PRIVACY IN MULTIMODAL FEDERATED HUMAN ACTIVITY RECOGNITION

Alex Iacob [1] **Pedro P. B. Gusmão** [1] **Nicholas D. Lane** [1] **Armand K. Koupai** [2] **Mohammud J. Bocus** [2]
Raúl Santos-Rodríguez [2] **Robert J. Piechocki** [2] **Ryan McConville** [2]

ABSTRACT

Human Activity Recognition (HAR) training data is often privacy-sensitive or held by non-cooperative entities. Federated Learning (FL) addresses such concerns by training ML models on edge clients. This work studies the impact of privacy in federated HAR at a user, environment, and sensor level. We show that the performance of FL for HAR depends on the assumed privacy level of the FL system and primarily upon the colocation of data from different sensors. By avoiding data sharing and assuming privacy at the human or environment level, as prior works have done, the accuracy decreases by 5-7%. However, extending this to the modality level and strictly separating sensor data between multiple clients may decrease the accuracy by 19-42%. As this form of privacy is necessary for the ethical utilisation of passive sensing methods in HAR, we implement a system where clients mutually train both a general FL model and a group-level one per modality. Our evaluation shows that this method leads to only a 7-13% decrease in accuracy, making it possible to build HAR systems with diverse hardware.

1 INTRODUCTION

Human Activity Recognition (HAR) involves classifying human actions (Vrigkas et al., 2015; Jobanputra et al., 2019), such as running or sitting, using data from personal devices like smartphones or environmental sensors. However, practical and legal considerations limit learning from HAR data. For example, using video cameras to simulate virtual bodily-worn movement sensors (Kwon et al., 2020) may generate divergent features from Wi-Fi signals. Furthermore, privacy requirements impose data collection limitations. In this work, privacy requirements refer to constraints on collecting or centralising data at three levels:

User(Human subject)-level Privacy For gyroscope or accelerometer data from smartphones and wearables, end-users may be unwilling to share personal information.

Environment-level Privacy For locations such as hospitals and internment facilities, sensitive information must often remain private from third parties. This constraint may prove challenging as data used for HAR is susceptible to environmental characteristics. For example, a sensor may produce varying features based on object placement or room size.

Modality-level Privacy Data generated from different groups of sensors may be owned by competing entities.

Traditional Machine Learning approaches tackle feature heterogeneity by colocating data and training with Multi-task Learning techniques. However, the privacy constraints above make centralisation unfeasible on a large scale in HAR. Instead, they require a Federated Learning (FL) approach to keep data encapsulated in clients at the necessary privacy level during training. Our work brings the following contributions to Federated Human Activity Recognition:

1. First, we evaluate the performance of multiple models trained in a federated fashion on a multimodal dataset keeping data privately stored on clients at increasing privacy levels. Unlike other works, we investigate the additive effects of privacy up to the complete separation of each user, environment, and modality combination.

2. Second, we show that privacy at the modality level results in the *highest* accuracy cost, followed by the environmental level and then the user level. To mitigate this, we propose mutual learning of group-level models alongside the standard FL model to cover modalities that cannot be colocated in a single client. Our results indicate that this method can significantly reduce accuracy degradation from 19-42% to just 7-13%.

[1]Department of Computer Science and Technology, University of Cambridge, United Kingdom [2]School of Computer Science, Electrical and Electronic Engineering, and Engineering Maths, University of Bristol, United Kingdom. Correspondence to: Alex Iacob <aai30@cam.ac.uk>, Pedro Porto Buarque de Gusmao <pp524@cam.ac.uk>, Nicholas Donald Lane <ndl32@cam.ac.uk>, Armand K. Koupai <uw20504@bristol.ac.uk>, Mohammud J. Bocus <junaid.bocus@bristol.ac.uk>, Raúl Santos-Rodríguez <enrsr@bristol.ac.uk>, Robert J. Piechocki <eerjp@bristol.ac.uk>, Ryan McConville <ryan.mcconville@bristol.ac.uk>.

