FedMFS: Federated Multimodal Fusion Learning with Selective Modality Communication

Liangqi Yuan, Dong-Jun Han, Vishnu Pandi Chellapandi, Stanislaw H. Żak, and Christopher G. Brinton

Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, 47907, USA
Email: {liangqiy, han762, cvp, zak, cgb}@purdue.edu

***Abstract*—Federated learning (FL) is a distributed machine learning (ML) paradigm that enables clients to collaborate without accessing, infringing upon, or leaking original user data by sharing only model parameters. In the Internet of Things (IoT), edge devices are increasingly leveraging multimodal data compositions and fusion paradigms to enhance model performance. However, in FL applications, two main challenges remain open: (i) addressing the issues caused by heterogeneous clients lacking specific modalities and (ii) devising an optimal modality upload strategy to minimize communication overhead while maximizing learning performance. In this paper, we propose Federated Multimodal Fusion learning with Selective modality communication (FedMFS), a new multimodal fusion FL methodology that can tackle the above mentioned challenges. The key idea is to utilize Shapley values to quantify each modality's contribution and modality model size to gauge communication overhead, so that each client can selectively upload the modality models to the server for aggregation. This enables FedMFS to flexibly balance performance against communication costs, depending on resource constraints and applications. Experiments on real-world multimodal datasets demonstrate the effectiveness of FedMFS, achieving comparable accuracy while reducing communication overhead by one twentieth compared to baselines.**

I. Introduction

Federated learning (FL) is a collaborative machine learning (ML) approach that provides users with significant privacy protection by exclusively sharing model parameters without granting access to user raw data, infringing upon the data, or leaking user raw data [1]. The conventional FL paradigm involves clients training ML models on local data, then uploading them to a central server for aggregation, and finally downloading and deploying the aggregated model for applications [2]. As edge devices in the Internet of Things (IoT), such as smartphones, robots, unmanned aerial vehicles (UAVs), and vehicles, are equipped with multimodal sensors and constrained resources, there has been an increasing interest in multimodal fusion federated learning (MFFL) frameworks. Examples of MFFL are various connected and automated vehicles (CAVs) with sensors such as cameras, LiDAR, and Radar [3], [4]. These multimodal sensors enable optimal control decisions in various driving scenarios, varying weather conditions, and restricted field of view. At the same time, CAVs also rely on FL to collaboratively learn between vehicles [5].

Relevant Literature. A variety of fusion algorithms have been proposed to improve the performance of MFFL have been proposed. Qi *et al.* [6] proposed a data-level fusion FL system tailored for the combination of wearable sensor signals and images for fall detection. Xiong *et al.* [7] implemented a feature-level fusion approach with attention modules. Feng *et al.* [8] presented two decision-layer fusion strategies using concatenation and attention modules, respectively, to address three types of client heterogeneities: modality absence, label absence, and label errors. Salehi *et al.* [9] incorporated the MFFL framework into CAVs and conducted real-world experiments. In their FLASH framework, clients randomly select and upload one of the three modality models or an ensemble model for aggregation. Chen *et al.* [10] integrated MFFL into decentralized FL [11], [12], aiming to facilitate collaborative training within client networks without server support.

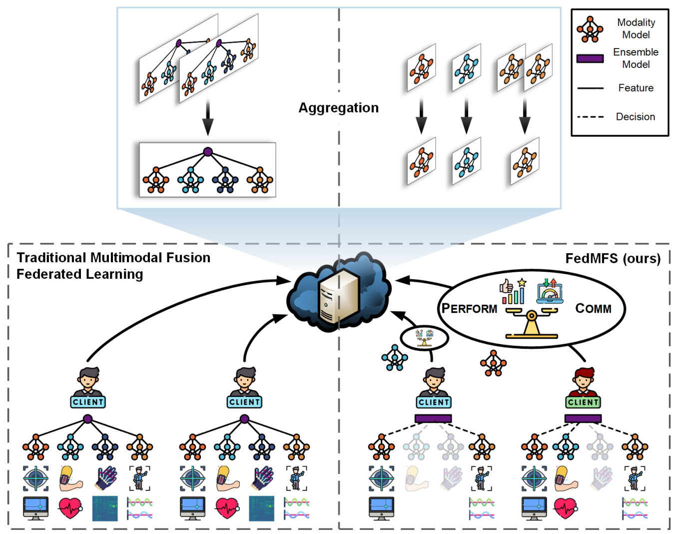

Fig. 1. Schematic representation of traditional multimodal fusion federated learning vs. the proposed FedMFS.

