# FedHARM: Harmonizing Model Architectural Diversity in Federated Learning

Anestis Kastellos[1], Athanasions Psaltis[1,2], Charalampos Z. Patrikakis[2], and Petros Daras[1]

[1] Centre for Research and Technology Hellas, Thessaloniki, Greece
`{kastellosa, psaltis, daras}@iti.gr`
[2] Dept. of Electrical and Electronics Engineering, University of West Attica, Athens, Greece
`{apsaltis, bpatr}@uniwa.gr`

**Abstract.** In the domain of Federated Learning (FL), the issue of managing variability in model architectures surpasses a mere technical barrier, representing a crucial aspect of the field's evolution, especially considering the ever-increasing number of model architectures emerging in the literature. This focus on architecture variability emerges from the unique nature of FL, where diverse devices or participants, each with their own data and computational constraints, collaboratively train a shared model. The proposed FL system architecture facilitates the deployment of diverse convolutional neural network (CNN) architectures across distinct clients, while outperforming the state-of-the-art FL methodologies. $FedHARM$[3] capitalizes on the strengths of different architectures while limiting their weaknesses by converging each local client on a shared dataset to achieve superior performance on the test set.

**Keywords:** Representation Learning · Federated Learning · Model-Agnostic

## 1 Introduction

Modern machine learning has witnessed an expansion of model architectures, each tailored to specific types of data and computational tasks. For instance, ResNets (Residual Networks) [11] have gained popularity for their ability to enable training of extremely deep networks by using skip connections. On the other hand, EfficientNets [37] offer a balanced approach, scaling different dimensions of the network in a compound manner to achieve remarkable efficiency and accuracy. Similarly, MobileNets [12] are designed to provide lightweight, yet effective, neural networks for mobile and edge devices, focusing on optimized performance in resource-constrained environments. Such architectures have exceled in a wide spectrum of computer vision tasks including classification [1, 23, 31], detection [9, 36], retrieval [7, 18].

[3] Code: https://github.com/Kastellos/FedHARM

In FL, diverse model architectures present both challenges and opportunities. Each architecture processes data differently, with varying computational complexities and task suitability. For example, ResNets are ideal for high-accuracy needs, while MobileNets are better for mobile devices due to efficiency. As detailed in previous studies [19], FL involves participants from powerful servers to resource-constrained edge devices collaborating on a shared model with unique datasets, leading to heterogeneity in computational and data characteristics. Data across devices are often non independently and identically distributed (non-IID), complicating model training and performance. Thus, choosing the right model architecture is crucial, requiring a tailored strategy to match each node's capabilities and data.

In previous studies [26, 33], communication efficiency and privacy preservation have been identified as core challenges in FL , where the decentralized architecture adds another layer of complexity. Optimizing model architectures for diverse environments is crucial for reducing communication overhead and accommodating the privacy and security needs of different nodes. This optimization, alongside the development of flexible models that can handle incremental learning and adapt to new data without forgetting prior information, is essential for maintaining learning efficiency and effectiveness. Additionally, the scalability and robustness of FL systems, capable of adapting to varying architectures across a broad range of domains and applications, are vital. Addressing these challenges requires innovative approaches in model aggregation, updating, and architecture variability management to ensure FL's practical deployment and success in fields ranging from healthcare to smart cities, and the exploration of hybrid models or meta-learning strategies for more versatile and powerful FL systems.

Considering all the aforementioned aspects, the exploration of model architecture variability in FL stands as a strategic response to the multitude of challenges and opportunities inherent in this innovative learning paradigm. It embodies a dedicated effort to enhance the efficiency, scalability, privacy, and real-world applicability of FL. This attempt builds upon the foundational insights extracted from preceding studies, pushing the boundaries of what is achievable in collaborative, decentralized learning environments. Notably, this study is the first, to the authors' knowledge, to step on this path. It pioneers in addressing the complexity and diversity of model architectures within the FL framework, making it a ground breaker in the field and setting a point of reference for future research in this area. The proposed FL system is illustrated in Figure 1.

