67%. Additionally, the result presents that some of the clients even have better F1-scores in the system in the case study, such as the clients represented with these ID numbers: 2, 7, 9, 10, 11, 12, 14, 17, 19, 21, and 22, with 11 out of 24 clients obtaining a better performance compared to the FedFitTech baseline. In conclusion, the overall recognition performance remains almost unchanged, while early stopping reduces communication costs and overall computational load.

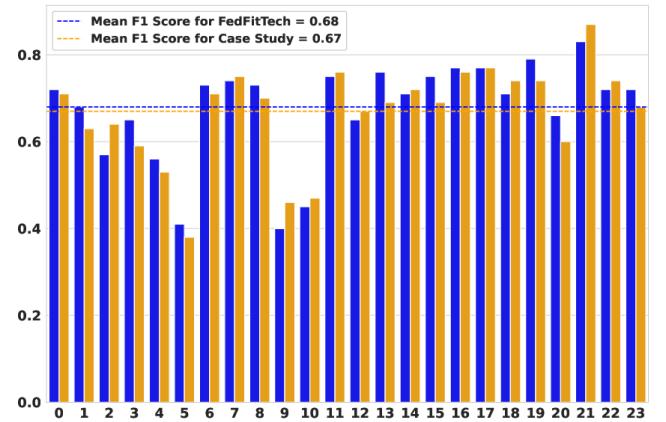


Figure 4: The different F1-scores (Y-axis) over all clients (X-axis) in the WEAR dataset, for the FedFitTech baseline and the early stopping case study experiment.

In addition, Fig. 5 and Fig. 6 depict the label-based F1 scores of the FedFitTech baseline and the case study, respectively. Actual label names in these figures are as follows: 'A': 'NULL', 'B': 'jogging', 'C': 'jogging (rotating arms)', 'D': 'jogging (skipping)', 'E': 'jogging (sidesteps)', 'F': 'jogging (butt-kicks)', 'G': 'stretching (triceps)', 'H': 'stretching (lunging)', 'I': 'stretching (shoulders)', 'J': 'stretching (hamstrings)', 'K': 'stretching (lumbar rotation)', 'L': 'push-ups', 'M': 'push-ups (complex)', 'N': 'sit-ups', 'O': 'sit-ups (complex)', 'P': 'burpees', 'Q': 'lunges', 'R': 'lunges (complex)', and 'S': 'bench-dips'.

Despite early stopping decreasing the overall mean F1-score slightly by 1% (see Fig. 4), some of the clients reached a remarkable performance increase for some of the labels. For instance, the push-up fitness activity (see label L in the confusion matrices) had an F1-score for Client 2 in FedFitTech of 0%, while it is 76% in the case study. Client 18, which was the one who stopped training the first (see Fig. 2), achieved a 47% reduction in communication costs, and its overall F1-score increased from 71% to 74%. Also, tracking some of the activities (e.g., 'stretching (lunging)': H) increased from 55% to 86%.