Problem Statement. In the context of the IoT, where communication constraints are particularly pronounced due to bandwidth limitations and diverse device capabilities, most existing MFFL frameworks emphasize massive parallel processing of extensive sensor data. They utilize multimodal fusion to boost model performance, especially in scenarios where heterogeneous clients lack certain modalities. However, there is still a need to devise strategies to reduce communication overhead and improve learning efficiency within the MFFL paradigm. In this paper, we aim to answer the following key question: ***In resource-constrained and heterogeneous MFFL settings, how should each client evaluate and select the best set of modalities to trade-off between performance and communication?***

Solution. As illustrated in Fig. 1, we propose a decision-level fusion approach, where predictions from *modality models* are used as inputs to the *ensemble model*. This allows the independent deployment of the modality models in various application scenarios, accommodating situations where *heterogeneous clients might lack specific modalities*. Furthermore, we consider classical ML models for our ensemble instead of the complex neural networks typically used in traditional MFFL, resulting in reduced computational overhead and improved interpretability. Lastly, considering communication constraints, clients do not always have the ability to upload all modality models to the server for aggregation. Therefore, we emphasize the need to trade-off between performance and communication to ensure accuracy, reduce communication overhead, and enhance learning efficiency. To measure the impact of each modality, we propose the use of the Shapley value [13]–[15] on the ensemble model to quantify their respective performances. Recent studies using Shapley values have demonstrated that, in data/sensor fusion, different input features have varying impacts on the model output [16]–[19]. Although many communication-efficient FL frameworks have been developed [20]–[22], our focus is on specific MFFL scenarios where further reduction in communication overhead is achieved through selective modality communication.

Contributions. This paper is concerned with a Federated Multimodal Fusion learning with Selective modality communication (FedMFS) framework, as illustrated in Fig. 1. The paper's contributions are:

- We introduce FedMFS, an interpretable, flexible, detachable, and modular MFFL framework, tailored for heterogeneous clients with absent modalities. A Shapley component incorporated within FedMFS enables the interpretation of the contribution of individual modalities in the final prediction.
- We provide a series of configurations to balance between the performance of the modality model and the communication overhead. The proposed FedMFS, through the adjustable configuration of three customizable parameters, allows users to minimize communication costs while ensuring optimal learning efficiency, which depends on the available communication resources.
- We evaluate the performance and communication overhead of our proposed FedMFS against four baseline methods on a real-world heterogeneous multimodal dataset. Experimental results demonstrate the superior performance of FedMFS, which, while achieving comparable accuracy, incurs only one twentieth of the communication overhead compared to the baselines.

II. Methodologies

A. Framework Formulation

We assume that there are K clients participating in the FL framework. Each client, denoted k, possesses a dataset $\mathbb{D}^k$ and a label set Y^k, comprising multimodal data represented as

$$\mathbb{D}^k = \{\mathcal{D}_1^k, \mathcal{D}_2^k, \ldots, \mathcal{D}_{M_k}^k\}, \tag{1}$$

where $\mathcal{D}_m^k$ denotes datasets corresponding to m modality, such as images, LiDAR, RF, etc., with $m = 1, 2, \ldots, M_k$ indicating specific data modality. We note that the heterogeneous client k can accommodate a different number of data modalities, represented by M_k. Each client has a *modality model*, θ_m^k, for every modality dataset $\mathcal{D}_m^k$, which is designed to capture the relationship between the input data and its corresponding labels. Therefore, each client possesses a set of models, represented as

$$\Theta^k = \{\theta_1^k, \theta_2^k, \ldots, \theta_{M_k}^k\}. \tag{2}$$

Each model, when trained on the dataset, yields a predicted label, denoted as

$$\hat{y}_m^k = \theta_m^k(\mathcal{D}_m^k), \tag{3}$$
$$\widehat{\mathbb{Y}}^k = \{\hat{y}_1^k, \hat{y}_2^k, \ldots, \hat{y}_{M_k}^k\}, \tag{4}$$

where $\hat{y}_m^k$ represents the predicted label generated from the model θ_m^k for the dataset $\mathcal{D}_m^k$, and $\widehat{\mathbb{Y}}^k$ is the collection of all the predicted labels for client k. Our goal is to fuse the outputs of all modality models at the decision layer through a post-processing *ensemble model* ω^k, represented as