The main contributions in addressing model architecture variability in FL can be summarized as follows:

a) **Model-Agnostic FL Framework:** This study presents a novel FL approach compatible with various model architectures, including ResNet, EfficientNet, and MobileNetV3, showcasing its adaptability to different computational capabilities and data needs across nodes.
b) **Hybrid Learning Approach with Focus on Representation:** This study innovatively blends supervised and self-supervised learning for rep-

resentation learning. It optimizes local models with specific data and aligns them with global patterns through self-supervised learning, moving away from traditional federated averaging to a representation-centric aggregation method.

c) **Efficient Handling of Model Architecture Variability:** The study tackles combining outputs from different model architectures in FL. By focusing on specific blocks or layers and using feature extraction, it effectively aggregates learned representations, ensuring coherence in the federated system.

## 2 Related Work

The landscape of FL has rapidly evolved, as evidenced by the plethora of research [3, 5, 6, 28, 41, 43, 44] that have tackled its fundamental issues and put forth inventive ways to improve its effectiveness, confidentiality, and expandability. The promise of this emerging sector to facilitate collaborative learning without sacrificing data privacy has garnered a great deal of attention.

Representation learning serves as a cornerstone in the efficacy of FL, underpinning the ability to distill and generalize knowledge from distributed data. This foundational technique facilitates the extraction of informative features, enhancing the collaborative intelligence of FL models, enabling them to perform robustly across all client datasets. The authors of [4] introduced an FL framework that aims to cultivate a common data representation among clients while maintaining distinct local heads for each. This approach capitalizes on the distributed computing resources and performs frequent local updates targeting low-dimensional parameters to achieve linear convergence and produce accurate representations. FedX [10] learns neutral representations from diverse local data by utilizing contrastive learning [8] to enable a bilateral knowledge distillation, allowing the system to function without the need for shared data features. Due to privacy constraints, FL schemes are required to excel on limited data on a per-client basis. Concurrently, as knowledge is centralized through model updates, it is essential for the aggregated model to assimilate and generalize information across the dataset from all clients, while preventing the loss of client-specific characteristics. The work of Psaltis *et al.* [30] demonstrates that contrastive learning can significantly increase the performance of the local clients even when the distribution is imbalanced and scattered across them. Li *et al.* [20] employ contrastive learning at the model level, leveraging the congruence among model representations to refine the local training processes of individual clients. FedCA [45] used a shared split of the dataset to align the representations of the images. The framework is distinguished by its integration of centralized sample representations from each client to ensure a uniform representation space accessible by all, and a module that synchronizes each client's representation with that of a foundational model trained on publicly available data.

Addressing the aggregation of learned representations on a global scale presents significant challenges. Despite numerous attempts to synchronize heterogeneous models across diverse clients, current approaches struggle to achieve seamless architecture alignment without modifying model structures or directly aggregating model weights. This limitation points to the untapped potential for innovative methods capable of integrating diverse architectures without the need for altering client models. Although existing research has explored methods like distillation techniques or various ensemble strategies [22,29,34,40,46], these approaches often necessitate mapping features into a common latent space to a shared latent space for knowledge integration and weight aggregation [13,14,25,38,39]. Such processes typically involve additional training of either entire networks or specific network components (*e.g.*, network heads), leading to increased training durations and reduced efficiency and scalability, thereby diminishing the practicality of these solutions.

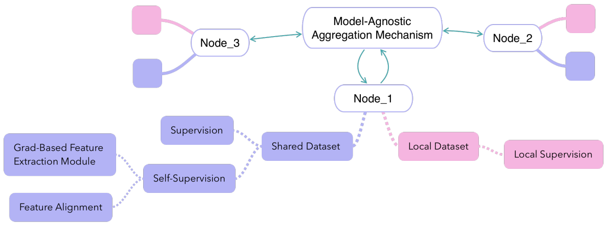

**Fig. 1:** Schematic of the proposed FL System Architecture illustrating the feature extraction and alignment modules on the local dataset of each client that harmonizes the heterogenous architectures and the fully-supervised training on the local dataset. The Model-Agnostic Aggregation Mechanism is the process performed on the main server to create an enriched representation of the local dataset, derived from all the clients