2 Multimodality in Federated Human Activity Recognition

Federated Learning, proposed by McMahan et al. (2017), trains ML models from distributed data on edge devices using efficient communication techniques for maintaining privacy. Although successful in training models from diverse users, such as keyboard prediction (Hard et al., 2018), and diverse hardware, such as medical applications (Sheller et al., 2020), data heterogeneity remains a significant challenge (Kairouz et al., 2021, sec 3.1). Due to privacy constraints, approaches like Multi-task Learning and Continual Learning, which handle feature heterogeneity, are limited in a Federated Learning context. For instance, Elastic Weight Consolidation (Kirkpatrick et al., 2017) estimates parameter variance using past data, while Learning Without Forgetting (Li & Hoiem, 2018) stores network outputs from past tasks.

Previous work on Federated Human Activity Recognition has not fully explored the heterogeneity emerging from independent data collection systems. For instance, Sozinov et al. (2018) considers skewed label distributions and noise across smartphone users while *colocating* gyroscope and accelerometer data. Similar partitioning schemes are investigated for feature extraction (Xiao et al., 2021) and clustering methods (Ouyang et al., 2021). Furthermore, such works may use artificially partitioned centralised datasets, as in Zhao et al. (2020), or contain only one modality, as in some datasets collected by Ouyang et al. (2021). To create adaptable HAR systems that can accommodate new clients with different sensor types in the federation, investigating Federated HAR with modalities split across clients is necessary, given the shifting hardware landscape of HAR sensors.

3 Federating Human Activity Recognition

We construct multiple partitions of the OPERAnet dataset published by Bocus et al. (2022) to assess Federated Human Activity Recognition under privacy at user, environment, and modality privacy levels. The dataset contains five different sensors; however, Bocus et al. (2022) indicate that only Channel State Information (CSI) from a Network Card Interface (NIC) and Passive Wi-Fi Radar (PWR) should be used for HAR. The data were collected synchronously, with the multiple channels—three for CSI and two for PWR—of RF data. They cover eight hours of surveying six participants performing six activities spread across two rooms. Because room activity distribution is non-uniform, separating clients by environment also provides skewness at the label level.

We transform the time-series data into image data in keeping with the original HAR preprocessing applied by Bocus et al. (2022) and previous works (Bocus et al., 2021; Li et al., 2020; 2022). We further increase the dataset's size and modality diversity by reusing the pipeline of Koupai et al. (2022). Based on the underlying CSI and PWR data, Koupai et al. (2022) construct spectrograms of the CSI and PWR data. The complete image set contains five data views for each underlying CSI or PWR channel. We use the three most effective view types reported by Koupai et al. (2022). Since different channels for CSI and PWR are physically colocated on the device, it is assumed that fusing images generated from separate channels would not be a violation of privacy at the sensor level. Consequently, the complete image types we shall refer to as modalities for the rest of this work contain concatenated images generated from each source channel. One such image type comes from CSI; two come from PWR.

The work of Koupai et al. (2022) offers two centralised baselines to compare against, a ResNet34 (He et al., 2016) model used for HAR and a Fusion Vision Transformer (FViT). While CNNs have been successfully applied to HAR by Bevilacqua et al. (2018); Ronald et al. (2021) and Tang et al. (2023), the novel FViT addresses the issue of multimodal HAR by adapting the Vision Transformer (ViT) architecture developed by Dosovitskiy et al. (2021) to operate over *fused* images. Crucially for our experiments, FViT has a parameter count invariant to the number of images combined, making the network capacity equivalent between fused and unfused modalities. In addition to the transformer and ResNet34, we use the smallest EfficientNetV2 constructed by Tan & Le (2021) as the communication costs and compute concerns in FL make the smaller network a practical choice.