$$\begin{aligned}\widehat{Y}^k &= \omega^k\left(\theta_1^k(\mathcal{D}_1^k), \theta_2^k(\mathcal{D}_2^k), \ldots, \theta_{M_k}^k(\mathcal{D}_{M_k}^k)\right) \\ &= \omega^k\left(\hat{y}_1^k, \hat{y}_2^k, \ldots, \hat{y}_{M_k}^k\right) \\ &= \omega^k\left(\widehat{\mathbb{Y}}^k\right),\end{aligned} \tag{5}$$

where $\widehat{Y}^k$ denotes the predicted label set corresponding to the dataset $\mathbb{D}^k$.

B. Proposed FedMFS Algorithm

The proposed FedMFS is described in Algorithm 1. The primary objective of FedMFS is the collaborative learning of the modality model θ_m for each modality m. For each client, individual and system heterogeneities (e.g., modality missing, noise, device errors, device malfunctions, etc.) are addressed through a personalized ensemble model ω^k. Furthermore, an intrinsic objective of FedMFS is to compensate for the communication constraints inherent in IoT edge devices by minimizing communication overhead, thus ensuring efficient learning efficacy for clients within the FL framework.

C. Client Learning

For each data modality for every client, the local objective is to minimize the difference between the predicted and the true labels. This objective can be achieved using various optimizers, such as stochastic gradient descent (SGD). Formally, we describe the optimization problem as $\min_{\theta_m^k} \mathcal{L}(Y^k, \hat{y}_m^k)$,

Algorithm 1 FedMFS: Federated Multimodal Fusion Learning with Selective Modality Communication

Input: Communication rounds (T), clients' dataset ($\mathbb{D}^k$), clients' label set (Y^k), local training epoch (E), initial models ($\theta^k_{m,0}$ & ω^k_0), loss function ($\mathcal{L}$), learning rate (η), modality model upload count (γ), performance and communication weights (α_s & α_c)
Output: Generalized global modality models (θ_m) and personalized local ensemble models for each client (ω^k)

 # Global Iteration

for $t = 0$ **to** $T-1$ **do**

  # Local Learning

  for each client k **in parallel do**

    for each data modality m **do**

      Backpropagate the loss function and update the modality models $\theta^{k,t}_m \leftarrow \arg\min_{\theta^{k,t}_m} \mathcal{L}(Y^{k,t}, \hat{y}^{k,t}_m)$.

      (Stage #1) Update the ensemble model $\omega^{k,t} \leftarrow \arg\min_{\omega^{k,t}} \mathcal{L}(Y^{k,t}, \widehat{Y}^{k,t})$.

    end for

  end for

  # Performance-Communication Trade-Off

  Compute the impact of each modality $\Phi^{k,t}$ on the prediction using the Shapley values. ▷ (6), (7)

  Compute the modality model size $\bar{\Theta}^k$. ▷ (8)

  Normalize and select top-γ priority modality models for uploading. ▷ (9), (10), (11)

  # Server Aggregation

  for each data modality m **do**

    Server calculates the weight $\beta^{k,t}_m$ for each client k and aggregates modality model θ^t_m. ▷ (13), (14)

  end for

  # Local Deploying

  for each client k **in parallel do**

    for each data modality m **do**

      Deploy the downloaded global modality model θ^t_m.

      Re-calculate $\widehat{Y}^{k,t}$ based on the global modality model θ^t_m.

      (Stage #2) Update the ensemble model $\omega^{k,t} \leftarrow \arg\min_{\omega^{k,t}} \mathcal{L}(Y^{k,t}, \widehat{Y}^{k,t})$.

    end for

  end for

  Update the modality models for each modality $\theta_m \leftarrow \theta^T_m$.

  Update the ensemble models for each client $\omega^k \leftarrow \omega^{k,T}$.

end for

where $\mathcal{L}$ is a loss function that measures the discrepancy between the true label set and the predicted label set. For the ensemble model ω^k, the learning objective is to minimize the discrepancy between the true label set and the predicted label sets across all modalities, as $\min_{\omega^k} \mathcal{L}(Y^k, \widehat{Y}^k)$.