## 3 Strategies for Managing Diverse Model Architectures

**Problem statement:** The effective integration and management of a wide array of diverse model architectures across various nodes in a federated network. This problem arises due to the heterogeneous nature of FL environments, where nodes, ranging from high-powered servers to resource-constrained edge devices, each come with their own unique data characteristics and computational capabilities. The proposed methodology for managing diverse model architectures in an FL systems is designed to be model-agnostic, leveraging the strengths of representation learning. This strategy is distinct from the conventional federated averaging, taking advantage of the potency of both supervised and self-supervised learning paradigms, but lacking the advantage to aggregate the weights of the models duo to the heterogeneity of the architectures. Traditional approaches, including FedAvg [27], are insufficient in confronting the challenges posed by the intrinsic diversity of machine learning models deployed throughout the network. Moreover, strategies such as model distillation and the integration of feature fusion at the upper layers, although aimed at ameliorating these difficulties, do not succeed in providing a substantive resolution.One of the key aspects of our methodology is the accommodation of various model architectures across local clients. The technique is particularly concentrated on CNN architectures, specifically: (i) ResNet, (ii) EfficientNet, and (iii) MobileNetV3. This flexibility ensures that each local node can select a model that best fits its computational capabilities and specific data requirements. Such diversity in model architecture is crucial for the adaptability and personalization of the FL system.

### 3.1 Local Supervision and Self-supervision Representation Learning

Within the methodology described in this research, there are two learning paradigms integrated, that are applied to distinct subsets of the datasets. Every client in the FL system has access to three distinct sets of data in terms of distribution: a test set, a private set, and a shared set. To foster a collaborative learning environment, all clients have access to the shared set, which is an image collection that is used for training. Conversely, the private set is a unique group of images for each client that is extracted from the original training dataset following the subtraction of the images assigned to the shared set. The training scheme applied to the private set employs a fully supervised learning procedure, enabling each client to learn and adapt to the distinct attributes and patterns that derive from the local data. A hybrid learning model is adopted for the shared set, encompassing both supervised and self-supervised learning mechanisms. It utilizes the labels to encourage the learning of specific patterns within the set, while simultaneously generating a descriptor for each image. The objective is to align the individual descriptors with a collectively aggregated embedding which is constructed by all the clients, thus boosting cohesiveness and efficacy of the learning process across the federated system. This shared resources plays a pivotal role in the latter stages of the training process, particularly during the self-supervised learning phase. It serves as a unifying element, bridging the diverse learning experiences of individual nodes and aligning them with the broader, global dataset context.

**Gradient-based Feature Extraction Module** The core of our excitation technique is grounded in the observation that the most influential features are highlighted by the gradients during the backpropagation process. Drawing inspiration from the insights provided in [42], where the authors underscore the significance of gradients in evaluating feature importance within their personalized FL system, the proposed methodology advances this understanding. Gradients within a model delineate the direction of optimization and effectively indicate the influence of each neural unit. The feature extraction method we propose builds upon this concept, harnessing the model's gradients relative to the ground truth to identify and extract pivotal features from each block's output. By selecting features that possess the largest absolute gradients, the algorithm ensures the inclusion of those with the most substantial impact on the model's predictive outcome, thereby enriching the feature descriptor with the most influential attributes for the task at hand. The formula 1 of the module is depicted below:

$$\text{grad}_j^b = \text{Top}k\left(\left|\frac{\partial p_j^c}{\partial F_j^b}\right|\right), k \in \{128, 256, 512, 1024\}, b \in \{1, 2, 3, 4\} \tag{1}$$

where $b$ represents the block of the model, $k$ number of features to be extracted, $j$ the training sample, $p_j^c$ the predictive output of the $c$-th category. This approach begins with initial training at each local node using its own private dataset, a critical phase for developing accurate data representations. These representations lay the groundwork for the method's subsequent phases. Following this initial training, every node processes the images from a shared subset. For each image, the algorithm selects the top features exhibiting the largest absolute gradients within each of the blocks of the network's feature map, as shown in the Algorithm 1 lines 6-12. It is these prominent features, considered pivotal for the image's representation, that are then forwarded to the central server. This process ensures that the most critical aspects of the data are emphasized and aggregated globally, enhancing the overall learning and representation capability of the system. In Figure 2, Grad-cam [32] was used to visually compare the gradient maps of each block across the network architectures