3.1 Partitioning by Privacy Level

The partitions we construct correspond to increasing privacy levels. For example, splitting by human subject implies that each client in that partition only contains data corresponding to one human participant and thus obeys *Subject(User)-level* privacy. Likewise, the partition splitting by participant and room implies that each participant and room combination is treated as a separate client and offers both *Subject(User)-level* and *Environment-level* privacy. The most heterogeneous partition we create treats each participant, room, and modality combination as one client and offers the previous two levels of privacy together with *Modality-level* privacy.

To create a meaningful test set for Federated HAR, we use the data of the sixth client. Since OPERAnet has not been used for FL before, our evaluation compares the accuracy of FL partitioned by subject and environment to the State of The Art centralised baselines using colocated fused modalities. Following this initial investigation, we explore privacy interactions at a subject (user), environmental and modality level when the modalities are never fused and not necessarily colocated. The *separated-modality* experiments are the ones we use to report findings, as they can cover all levels of

Table 1. The partitions in our experimental setup. They include the centralised baseline (*Centralised*), those partitioned by human subject (*Subj*), by subject and environment (*Subj+Env*), or by subject, environment, and modality (*Subj+Env+Mod*). A partition can contain fused (*F*) or separated (*S*) modalities.

Partition	Avg Samples	#Samples Subj 6	#Train Clients	#Clients/R
Centralised (F)	1947.0 ± 0.0	463	1	1
Subj (F)	324.3 ± 32.3	463	6	2
Subj + Env (F)	194.6 ± 194.6	463	10	3
Centralised (S)	5841.0 ± 0.0	1389	1	1
Subj (S)	973.0 ± 973.0	1389	6	2
Subj + Env (S)	583.8 ± 583.8	1389	10	3
Subj + Env + Mod (S)	194.6 ± 194.6	1389	30	9

Table 2. Accuracy results (mean and standard deviation) for model and partition combinations on the test set of OPERAnet. Note the impact of partitioning by modality compared to the subject or environment and the smoother decline in the performance of FViT compared to the CNNs. The "Ensemble" uses the three group models to predict the data label belonging to their modality. The results of F1-Score, shown in Table 3, follow the same trend.

Partition	FViT	ResNet34	EffNetB0	Ensemble
Centralised (Fused)	0.90±0.01	**0.93±0.01**	0.91±0.01	-
Subj (Fused)	0.85±0.02	**0.90±0.02**	0.88±0.01	-
Subj+Env (Fused)	0.83±0.03	**0.84±0.07**	0.79±0.04	-
Centralised (S)	0.83±0.01	**0.89±0.02**	0.87±0.01	-
Subj (S)	0.81±0.01	0.84±0.03	**0.85±0.01**	-
Subj+Env (S)	0.78±0.02	**0.84±0.02**	0.80±0.01	-
Subj+Env+Mod (S)	**0.64±0.04**	0.47±0.03	0.55±0.05	**0.76±0.02**

privacy. Table 1 presents the constructed partitions and their statistics. As we intended to use 30% of total clients each round, we split the data of one participant into two clients based on their room when federating by subject. For the centralised baseline, we follow Koupai et al. (2022) and train for 100 local epochs, while FL trains for 10 rounds with 10 local epochs. The optimiser parameters are kept constant and at parity with Koupai et al. (2022)—see Appendix A.

3.2 Mutual Global and Group Model Learning

To handle separating modalities across clients in a federated network, we propose a group FL structure utilising Deep Mutual Learning (Zhang et al., 2018). Two models are trained on each client and distil knowledge into each other. One model is a globally federated model trained on all clients. The other is a group-level one trained only on clients with a specific modality. The server maintains one model per modality group, providing flexibility for integrating new sensors. We chose the FViT as the global federated model because of its resilience to high heterogeneity in previous experiments. Furthermore, we use the small EfficientNetV2 as the group-level model for future scalability. We present the performance of an ensemble of group-level models, each predicting the relevant modality. Hyperparameters were optimised via Bayesian search, resulting in global and group-level distillation weights of 0.33 and 0.75, respectively.