D. Performance-Communication Trade-Off

Due to the resource constraints on edge devices serving as clients, they may not possess adequate storage capacity to house extensive multimodal data, computational capability to learn on multimodal datasets, or communication bandwidth to upload several models to the server. Thus, we introduce two metrics to assist clients to determine whether to upload their models to the server:

- **Shapley value** (φ) represents the impact of a modality model on the final prediction, where a higher value of φ is preferred ($\uparrow$).
- **Modality model size** ($|\theta|$) pertains to the communication overhead, where a lower value of $|\theta|$ is preferred ($\downarrow$).

Shapley Value (Impact). During each communication round, clients evaluate the impact of the models Θ^k on the outcomes utilizing interpretability techniques and choose to upload only a singular selected model θ^k. We consider using the Shapley value as an assessment to evaluate the relationship between input $\widehat{\mathbb{Y}}^k$ and output $\widehat{Y}^k$ of the ensemble model ω^k:

$$\varphi^k_m = \sum_{\mathcal{Y} \subseteq \widehat{\mathbb{Y}}^k \setminus \{m\}} \frac{|\mathcal{Y}|!(|\widehat{\mathbb{Y}}^k| - |\mathcal{Y}| - 1)!}{|\widehat{\mathbb{Y}}^k|!} \left(\omega^k(\mathcal{Y} \cup \{m\}) - \omega^k(\mathcal{Y})\right), \tag{6}$$

where φ^k_m is the Shapley value of input modality m, $\mathcal{Y}$ is a subset of $\widehat{\mathbb{Y}}^k$ excluding modality m, and $\omega^k(\mathcal{Y})$ is the predicted value using only modalities in set $\mathcal{Y}$. For all modalities, we assess the magnitude of each Shapley value by taking its absolute value and construct the following set:

$$\Phi^k = \left\{|\varphi^k_1|, |\varphi^k_2|, \ldots, |\varphi^k_{M_k}|\right\}. \tag{7}$$

Modality Model Size (Communication Overhead). Given modality models with parameters $\Theta^k = \{\theta^k_1, \theta^k_2, \ldots, \theta^k_{M_k}\}$, the communication overhead for each modality model is directly proportional to the model size given by

$$\bar{\Theta}^k = \{|\theta^k_1|, |\theta^k_2|, \ldots, |\theta^k_{M_k}|\}. \tag{8}$$

Priority (Composite Score). Considering the impact of the modality model, as quantified by the Shapley value and the communication overhead as characterized by the modality model size, we propose priority P as a composite score. To derive the priority, we proceed with individual normalization for each criterion:

$$\begin{cases} \tilde{\varphi}^k_m = \dfrac{\varphi^k_m - \min(\Phi^k)}{\max(\Phi^k) - \min(\Phi^k)}, \\ |\tilde{\theta}^k_m| = \dfrac{\theta^k_m - \min(\bar{\Theta}^k)}{\max(\bar{\Theta}^k) - \min(\bar{\Theta}^k)}, \end{cases} \text{for } m = 1, 2, \ldots, M_k, \tag{9}$$

where $\tilde{\varphi}^k_m$ represents the normalized Shapley Value and $|\tilde{\theta}^k_m|$ denotes the normalized communication overhead. With a focus on identifying modality models for server communication, we devise the priority P^k_m for each modality and the corresponding set $\mathcal{P}^k$ to determine whether modality models should be sent to the server. They are formulated as

$$\begin{aligned} P^k_m &= \alpha_s \times \tilde{\varphi}^k_m + \alpha_c \times (1 - |\tilde{\theta}^k_m|), \\ \mathcal{P}^k &= \left\{P^k_1, P^k_2, \ldots, P^k_{M_k}\right\}, \end{aligned} \tag{10}$$

where α_s and α_c are the predetermined metric weights, satisfying $\alpha_s + \alpha_c = 1$. Naturally, a modality with the maximal priority is considered optimal.

