ROBUST AND PRIVATE MULTIMODAL FEDERATED HUMAN ACTIVITY RECOGNITION

Alex Iacob [1] **Pedro P. B. Gusmão** [1] **Nicholas Donald Lane** [1]

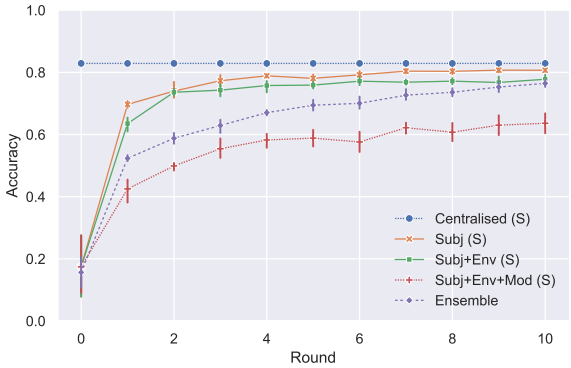

Figure 1. Accuracy of the model at increasing privacy levels. Ensemble refers to the accuracy of group-level models trained/tested on their afferent modality on the "Subj+Env+Mod" partition.

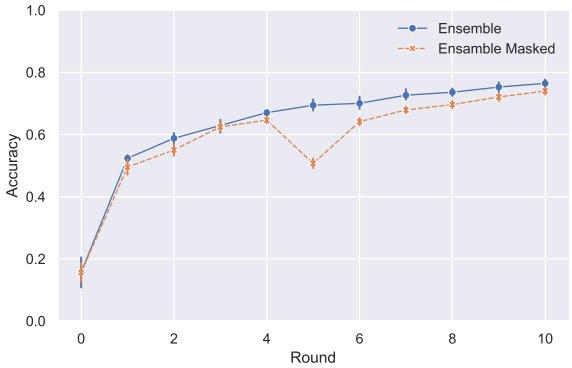

Figure 2. Default performance of the ensemble vs the average performance we obtain when successively masking clients from each of the three modalities up to the 5th round.

Human Activity Recognition (HAR) involves classifying human actions, such as running or sitting, using data from personal devices like smartphones or environmental sensors. However, privacy requirements impose data collection limitations. In this work, privacy requirements refer to constraints on collecting or centralising data at three levels:

User-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.

Modality-level Privacy Data generated from different groups of sensors may be owned by competing entities or raise ethical concerns if collected in public spaces.

A Federated Learning (FL) approach keeps 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 a Multimodal Vision Transformer [2] trained in a federated fashion on the recent multimodal OPERAnet [1] dataset. 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 smaller group-level models (EfficientNetV2B0) alongside the FL model to cover modalities that cannot be colocated in a single client. Our results in Fig. 1 indicate that this method can significantly reduce accuracy degradation. Furthermore, Fig. 2 shows that our approach is resilient to adding new modalities during training.

[1]University of Cambridge: {aai30,pp524,ndl32}@cam.ac.uk.

REFERENCES

[1] M. J. Bocus, W. Li, S. Vishwakarma, R. Kou, C. Tang, K. Woodbridge, I. Craddock, R. McConville, R. Santos-Rodriguez, K. Chetty, and R. Piechocki. 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.

[2] A. K. Koupai, M. J. Bocus, R. Santos-Rodriguez, R. J. Piechocki, and R. McConville. 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.