A comprehensive evaluation identified the most informative characteristics for constructing a meaningful feature descriptor, using targeted feature cropping from the feature maps and various adaptive pooling strategies, ultimately finding that the proposed Gradient-based Feature Extraction Module outperformed these techniques in feature selection efficacy.

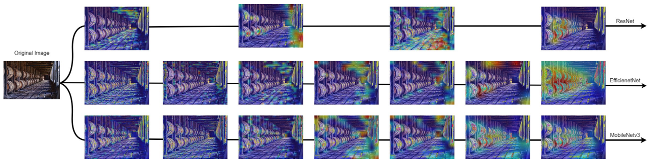

**Fig. 2:** Visual representation of the gradient visualizations for three different neural network architectures applied to the same original image. On the left, the original image is depicted. Progressing to the right, the subsequent images represent gradient heatmaps as interpreted by ResNet, EfficientNet, and MobileNetV3, respectively. These heatmaps highlight areas of the image that contribute most significantly to the models' predictions, with warmer colors indicating higher gradient values and thus greater importance in the decision-making process of each network.

### 3.2 Model-Agnostic Aggregation Mechanism

The central node in this architecture plays a vital role in aggregating the representations from each participating node. This aggregation focuses on the outputs from specific blocks or layers of the models, going beyond the individual differences in architectures. The process is designed to be model-agnostic, accommodating the wide variety of models without requiring uniformity. A significant departure from traditional approaches in our methodology is the avoidance of weight aggregation algorithms. Instead, we focus on a representation-centric aggregation approach. This shift ensures that the learning process is more coherent, communication efficient and effective, particularly suitable for environments with heterogeneous models. The ultimate goal is to align the local representations of the shared dataset with the global representation. The global representation is the aggregation of the local descriptors of each image in the shared subset, contributed by all participating clients in the FL system. This alignment enhances the model's ability to generalize and adapt to new data, improving its performance across the federated network. The hybrid training approach – combining supervised and self-supervised learning – ensures that local models are fine-tuned to both their specific data characteristics and the aggregated knowledge from the shared dataset. In conclusion, this proposed aggregation methodology offers a flexible and effective approach to managing diverse model architectures in FL. By focusing on representation learning and adopting a model-agnostic aggregation approach, it ensures efficient and robust learning across the federated network, paving the way for more adaptable and powerful FL systems.

**Iterative Representation Refinement and Dual-Loss Aggregation** At the central server, an average representation for each image is computed by aggregating the feature information received from all clients. This global representation encapsulates the collective insights of all participating nodes, enriching the understanding of each image in the shared dataset. The training of the network then proceeds with a cosine similarity loss function. In each epoch of this training phase, two views of the data are considered: the global representation and the current local representation of the network being trained. The first epoch plays a pivotal role as it sets the baseline for the representations used in a knowledge distillation-inspired learning process. The representations for the subsequent epochs are then recalculated based on the top feature indices identified in the previous epoch. This iterative process is refined further by incorporating both cosine similarity and cross-entropy loss functions into the global aggregation step. This dual-loss approach allows the model to benefit from the strengths of both supervised and self-supervised learning paradigms, leading to more robust and well-rounded representations. By continuously refining the representations and the model through this iterative process, the FL system efficiently leverages the most relevant features of the data, enhancing the model's performance and adaptability to diverse datasets. Algorithm 1 provides a detailed illustration of the proposed aggregation technique in the suggested FL framework. The algorithm's goal is to reduce inconsistencies between the local and global features, promoting uniformity and coherence in the learned representations across the network. Upon successful alignment of the local models with the global representation, the updated models are aggregated at the central server, enhancing the global model with enriched insights from the network's distributed learning experience. This strategy ensures that local models not only perform well on their data but also contribute effectively to the collective intelligence of the federated system.