To streamline our decision-making, we focus on modalities with scores among the top-γ priority:

$$\begin{aligned} \mathcal{P}^k_\gamma &= \operatorname{top}\max_\gamma(\mathcal{P}^k) \\ &= \{x : x \in \mathcal{P}^k \text{ and } |\mathcal{P}^k \cap \{y \,|\, y \geq x\}| \leq \gamma\}. \end{aligned} \tag{11}$$

Hence, the set of modalities that client k communicates to the server becomes:

$$\Theta_{\gamma}^{k}=\left\{\theta_{m}^{k}:\theta_{m}^{k}\in\Theta^{k}\text{ and }P_{m}^{k}\in\mathcal{P}_{\gamma}^{k},\ \text{ for }m=1,2,\ldots,M_{k}\right\}, \tag{12}$$

where the set Θ_{γ}^{k} represents the modality models corresponding to the top-γ priority from $\mathbb{S}^{k}$. Each client will upload a data packet with various details to the server for aggregation, including model parameters θ^{k}, modality information m, the number of samples $|\mathcal{D}_{m}^{k}|$, among others. Likewise, upon downloading from the server, this information will also be retrieved. Note that only the model θ^{k} will be uploaded/downloaded to/from the server. The ensemble model ω^{k} varies across clients, determined by the unique deployment scenarios of each client, such as geographical location, operational duration, external interference, etc.

E. Model Aggregation

Upon receiving data packets from the clients, the server performs a weighted aggregation of the models based on the number of samples for each data modality. For a given data modality m, the server aggregates the model parameters from client k with modality m. The update is given by

$$\theta_{m}\leftarrow\sum_{\theta_{m}^{k}\in\Theta_{p}^{k}}\beta_{m}^{k}\theta_{m}^{k}, \tag{13}$$

where β_{m}^{k} represents the aggregation weight coefficient. Following the methodology adopted in FedAvg [1], these weights are determined based on the number of samples, and can be expressed as

$$\beta_{m}^{k}=\frac{|\mathcal{D}_{m}^{k}|}{\sum_{k=1}^{K_{m}}|\mathcal{D}_{m}^{k}|}, \tag{14}$$

where K_{m} denotes the number of client models received by the server for modality m.

F. Deployment

In the FedMFS framework, only the modality models θ_{m} are aggregated, while the ensemble models ω_{k} remain separate. The purpose is to give the modality models a global perspective, ensuring a wider generalization. These modality models adhere to the classical FL iterative process and are deployed as global modality models. On the contrary, the personalized ensemble model undergoes a two-stage update. In Stage #1, the ensemble model serves as an intermediate state, primarily facilitating the calculation of the Shapley values. It is not deployed in any application (i.e., it is not tested on any test set). In Stage #2, after receiving the global modality models from the server, the clients subsequently update their ensemble model using global modality models in conjunction with local data to achieve the final state.

III. Experiment and Results

A. Experiment Setup

Dataset. We use the multimodal dataset, ActionSense [23], to validate the proposed FedMFS. ActionSense is a comprehensive multimodal dataset that captures human daily activities, integrating various wearable and environmental sensors. It captures data of human interactions with objects and the environment in a kitchen setting as tasks are performed. This experiment illustrates the application of FL leveraging wearable sensors to implement FL and protect user privacy. We utilize six modalities of sensors from ActionSense, with their descriptions shown in Table I. To ensure fairness with comparison, we adopted the sample code provided by ActionSense for preprocessing sensor data, including filtering, resampling (since the sensors have distinct sampling rates), normalization, and so forth.

TABLE I
Description of ActionSense Dataset

Sensor	Type	Position	Feature	Heterogeneity (Missing Data)
Eye Tracking	Position	Head	2	
Myo	EMG	Left Arm	8	
Myo	EMG	Right Arm	8	
Tactile Glove	Pressure	Left Hand	32×32	S06 – S09 [1]
Tactile Glove	Pressure	Right Hand	32×32	S06 – S09 [1]
Xsens	Rotation	Body	22×3	

[1] S06 – S09 refers to subjects 06 through 09.

Base Models. To ensure a fair comparison, for all the aforementioned data modalities, we initially reshape them into two dimensions, i.e., time $\times$ features. We employ a consistent Long Short-Term Memory (LSTM) network structure for all modalities, comprising a single LSTM layer with 64 hidden units, followed by a fully connected layer, and using a LogSoftmax for output. We adopt negative log likelihood loss (NLLLoss), with SGD as optimizer, a learning rate $\eta=0.1$, a batch size of 32, a local training epoch $E=5$, and a total of $T=100$ communication rounds. Four base models for the six modalities; Eye, Myo, Tactile, and Xsens have sizes of 0.07 MB, 0.08 MB, 1.07 MB, and 0.13 MB, respectively.