4 Evaluation

Our evaluation reveals that Federated Human Activity Recognition is sensitive to sensor heterogeneity but partially resilient to subject characteristics and room structure. In Table 2, we present accuracy results of all partition and model combinations using fused (F) or separated (S) modalities. Fused modality experiments establish a baseline of comparison between federated and centralised training on OPERAnet. Separated modality experiments allow us to investigate more granular levels of privacy and will provide most of the notable figures. Accuracy convergence curves are available per model in Fig. 1; however, they only show unfused modalities to emphasise results for modality-level privacy. In the appendix, convergence curves for fused modalities are presented in Fig. 2 and follow similar trends. All experiments used the Flower (Beutel et al., 2020) FL framework.

4.1 Subject(User)-level Privacy

Experiments with user-level privacy showed that all three models achieved results within 3-5% of the centralised baseline, regardless of modality fusion. As shown in Fig. 1, this small gap to the centralised unfused baseline was consistently observed across rounds. In addition, a study by Sozinov et al. (2018) found a similar 4-6% accuracy gap when treating each person as a separate partition, indicating that body characteristics and slight movement differences are not significant enough to generate highly divergent features. These findings suggest that FL is a practical solution for accessing extensive private data from end users. However, unfused modalities reduced accuracy for centralised and subject-level partitions nearly uniformly compared to fused ones—showcasing the benefits of centralisation.

4.2 Environment-level Privacy

A further slight-to-medium accuracy degradation is perceptible in the experimental partitions applying subject-level and environmental privacy in Table 2. It is worth noting that data from OPERAnet may have reduced environmental heterogeneity as the same hardware, procedure, and subjects were used in a controlled setting. However, this is a common issue for all HAR systems (Vrigkas et al., 2015, sec.6). Fused modalities saw an additional drop in accuracy of 2-9%, while unfused modalities saw a less significant impact, with the additional maximum drop never exceeding 5%. Notably, ResNet34 operating on unfused modalities did not exhibit a significant accuracy drop when privacy was increased. Fused modalities contain more information about the environment per sample, aiding in distinguishing between rooms. However, the additional information becomes

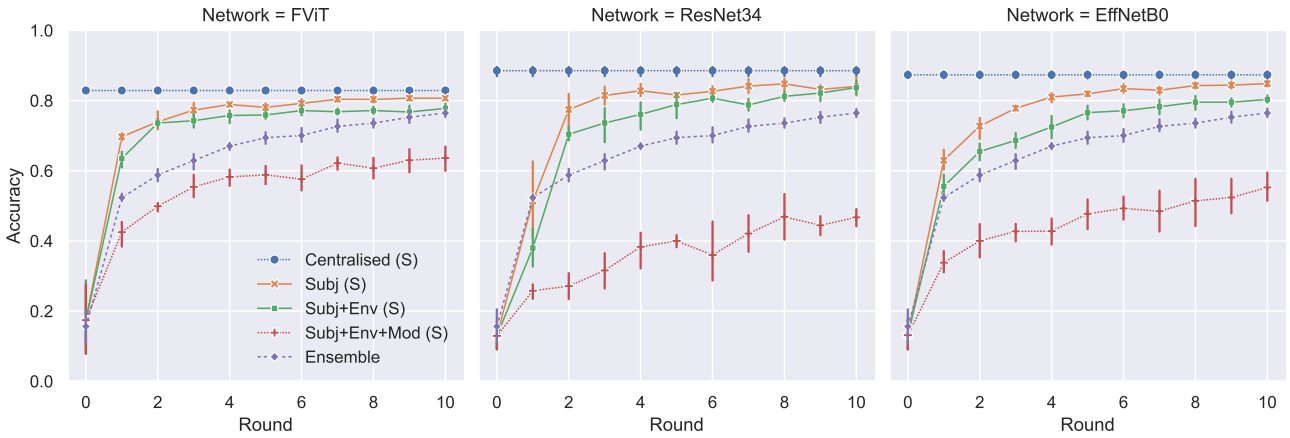

Figure 1. Model per-round accuracy on the sixth subject's dataset at all privacy levels. For low-heterogeneity, CNNs generally outperform FViT, but FViT gains a significant advantage when modality partitioning is introduced. Notably, FViT converges faster in the initial rounds and reaches higher accuracy with fewer data in the first two rounds. The "Ensemble" results reverse the effects of modality-level privacy and outperform all non-grouped federated models.

less valuable after samples are split into clients. Figure 1 highlights that the small EfficientNetV2 suffers more from not having examples from both rooms available.