**Image Descriptor - Feature Alignment** The core innovation of the proposed approach is based on the hypothesis that we can directly map blocks of features across different model architectures, owing to the similarity in the information they encode, and then aggregate them accordingly. This strategy suggests a more streamlined and potentially effective method for combining diverse models by capitalizing on the inherent parallels in their feature representations. However, in the journey of aggregating learned representations from different blocks of each node in our FL system, we encountered several significant challenges, particularly with the alignment metrics between node representations, especially with the implementation of cosine similarity as a measure of alignment. Unexpectedly, the use of cosine similarity not only failed to enhance the network's performance but rather led to a decrease in effectiveness. This issue was further complicated by the observation that for the majority of the layers, except the last one, the cosine similarity consistently equated to zero. The equation for cosine similarity is as follows:

$$\text{cosine similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

where $A$ and $B$ are two vectors for which you are calculating the cosine similarity. The dot product of $A$ and $B$ is divided by the product of their magnitudes (or Euclidean norms). The magnitudes are calculated as the square root of the sum of the squared elements of each vector.

The construction of an image embedding derived from the entirety of a client's model, and the utilization of solely the last flattened embedding, did not yield a notable enhancement in accuracy. This outcome was observed despite each client's embedding being a high-dimensional 1920-vector, which was meticulously generated through the Gradient-based Feature Excitation Algorithm, a variance attributable to the disparate model architectures and the dimensions of the latent spaces involved.

The main descriptor's experimentation began with ResNet architectures — ResNet18, ResNet34, and ResNet50—due to their prevalence in FL literature and computational constraints. The descriptor is constructed by aggregating specific features from each primary block of the network. The technique encompasses the strategic extraction of features from all the blocks of the ResNet architectures to frame a thorough 1920-dimensional embedding, utilizing the Gradient-based Feature Extraction Module. This process initiates with the selection of 128 features from the initial main block, followed by the extraction of 256 features from the second block, 512 from the third, and culminating with 1024 from the fourth block. When trying to align heterogeneous network architectures, first the block-alignment process is performed and the selected blocks will create the image descriptor

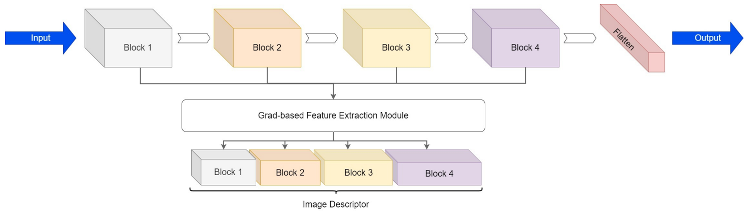

**Fig. 3:** Diagram of the descriptor construction process from a ResNet architecture, illustrating the sequential extraction of features from the four distinct blocks of the model. The feature maps are processed through a Feature Extraction Module, resulting in a composite image descriptor that encapsulates multi-scale representations of the input.

In particular, the process of feature elicitation serves as a pivotal step in harmonizing the local models with a central, aggregated global accumulation of the descriptors. It starts with an image passing through a neural network, which is segmented into distinct blocks, each responsible for extracting features at varying levels of complexity. The early blocks, such as *Block1*, typically capture elementary features like edges, while the deeper blocks, such as *Block4*, discern more intricate patterns. The outputs of these blocks undergo a selective process where a specific number of features are chosen based on their activation levels, which are indicative of their importance. These chosen features are then combined, with the deeper blocks contributing a larger share of features, reflecting their increased complexity and importance in characterizing the image. This process is illustrated in Figure 3. The reasoning behind this specific excitation approach is grounded in empirical findings. Experiments have revealed that similarity metrics for features in the initial layers of both trained and untrained models are remarkably high, indicating homogeneity in the primitive features extracted by these layers. Consequently, as the network delves deeper, the need for capturing a broader and more complex range of features grows, prompting the selection of larger feature maps from the deeper layers—(128, 256, 512, 1024) respectively—to ensure a comprehensive representation.