Baselines. At the system level, we consider three traditional multimodal sensor fusion frameworks for FL, encompassing data-level [6], feature-level [7], and decision-level [8] fusion as our baselines. Additionally, FLASH [9] with uniform model selection probabilities serves as another baseline. To ensure a fair systematic comparison within the FL context, we do not incorporate various specialized techniques present in these baselines, such as co-attention mechanisms. All these baselines employ a uniform network architecture, specifically, an LSTM layer followed by a fully connected layer, utilizing a concatenate strategy for fusion.

Proposed FedMFS. For the proposed FedMFS framework, in addition to the fundamental configurations mentioned above, these modality models do not output logarithmic probabilities but rather provide definitive predicted categories ($\widehat{\mathbb{Y}}$) for the ensemble model (ω). The ensemble model can adopt various choices depending on the specific use case and the resources available on the client, such as voting methods, linear models, k-nearest neighbors (k-NN), etc. Here, we use the Random Forest (RF) as our ensemble model because of its robust interpretability. We perform a subsampling on the dataset,

TABLE II
COMPARISON OF ACCURACY AND COMMUNICATION OVERHEAD AT CUMULATIVE CONSUMPTION OF 50 MB

Method	γ	α_s	α_c	Acc. [1] (%) (↑)	Comm. [2] (MB) (↓)	Comm. Round
Data-level [6]				47.89	19.36	2
Feature-level [7]				50.45	13.97	3
Decision-level [8]				36.44	14.01	3
FLASH [9]				54.81	2.08	24
		1	0	82.90	3.39	14
		0.8	0.2	81.98	2.68	19
	1	0.5	0.5	95.95	0.85	60
		0.2	**0.8**	**97.34**	**0.72**	**69**
		0	1	70.43	0.64	77
		1	0	79.52	5.55	8
		0.8	0.2	86.94	4.63	11
	2	0.5	0.5	92.36	1.67	29
		0.2	0.8	90.31	1.55	32
		0	1	81.62	1.34	37
		1	0	82.61	7.22	6
		0.8	0.2	86.60	6.50	8
	3	0.5	0.5	87.89	2.55	19
FedMFS proposed by us		0.2	0.8	85.48	2.23	22
		0	1	81.94	2.03	24
		1	0	80.68	8.43	5
		0.8	0.2	85.06	8.11	5
	4	0.5	0.5	83.35	5.52	8
		0.2	0.8	83.15	3.24	15
		0	1	84.22	3.24	15
		1	0	82.05	8.93	5
		0.8	0.2	82.65	10.98	4
	5	0.5	0.5	83.21	8.57	5
		0.2	0.8	82.74	8.58	5
		0	1	81.58	8.58	5
		1	0	81.86	9.16	5
		0.8	0.2	74.88	13.93	3
	6	0.5	0.5	73.61	13.93	3
		0.2	0.8	71.84	13.93	3
		0	1	72.69	13.93	3

[1] Average accuracy between clients, ↑ refers to the higher (preferred).
[2] Communication overhead per iteration, ↓ refers to the lower (preferred).

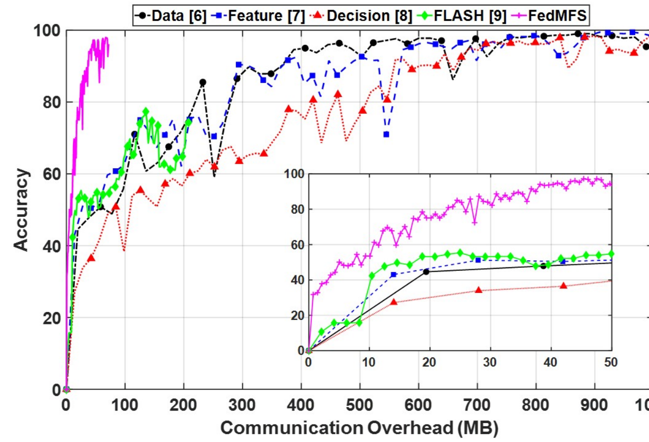

Fig. 2. Comparison of accuracy between FedMFS with the configuration $\gamma = 1, \alpha_s = 0.2, \alpha_c = 0.8$ and four baselines on a communication overhead scale. Only up to 1000 MB of communication overhead is depicted, while the data-level fusion approach requires close to 2000 MB to complete all iterations.