0	0.58	0.89	0.78	0.78	0.93	0.85	0.91	0.82	0.77	0.67	0.71	0.48	0.50	0.61	0.60	0.57	0.81	0.88	0.55
1	0.72	0.00	0.64	0.74	0.75	0.80	0.90	0.82	0.74	0.95	0.67	0.59	0.42	0.75	0.54	0.50	0.75	0.84	0.77
2	0.72	0.90	0.00	0.67	0.93	0.80	0.83	0.54	0.00	0.60	0.89	0.00	0.48	0.78	0.69	0.70	0.38	0.73	0.18
3	0.78	0.86	0.82	0.73	0.89	0.75	1.00	0.00	0.82	0.27	0.76	0.52	0.67	0.73	0.91	0.30	0.12	0.62	0.82
4	0.62	0.93	1.00	0.82	0.80	0.20	0.86	0.93	0.67	0.42	0.00	0.78	0.63	0.91	0.54	0.00	0.13	0.00	0.37
5	0.55	0.00	0.55	0.88	0.86	0.00	0.96	0.92	0.00	0.00	0.94	0.46	0.00	0.62	0.71	0.00	0.19	0.17	0.00
6	0.68	0.88	0.90	1.00	0.89	0.95	1.00	0.95	0.46	0.12	0.95	0.59	0.50	0.52	0.52	0.33	0.96	0.90	0.82
7	0.80	0.78	0.82	0.72	0.90	0.00	0.71	0.78	0.75	0.95	0.95	0.48	0.33	0.90	0.87	0.90	0.62	0.95	0.93
8	0.65	0.90	0.90	0.89	0.95	1.00	0.80	0.89	0.46	0.95	0.90	0.56	0.18	0.74	0.80	0.80	0.64	0.38	0.57
9	0.41	0.89	0.89	0.63	0.17	0.82	0.71	0.00	0.00	0.60	0.40	0.00	0.70	0.29	0.43	0.20	0.38	0.00	0.00
10	0.54	0.90	0.87	1.00	0.28	0.25	0.84	0.67	0.00	0.00	0.41	0.00	0.82	0.74	0.33	0.00	0.00	0.00	0.00
11	0.55	0.89	0.95	0.93	0.95	0.88	0.78	0.33	0.15	0.90	1.00	0.55	0.00	0.87	0.92	0.94	0.84	0.82	0.89
12	0.78	0.46	0.70	0.90	1.00	0.90	0.82	0.56	0.86	0.82	0.78	0.35	0.00	0.78	0.80	0.57	0.57	0.00	0.80
13	0.85	0.91	0.84	0.94	0.91	0.89	0.78	0.00	0.89	0.95	0.89	0.56	0.40	0.91	0.64	0.84	0.82	0.71	0.71
14	0.69	0.94	0.76	0.75	1.00	0.88	0.95	0.78	1.00	0.86	0.90	0.39	0.15	0.52	0.75	0.55	0.57	0.78	0.31
15	0.78	0.94	0.83	1.00	0.95	0.94	0.00	0.80	0.86	0.72	0.59	0.70	0.62	0.64	0.57	0.78	0.90	0.75	0.80
16	0.77	0.94	0.90	0.95	0.94	0.89	0.84	0.70	0.80	0.49	0.80	0.62	0.80	1.00	0.56	0.55	0.82	0.80	0.80
17	0.72	0.90	0.86	0.86	0.82	0.95	0.94	0.95	0.89	0.90	1.00	0.56	0.00	0.49	0.43	0.77	0.90	0.95	0.84
18	0.70	1.00	0.80	0.82	0.82	0.76	0.70	0.55	0.88	0.31	0.75	0.73	0.86	0.91	0.53	0.82	0.00	0.61	0.90
19	0.80	0.95	0.95	0.86	1.00	0.71	0.84	0.80	0.90	0.80	0.61	0.55	0.67	0.75	0.88	0.73	0.47	1.00	0.75
20	0.77	0.80	0.94	0.64	0.55	0.55	0.33	0.70	0.88	0.75	0.89	0.58	0.00	0.43	0.64	0.62	0.88	0.64	0.88
21	0.73	0.88	0.94	0.89	0.94	0.94	0.88	0.95	0.89	0.95	1.00	0.45	0.22	1.00	0.89	0.89	0.74	0.94	0.62
22	0.62	0.95	0.94	0.89	1.00	0.94	1.00	0.00	0.18	0.95	0.95	0.89	0.95	1.00	0.84	0.90	0.60	0.00	0.00
23	0.77	1.00	0.94	0.88	0.95	0.93	0.95	0.27	0.57	0.56	0.88	0.57	0.53	0.82	0.59	0.63	0.57	0.84	0.73

Figure 5: The client-label F1-scores of FedFitTech, with on the Y-axis the client IDs and on the X-axis the activity labels (as described in section 4.2).

4.3 Discussion

Fig. 2 and 3 show that this client attrition, while seemingly modest in this small-scale experiment, represents a significant concern in real-world Federated Learning deployments involving a large number of clients, as 37.5% of clients decreased their communication loads. Besides, these findings suggest that client-specific factors, such as data heterogeneity, resource constraints, or local model convergence, may impact the stability of participation within the Federated Learning framework. However, as shown in Fig. 3, the global model continues to improve despite some clients no longer participating in training. Also, Fig. 4 demonstrates that in the case study, 45.8% of the clients increased their overall mean F1-scores, while communication costs decreased by 13% (see Fig. 2).

5 Conclusions and Future Work

Wearable devices provide rich data that enable in-depth insights into user fitness activities within the domain of Fitness Technology (FitTech). However, manual data annotation in FitTech is both time-consuming and prone to error, creating a significant bottleneck. Users can offload the tedious task of annotating activities while benefiting from diverse and high quality labels. Although Centralized Learning approaches can harness all this annotated data, they often struggle with latency, limited scalability, and heightened privacy risks.