4.3 Modality-level Privacy

The experimental results reveal an unexpected pattern when applying privacy at the modality level. The previously top-performing ResNet34 experiences a 37% further drop in accuracy with a 42% total, while the EfficientNetV2 suffers a 25% further drop in accuracy with a 32% total. By contrast, the Fusion Vision Transformer (FViT), which produced worse results in previous experiments, only experiences a 14% further drop in accuracy with 19% total and emerges as the model with the most significant performance advantage for a given privacy level. To better understand this outcome and the interplay between different model types, we turn to the plot in Fig. 1, which shows the convergence of models for different partitions. The plot immediately reveals the steeper slope of improvement that FViT obtains in the first few rounds. Furthermore, this pattern of performance aligns with the fine-tuning experiments reported in (Koupai et al., 2022), where FViT outperformed ResNet when both were trained on a small amount (1-20%) of data.

The increasing prevalence of IoT devices, surveillance cameras, personal smartphones, and passive RF sensors has led to extensive human activity recognition (HAR) data collection. However, with no uniform regulation or competitive environment, it is critical to prioritise privacy preservation and address the afferent accuracy degradation.

4.4 Mutual Learning with Per-modality Group Models

We evaluate the effectiveness of our ensemble, which uses mutual learning to handle the challenges of federated learning across modalities. As demonstrated in both Fig. 1 and Table 2, the ensemble achieves near-equivalent accuracy to FViT on the "Subj+Env" partition with colocated modalities. However, training the federated learning and group-level models simultaneously is costly and difficult to optimise. Moreover, our hyperparameter search, which explored 79 combinations of distillation weights, revealed that the ensemble's performance is sensitive to hyperparameter changes. Meanwhile, the federated model failed to surpass the "Subj+Env+Mod(S)" result in Table 2 through mutual learning, primarily due to the inherent difficulty of multimodal training on the same network without employing specific multi-task techniques.

5 Conclusion

We investigated the performance of Multimodal Federated Human Activity Recognition under privacy levels that may arise in practice, such as the subject(user), environmental, and modality levels. Our results show that performance degrades with each additional privacy layer starting with 5-7% for the subject and environmental levels. Remarkably, we observed an *overall* accuracy drop of 32-42% for CNNs when modality-level privacy is assumed. Nevertheless, our experiments determined that a Fusion Vision Transformer architecture performs well in extreme scenarios. Its fast initial convergence with few samples led to an *additional* drop of only 14% with a 19% *overall* drop for modality-level privacy. Furthermore, constructing small group-level models

for each modality type trained in a mutual-learning fashion with a global one can limit the *overall* degradation to 7-13%. Such a system can adjust to shifting hardware conditions by incorporating new group-level models and utilising the global model's knowledge for bootstrapping. Despite the clear trends, this work is limited by the size of OPERAnet. Besides larger datasets, other potential future research avenues include hierarchical FL with layered aggregation and creating sparse models with task-based subnetworks.

REFERENCES

Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., and Lane, N. D. Flower: A friendly federated learning research framework. *CoRR*, abs/2007.14390, 2020. URL https://arxiv.org/abs/2007.14390.