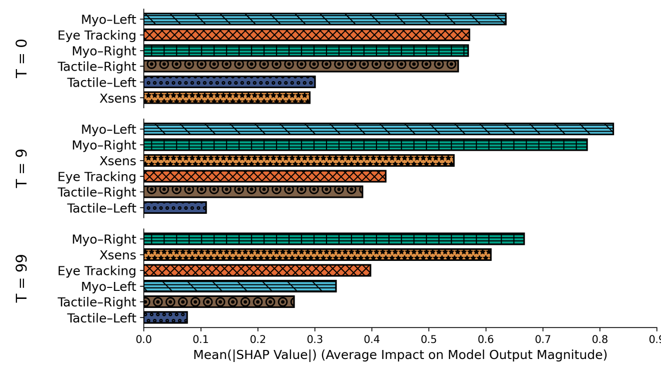

Fig. 3. The impact of modality models on the ensemble model's final prediction throughout the FedMFS iteration, exemplified with the configuration $\gamma = 1, \alpha_s = 0.2, \alpha_c = 0.8$.

selecting 50 samples to compute the Shapley values to reduce computational complexity. Note that the ensemble model, Shapley values, as well as the modality model sizes are kept private by the client and used for modality model selection. They are neither uploaded to the server nor does the server possess knowledge of the client's computing methodology.

B. Results

The results of the proposed FedMFS, in comparison with four baselines on the ActionSense dataset, are presented in Table II and Fig. 2.

Trade-Off Analysis. Considering the communication constraints, our results underscore the need to find a compromise between α_s and α_c to optimize performance when γ is constant. Increasing γ does not always lead to better results, as it can exacerbate communication overhead. In some instances, communicating only a select few informative modalities yields enhanced performance. In this context, the proposed FedMFS framework facilitates a flexible determination of the number of modality models to upload, balancing model performance, communication overhead, and learning efficiency.

Specifically for ActionSense, the configuration of $\gamma = 1, \alpha_s = 0.2, \alpha_c = 0.8$ yields the best accuracy with the least communication overhead. This configuration leads most clients to predominantly upload modalities such as Eye Tracking, Myo–Right, and Xsens, which are characterized by fewer input features and thus they are more compact. From extensive experimentation, we observed that in the initial stages of FL, the Eye Tracking modality consistently had a higher upload frequency due to its minimal modality model size, granting it an advantageous position in the trade-off. However, as communication rounds progress, the selection frequency of Eye Tracking decreases, while that of Myo–Right or Xsens increases. This trend can be attributed to the fact that, although the Eye Tracking modality results in the least communication overhead, its limited feature set captures a lower degree of information, giving a slight lag in recognition accuracy.

Comparison with Baselines. The proposed FedMFS not only exceeds the four baselines in terms of accuracy but also manages to reduce communication overhead by nearly an order of magnitude, thanks to its selective upload strategy.

Insights obtained from FedMFS's selection approach are:

(i) Different data modalities contribute distinctively to recognition accuracy.
(ii) Aggregation of all modal models is not always necessary.
(iii) Utilizing all modalities for fusion is not imperative.

Although the FLASH-adopted random upload strategy effectively reduces communication costs to approximately $\frac{1}{M_k+1}$, it lacks a performance and communication overhead-based selection mechanism, leading to suboptimal results. It is worth noting that the sizes of the modality models vary, which accounts for the reduced communication overhead of FedMFS compared to the FLASH framework. Fig. 2 illustrates that FedMFS can achieve much faster convergence with minimal communication overhead, superior performance, and high learning efficiency.

Interpretability of FedMFS. In addition to helping clients select the modality models, the Shapley value also offers an interpretative approach to quantify the modality models. During the FL process, we can see the clients' favor and the efficacy of each modality model within the FedMFS framework. Fig. 3 illustrates this dynamics, showing the impact of each data modality on the final prediction of the ensemble model across different communication rounds T. In the initial stages of FL, most modalities exhibit similar impacts. As communication rounds progress, modalities with larger feature sets and complex models, due to their higher communication overhead, take on a subordinate role in their selection within FedMFS. As FL advances, more straightforward modalities that still convey ample information, such as Myo–Right, emerge as primary contributors.