We argue in this paper that Federated Learning aligns well with the characteristics of FitTech, as clients have largely similar sensors, similar processing power, and mostly common activity classes, while requiring relatively infrequent communication between a global model and the clients. This work thus introduces FedFitTech, designed to explore the FitTech domain within Federated Learning, built upon the user-friendly, widely adopted, and open-source Flower framework. Moreover, we shared a case study as an example usage of this baseline, which involves a client-side early stopping strategy. The case study is compared with the FedFitTech baseline, considering communication loads and model performance metrics.

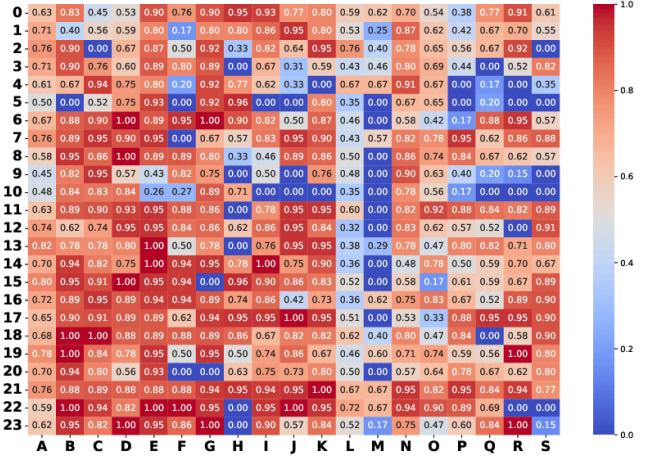


Figure 6: The client-label F1-scores for this paper’s case study of early stopping. When comparing with Fig. 5, it can be seen that overall classification performance remains similar, with certain activities (such as L - push-ups and M - complex push-ups) remaining challenging across users.

The findings show that our case study mitigates the communication burden with a negligible mean F1-score drop by balancing model generalization.

The FedFitTech baseline provides various promising directions for future research in Federated Learning for the FitTech domain. For instance, exploring transfer learning implementations can accelerate convergence and enable clients to benefit from pre-existing patterns from the outset. Additionally, integrating differential privacy techniques is crucial for improving data privacy and security, addressing a fundamental concern in FitTech. Furthermore, since the utilized WEAR dataset in FedFitTech consists of both inertial and ego-perspective camera data, investigating multimodal mechanisms has significant potential for handling diverse sensor data and mitigating data heterogeneity.

Moreover, exploring efficient local model clustering methods and then sharing cluster-specific global models with clients could reduce performance differences across clients and increase overall recognition performance by facilitating collaboration among users with similar behavioral patterns (e.g., activity routines, age, weight, or gender). In conclusion, we hope that our FedFitTech baseline will leverage the publicly available and widely used Flower framework, opening up a new research domain and development opportunities in FitTech.

The FedFitTech baseline is available as an open source repository at: <https://github.com/shreyaskorde16/FedFitTech>

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-IDs 506589320 and 520256321.