Bevilacqua, A., MacDonald, K., Rangarej, A., Widjaya, V., Caulfield, B., and Kechadi, M. T. Human activity recognition with convolutional neural networks. In Brefeld, U., Curry, E., Daly, E., MacNamee, B., Marascu, A., Pinelli, F., Berlingerio, M., and Hurley, N. (eds.), *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part III*, volume 11053 of *Lecture Notes in Computer Science*, pp. 541–552. Springer, 2018. doi: 10.1007/978-3-030-10997-4_33. URL https://doi.org/10.1007/978-3-030-10997-4_33.

Bocus, M. J., Li, W., Paulavicius, J., McConville, R., Santos-Rodriguez, R., Chetty, K., and Piechocki, R. Translation resilient opportunistic wifi sensing. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 5627–5633, 2021. doi: 10.1109/ICPR48806.2021.9412263.

Bocus, M. J., Li, W., Vishwakarma, S., Kou, R., Tang, C., Woodbridge, K., Craddock, I., McConville, R., Santos-Rodriguez, R., Chetty, K., and Piechocki, R. Operanet, a multimodal activity recognition dataset acquired from radio frequency and vision-based sensors. *Scientific Data*, 9(1):474, 2022. doi: 10.1038/s41597-022-01573-2. URL https://doi.org/10.1038/s41597-022-01573-2.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Hard, A., Rao, K., Mathews, R., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. Federated learning for mobile keyboard prediction. *CoRR*, abs/1811.03604, 2018. URL http://arxiv.org/abs/1811.03604.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Jobanputra, C., Bavishi, J., and Doshi, N. Human activity recognition: A survey. *Procedia Computer Science*, 155:698–703, 2019. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2019.08.100. URL https://www.sciencedirect.com/science/article/pii/S1877050919310166. The 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2019),The 14th International Conference on Future Networks and Communications (FNC-2019),The 9th International Conference on Sustainable Energy Information Technology.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K. A., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2):1–210, 2021. doi: 10.1561/2200000083. URL https://doi.org/10.1561/2200000083.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1611835114.

Koupai, A. K., Bocus, M. J., Santos-Rodriguez, R., Piechocki, R. J., and McConville, R. Self-supervised

multimodal fusion transformer for passive activity recognition. *IET Wireless Sensor Systems*, 12(5-6):149–160, 2022. doi: https://doi.org/10.1049/wss2.12044. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/wss2.12044.

Kwon, H., Tong, C., Haresamudram, H., Gao, Y., Abowd, G. D., Lane, N. D., and Ploetz, T. Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition, 2020. URL https://arxiv.org/abs/2006.05675.

Li, W., Bocus, M. J., Tang, C., Vishwakarma, S., Piechocki, R. J., Woodbridge, K., and Chetty, K. A taxonomy of wifi sensing: Csi vs passive wifi radar. In *2020 IEEE Globecom Workshops (GC Wkshps*, pp. 1–6, 2020. doi: 10.1109/GCWkshps50303.2020.9367546.

Li, W., Bocus, M. J., Tang, C., Piechocki, R. J., Woodbridge, K., and Chetty, K. On csi and passive wi-fi radar for opportunistic physical activity recognition. *IEEE Transactions on Wireless Communications*, 21(1):607–620, 2022. doi: 10.1109/TWC.2021.3098526.

Li, Z. and Hoiem, D. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, 2018. doi: 10.1109/TPAMI.2017.2773081. URL https://doi.org/10.1109/TPAMI.2017.2773081.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In Singh, A. and Zhu, X. J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL http://proceedings.mlr.press/v54/mcmahan17a.html.

Ouyang, X., Xie, Z., Zhou, J., Huang, J., and Xing, G. Clusterfl: a similarity-aware federated learning system for human activity recognition. In Banerjee, S., Mottola, L., and Zhou, X. (eds.), *MobiSys '21: The 19th Annual International Conference on Mobile Systems, Applications, and Services, Virtual Event, Wisconsin, USA, 24 June - 2 July, 2021*, pp. 54–66. ACM, 2021. doi: 10.1145/3458864.3467681. URL https://doi.org/10.1145/3458864.3467681.