IV. Conclusion

In this paper, we presented the FedMFS framework, leveraging Shapley values and modality model sizes to quantify the performance and communication overhead of each modality. In addition to achieving considerable recognition accuracy and reducing communication costs by almost one twentieth, the proposed FedMFS is suitable for heterogeneous clients, features detachable modular modality models, and offers interpretability for data modalities. Our future work will focus on enhancing the adaptability of FedMFS with customizable configurations to fully exploit scenarios where clients might possess dynamic communication capabilities, such as higher bandwidth. Furthermore, Shapley values can also aid in refining the training process of modality models, for example, by potentially discarding underperforming modalities like Myo–Left, thus optimizing computational efficiency.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[2] W. Fang, Z. Yu, Y. Jiang, Y. Shi, C. N. Jones, and Y. Zhou, "Communication-efficient stochastic zeroth-order optimization for federated learning," *IEEE Transactions on Signal Processing*, vol. 70, pp. 5058–5073, 2022.

[3] V. P. Chellapandi, L. Yuan, S. H. Zak, and Z. Wang, "A Survey of Federated Learning for Connected and Automated Vehicles," *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2023.

[4] V. P. Chellapandi, L. Yuan, C. G. Brinton, S. H. Zak, and Z. Wang, "Federated Learning for Connected and Automated Vehicles: A Survey of Existing Approaches and Challenges," *arXiv preprint arXiv:2308.10407*, 2023.

[5] L. Yuan, Y. Ma, L. Su, and Z. Wang, "Peer-to-peer federated continual learning for naturalistic driving action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5249–5258.

[6] P. Qi, D. Chiaro, and F. Piccialli, "FL-FD: Federated learning-based fall detection with multimodal data fusion," *Information Fusion*, p. 101890, 2023.

[7] B. Xiong, X. Yang, F. Qi, and C. Xu, "A unified framework for multi-modal federated learning," *Neurocomputing*, vol. 480, pp. 110–118, 2022.

[8] T. Feng, D. Bose, T. Zhang, R. Hebbar, A. Ramakrishna, R. Gupta, M. Zhang, S. Avestimehr, and S. Narayanan, "Fedmultimodal: A benchmark for multimodal federated learning," *arXiv preprint arXiv:2306.09486*, 2023.

[9] B. Salehi, J. Gu, D. Roy, and K. Chowdhury, "Flash: Federated learning for automated selection of high-band mmwave sectors," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 1719–1728.

[10] J. Chen and A. Zhang, "Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 87–96.

[11] L. Yuan, L. Sun, P. S. Yu, and Z. Wang, "Decentralized federated learning: A survey and perspective," *arXiv preprint arXiv:2306.01603*, 2023.

[12] V. P. Chellapandi, A. Upadhyay, A. Hashemi, and S. H. Żak, "On the Convergence of Decentralized Federated Learning Under Imperfect Information Sharing," *IEEE Control Systems Letters*, 2023.

[13] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," in *International conference on machine learning*. PMLR, 2020, pp. 9269–9278.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[15] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[16] L. Yuan, J. Andrews, H. Mu, A. Vakil, R. Ewing, E. Blasch, and J. Li, "Interpretable passive multi-modal sensor fusion for human identification and activity recognition," *Sensors*, vol. 22, no. 15, p. 5787, 2022.

[17] L. Yuan, H. Chen, R. Ewing, E. Blasch, and J. Li, "Three dimensional indoor positioning based on passive radio frequency signal strength distribution," *IEEE Internet of Things Journal*, vol. 10, no. 15, pp. 13 933 – 13 944, 2023.

[18] L. Yuan, H. Chen, R. Ewing, and J. Li, "Passive radio frequency-based 3d indoor positioning system via ensemble learning," *arXiv preprint arXiv:2304.06513*, 2023.

[19] S. Yang, L. Yuan, and J. Li, "Extraction and denoising of human signature on radio frequency spectrums," in *2023 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2023, pp. 1–6.

[20] S. M. Shah and V. K. Lau, "Model compression for communication efficient federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[21] L. Yuan, L. Su, and Z. Wang, "Federated transfer-ordered-personalized learning for driver monitoring application," *IEEE Internet of Things Journal*, May 2023.

[22] C. Wu, F. Wu, L. Lyu, Y. Huang, and X. Xie, "Communication-efficient federated learning via knowledge distillation," *Nature communications*, vol. 13, no. 1, p. 2032, 2022.

[23] J. DelPreto, C. Liu, Y. Luo, M. Foshey, Y. Li, A. Torralba, W. Matusik, and D. Rus, "Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 800–13 813, 2022.