In the Algorithm 1 presented before, $L_{CE}$ and $L_{COS}$ represent the Cross-Entropy and Cosine Similarity loss functions that are employed in the training process of the proposed method. $b_{grad}$ is the gradient map of the block of the models, $P_{D_i}$ represent the predictions of the $i$-th dataset of the $i$-th client's model. The variables $w$ and $h$ are the spatial dimensions of the feature maps, $F_E^i$ represent the extracted features from the $i$-th client's model, while $F_{E_j}$ is the aggregated descriptor manufactured in the main server and $z_{local}$ is the prediction of the model of the $x_j$ image and finally, the factor $\theta$ denote the weights of the model.

**Algorithm 1** Federated Learning with Grad-based Feature Extraction

```
Require: $D_i$ local datasets, $f^i(\theta)$ model of the local clients, $C$ common subset, $R$ total
    communication rounds, $E$ total epochs, $lr$ learning rate
for $r \leftarrow 1$ to $R$ do
    Local training
    for $i \leftarrow 1$ to $N$ do
        $f^i(\theta) \leftarrow \text{train}(f^i(\theta), D_i, L_{CE})$
    end for
    Grad-based Feature Extraction Module
    for $i \leftarrow 1$ to $N$ do
        for all $b$ in $f^i(\theta)_{blocks}$ do
            $b_{\text{grad}} \leftarrow \frac{\partial P_{D_i}}{\partial f^i(\theta)_{b,(wh)}}$
            $F^i_E \leftarrow \text{ExtractTopK}(|b_{\text{grad}}|), K \in \{128, 256, 512, 1024\}$
        end for
    end for
    Aggregation on Central server
    for $j \leftarrow 1$ to $C$ do
        $F_{Ej} \leftarrow \frac{1}{N} \sum_i^N F^i_{Ej}$
    end for
    for $i \leftarrow 1$ to $N$ do
        for $e \leftarrow 1$ to $E$ do
            for $j \leftarrow 1$ to $C$ do
                $F^i_{Ej} \leftarrow \text{Grad-based Feature Extraction Module}(f^i(\theta), P_j)$
                $z^j_{\text{local}} \leftarrow f^i(\theta)(\text{x}_j)$
                $L \leftarrow \alpha L_{CE}(z^j_{\text{local}}, P_j) + (1-\alpha) L_{COS}(F^j_{iE}, F^E_j)$
                $\theta \leftarrow \theta - lr \cdot \nabla L$
            end for
        end for
    end for
end for
```

**Harmonizing diverse architecture through block alignment strategies:** Expanding the scope of architectures to include EfficientNet and MobileNetV3, specifically EfficientNetB0, EfficientNetB1 and MobileNetV3-Small, MobileNetV3-Large, alongside the ResNet family, a nuanced approach to integration is necessitated by the variance in the total number of blocks within these models—EfficientNet and MobileNetV3 comprising seven blocks in contrast to ResNet's four. The incorporation of a diverse array of architectures, demanded a sophisticated alignment policy to accommodate the variations in block numbers—seven in EfficientNet and MobileNetV3 as opposed to four in ResNet. This alignment was meticulously conducted through an analysis of representational similarity and angular divergence among the models, utilizing cosine similarity metrics to achieve uniformity in representation spaces. The process entailed the synchronization of blocks that demonstrated the least angular divergence, taking into account the orientation and dimensionality of their feature maps. Despite challenges in direct comparison due to activation function and dimensionality differences, absolute feature grouping and KL-Divergence analysis enabled precise block alignment across architectures, fostering a unified FL framework. Consequently, the harmonization of block connections across ResNet, EfficientNet, and MobileNetV3 architectures is achieved, paving the way for a cohesive and aligned FL system that leverages the strengths of diverse model architectures while maintaining the integrity of their unique representational capacities. The primary computational expense stems from the gradient-based feature extraction module; however, this cost is effectively mitigated by the communication overhead, which is further analyzed in the supplementary material.