Ronald, M., Poulose, A., and Han, D. S. isplinception: An inception-resnet deep learning architecture for human activity recognition. *IEEE Access*, 9:68985–69001, 2021. doi: 10.1109/ACCESS.2021.3078184.

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., and Bakas, S. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1):12598, 2020. doi: 10.1038/s41598-020-69250-1. URL https://doi.org/10.1038/s41598-020-69250-1.

Sozinov, K., Vlassov, V., and Girdzijauskas, S. Human activity recognition using federated learning. In Chen, J. and Yang, L. T. (eds.), *IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications, ISPA/IUCC/BDCloud/SocialCom/SustainCom 2018, Melbourne, Australia, December 11-13, 2018*, pp. 1103–1111. IEEE, 2018. doi: 10.1109/BDCloud.2018.00164. URL https://doi.org/10.1109/BDCloud.2018.00164.

Tan, M. and Le, Q. V. Efficientnetv2: Smaller models and faster training. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10096–10106. PMLR, 2021. URL http://proceedings.mlr.press/v139/tan21a.html.

Tang, Y., Zhang, L., Min, F., and He, J. Multiscale deep feature learning for human activity recognition using wearable sensors. *IEEE Transactions on Industrial Electronics*, 70(2):2106–2116, 2023. doi: 10.1109/TIE.2022.3161812.

Vrigkas, M., Nikou, C., and Kakadiaris, I. A. A review of human activity recognition methods. *Frontiers Robotics AI*, 2:28, 2015. doi: 10.3389/frobt.2015.00028. URL https://doi.org/10.3389/frobt.2015.00028.

Xiao, Z., Xu, X., Xing, H., Song, F., Wang, X., and Zhao, B. A federated learning system with enhanced feature extraction for human activity recognition. *Knowl. Based Syst.*, 229:107338, 2021. doi: 10.1016/j.knosys.2021.107338. URL https://doi.org/10.1016/j.knosys.2021.107338.

Zhang, Y., Xiang, T., Hospedales, T. M., and Lu, H. Deep mutual learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4320–4328. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00454. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_Deep_Mutual_Learning_CVPR_2018_paper.html.

Zhao, Y., Liu, H., Li, H., Barnaghi, P. M., and Haddadi, H. Semi-supervised federated learning for activity recognition. *CoRR*, abs/2011.00851, 2020. URL https://arxiv.org/abs/2011.00851.

A APPENDIX

Table 3. The F1-Score results of partition-model combinations. The same trends from the accuracy comparisons repeat themselves with higher privacy requirements leading to worse performance. A similar strong decline can be observed when clients are partitioned by modality, with the FViT performing the best in the most heterogeneous condition despite trailing behind the CNNs for all other partitions. The ensemble group models also successfully recovered performance near the FViT levels when the partitioning was based only on subject and environment.

Partition	FViT	ResNet34	EffNetB0	Ensemble
Centralised (Fused)	0.80± 0.03	**0.86 ± 0.02**	0.83± 0.02	-
Subj (Fused)	0.73± 0.03	**0.82 ± 0.05**	0.78± 0.02	-
Subj+Env (Fused)	0.70± 0.05	**0.74 ± 0.08**	0.64± 0.04	-
Centralised (Split)	0.71± 0.01	**0.80± 0.03**	0.78± 0.02	-
Subj (Split)	0.68± 0.01	0.72± 0.06	**0.73± 0.02**	-
Subj+Env (Split)	0.64± 0.03	**0.71± 0.04**	0.67± 0.02	-
Subj+Env+Mod (Split)	**0.50± 0.04**	0.35± 0.03	0.39± 0.04	**0.60± 0.03**