References

- [1] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3):1–33, 2014.
- [2] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed.*, 10(3152676), 2017.
- [3] Lydia de la Torre. A guide to the california consumer privacy act of 2018. Available at SSRN 3275571, 2018.
- [4] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [5] Jessica Chia Liu, Jack Goetz, Srijan Sen, and Ambuj Tewari. Learning from others without sacrificing privacy: Simulation comparing centralized and federated machine learning on mobile health data. *JMIR mHealth and uHealth*, 9(3):e23728, 2021.
- [6] Sannara Ek, François Portet, Philippe Lalanda, and German Vega. Evaluation and comparison of federated learning algorithms for human activity recognition on smartphones. *Pervasive and Mobile Computing*, 87:101714, 2022.
- [7] KS Arikumar, Sahaya Beni Prathiba, Mamoun Alazab, Thippa Reddy Gadekallu, Sharmin Pandya, Javed Masood Khan, and Rajalakshmi Shenbaga Moorthy. Fl-pmi: federated learning-based person movement identification through wearable devices in smart healthcare systems. *Sensors*, 22(4):1377, 2022.
- [8] Chenglin Li, Di Niu, Bei Jiang, Xiao Zuo, and Jianming Yang. Meta-har: Federated representation learning for human activity recognition. In *Proceedings of the web conference 2021*, pages 912–922, 2021.
- [9] Dongzhou Cheng, Lei Zhang, Can Bu, Xing Wang, Hao Wu, and Aiguo Song. Protohar: Prototype guided personalized federated learning for human activity recognition. *IEEE Journal of Biomedical and Health Informatics*, 27(8), 2023.
- [10] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchu Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Hei Li Kwing, Titouan Parcollet, Pedro PB de Gusmão, and Nicholas D Lane. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020.
- [11] Jie Feng, Can Rong, Funing Sun, Diansheng Guo, and Yong Li. Pmf: A privacy-preserving human mobility prediction framework via federated learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–21, 2020.
- [12] Zhiwen Xiao, Xin Xu, Huanlai Xing, Fuhong Song, Xinhua Wang, and Bowen Zhao. A federated learning system with enhanced feature extraction for human activity recognition. *Knowledge-Based Systems*, 229:107338, 2021.
- [13] Xiaokang Zhou, Wei Liang, Jianhua Ma, Zheng Yan, I Kevin, and Kai Wang. 2d federated learning for personalized human activity recognition in cyber-physical-social systems. *IEEE Transactions on Network Science and Engineering*, 9(6):3934–3944, 2022.
- [14] Pengyuan Zhou, Hengwei Xu, Lik Hang Lee, Pei Fang, and Pan Hui. Are you left out? an efficient and fair federated learning for personalized profiles on wearable devices of inferior networking conditions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–25, 2022.
- [15] Yidong Chai, Haixin Liu, Hongyi Zhu, Yue Pan, Anqi Zhou, Hongyan Liu, Jianwei Liu, and Yang Qian. A profile similarity-based personalized federated learning method for wearable sensor-based human activity recognition. *Information & Management*, 61(7):103922, 2024.
- [16] Can Bu, Lei Zhang, Hengtao Cui, Dongzhou Cheng, Hao Wu, and Aiguo Song. Learn from others and be yourself in federated human activity recognition via attention-based pairwise collaborations. *IEEE Transactions on Instrumentation and Measurement*, 73:1–15, 2024.
- [17] Xiaomin Ouyang, Zhiyuany Xie, Jiayu Zhou, Jianwei Huang, and Guoliang 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*, pages 54–66, 2021.
- [18] Riccardo Presotto, Gabriele Civitarese, and Claudio Bettini. Fedclar: Federated clustering for personalized sensor-based human activity recognition. In *2022 IEEE international conference on pervasive computing and communications (PerCom)*, pages 227–236. IEEE, 2022.
- [19] Linlin Tu, Xiaomin Ouyang, Jiayu Zhou, Yuze He, and Guoliang Xing. Feddl: Federated learning via dynamic layer sharing for human activity recognition. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 15–28, 2021.
- [20] Lifang Shao. Protecting health monitoring privacy in fitness training: A federated learning framework based on personalized differential privacy. *Internet Technology Letters*, 7(6):e499, 2024.
- [21] Ziru Niu, Hai Dong, A Kai Qin, and Tao Gu. Flrc: Resource-efficient federated learning with early-stopping strategy. *IEEE Trans. on Mobile Computing*, 2024.
- [22] Kunjal Panchal, Sunav Choudhary, Subrata Mitra, Koyel Mukherjee, Somdeb Sarkhel, Saayan Mitra, and Hui Guan. Flash: Concept drift adaptation in federated learning. In *International Conference on Machine Learning*, 2023.
- [23] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.
- [24] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.
- [25] Yu Guan and Thomas Plötz. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 1(2):1–28, 2017.
- [26] Yexu Zhou, Haibin Zhao, Yiran Huang, Till Riedel, Michael Hefenbrock, and Michael Beigl. Tinyhar: A lightweight deep learning model designed for human activity recognition. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers*, pages 89–93, 2022.
- [27] Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. Wear: An outdoor sports dataset for wearable and egocentric activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–21, 2024.
- [28] Daniel Geissler, Dominique Nshimiyimana, Vitor Fortes Rey, SungHo Suh, Bo Zhou, and Paul Lukowicz. Beyond confusion: A fine-grained dialectical examination of human activity recognition benchmark datasets. *arXiv:2412.09037*, 2024.
- [29] Majid Kundroo and Taehong Kim. Demystifying impact of key hyper-parameters in federated learning: A case study on cifar-10 and fashionmnist. *IEEE Access*, 2024.
- [30] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.