The preprocessing pipeline we use is precisely described in Koupai et al. (2022); however, we shall offer a brief summary here. First, the CSI signal is denoised using a discrete wavelet transform and median filtering before applying PCA and generating a spectrogram through the STFT. Then, for the PWR data, the authors apply the cross ambiguity function to the PWR data followed by the CLEAN algorithm and the outputting of a Doppler spectrogram. We use three of the image types they generate. First, we use the concatenated spectrograms generated from the three-receiver surveillance channels of the PWR data. The combined images from the three channels have a dimension of 224×672. Second, we use the spectrograms generated using STFT on amplitude CSI data from two receivers with a concatenated size of 224×448. Third, we use the phase-difference spectrograms generated via STFT from the phase-difference CSI data from the two receivers with a concatenated size of 224×448. Combined in the fused partitions, they add up to 224×1568 images. Finally, we take the largest image type (224×672) in unfused partitions and pad the rest.

Client data partitions are generated in order of person index for split-subject modalities, person index and then room index for subject and environment, and subject, room, and modality index for the final partitioning. Our indexing assumes the human subjects are ordered from one to six, the rooms from one to two, and the modalities from one to three in the above order. The subject and room indexes are directly available in the dataset. Each model and partition combination was run using five distinct seeds generating the same client sequence across models. Thus differences in performance between models are not due to randomness in client selection. The seeds we use are $42, 1337, 3407, 8711, 9370$, and the client sequence is generated by calling `np.random.choice` for the given number of clients per round out of the entire population for each of the ten rounds at the start of the script right after the seed has been set. The mean and standard deviation are reported based on the five seeds in and Table 2 and Table 3. The per-round values in Fig. 1 and Fig. 2 have their mean and standard deviation calculated based on the accuracy of the models on each of the five seeds at the specific round. Before every experiment, the same seeds are used to set the random, NumPy, and Torch modules in Python.

All models have been trained as in Koupai et al. (2022) using AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with a weight decay of 0.01 and batch size of 10 rather than 64 due to the small size of the federated partitions. The computational resources involved four Nvidia A40s and were extensively used during parameter tuning.

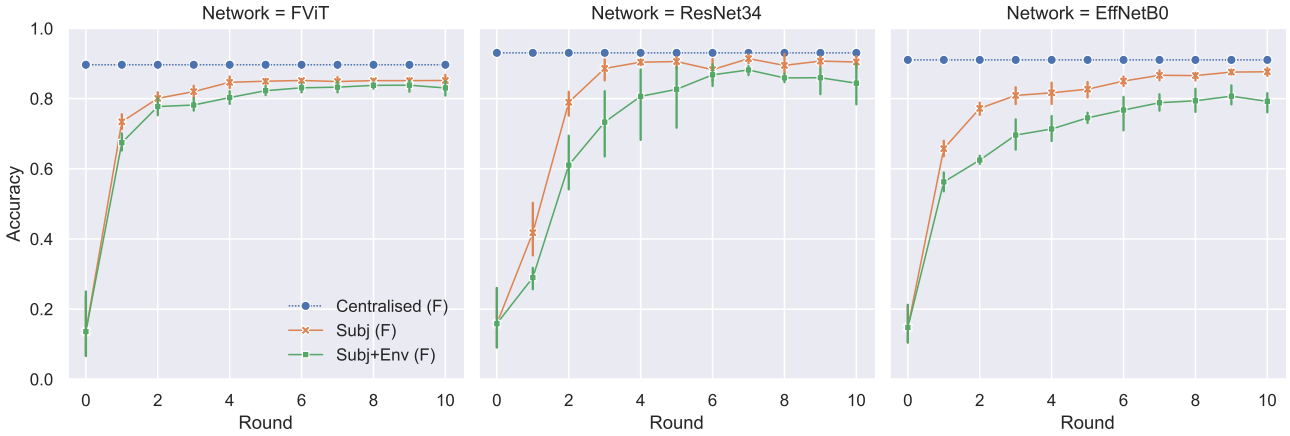

Figure 2. Model per-round accuracy on the fused-modality dataset. The trends observed resemble those for the split partitions with only a uniform decrease in accuracy by comparison. The only major change in results is the sensitivity of ResNet34 to environmental privacy.