## 4 Experimentation in Various Architectural Scenarios

The exploration of different model architectures in FL aims to understand how diverse neural network structures can affect the learning process when distributed across various nodes. In the series of experiments conducted, the focus was on integrating three core CNN-based architectures-ResNet, EfficientNet, and MobileNetV3-randomly initialized, to evaluate their performance in FL settings. The initial set of experiments deployed various versions of ResNet, specifically ResNet18, ResNet34, and ResNet50, but memory constraints limited the exploration to these three models, highlighting the practical challenges of deploying larger and more complex networks in FL.

### 4.1 Experiment Setup

**Datasets** The datasets used to validate the Fl system are CIFAR-10, CIFAR-100 [16] and MNIST [17], with data distributed to clients in both IID and non-IID formats. Regarding the shared subset, experiments are conducted in a range 5000 - 10000 data instances, to discover the optimal number. It was revealed that despite the fact that the alignment process performed better the bigger the share subset was, the remaining data that are divided to the clients proved insufficient for the models to effectively capture their unique patterns, so the total number of the shared division of the dataset is set to 5000.

**Data Augmentations** In the preprocessing phase of the study, a series of image augmentation techniques are implemented to enhance the diversity of the training dataset. The specific augmentations applied include horizontal and vertical flip, random crop, brightness adjustment of $+/-20\%$ and Gaussian blur.

**Distribution of the Networks to Clients** The models are allocated to clients utilizing a standardized approach, where each variation of the networks is selected according to a uniform distribution method, with a requirement that the method contains at least one of each different network architectures.

**Hyperparameters** The total communication rounds of our experiments are set to 10, while the epochs of training of each round are set to 25, the batch size is set to 128, the initial learning rate of the models was 0.01, using One-Cycle LR [35] with the minimum learning rate being 0.0001 and the optimized of the networks was the Adam [15]. The total number of variation to the number of clients is 5, 10, 20.

**Compared Methodologies** To facilitate a meticulous comparative analysis of the proposed framework's performance, we have chosen benchmark methods grounded in the principles of representation learning. The FL schemes selected for this purpose include $PerFCL$ [47], $FedCon$ [24], $MOON$ [21], $FedSimCLR$, $FedCA$ [45] and $FedSimSiam$ [2]. However, it is important to highlight that the proposed approach diverges from the existing state-of-the-art methodologies by incorporating a varied architecture at each federated node. This distinction is a critical factor that should be considered in any comparative analysis below.

### 4.2 Evaluation and Results

The initial goal was to evaluate the efficacy of various ResNet models disseminated among the clients. The resulting performance is the mean accuracy of the corresponding method across all the clients on the test set of each database.For the evaluation metric the accuracy is the mean accuracy of the corresponding method across all the clients on the validation set of each database. $FedHARM_{res}$ denotes the adaptation of this approach, incorporating ResNet architectures for client distribution, and it outperformed all competing methodologies on the CIFAR-10 dataset across every client configuration as shown in Table 1. A slight decline in performance was observed with an increase in the number of clients, attributable to the correspondingly reduced data available to each client. The integration of the alignment module significantly enhanced the framework's effectiveness, demonstrating its utility and success by achieving an accuracy of 82.88% in a scenario with 5 clients and 81.51% in a setting with 10 clients. The added value is further underscored by the fact that, unlike traditional FL strategies, the conventional solution avoids using weight averaging algorithms and instead opts for dominant feature utilization, showcasing its flexibility in extracting knowledge from a variety of model architectures.

Transitioning to the CIFAR-100 dataset, the proposed method exhibited superior performance in the 5-client configuration compared to the benchmark methods achieving 52.38% accuracy, but did not achieve the same level of success in the settings with 10 and 20 clients. This reduction in performance can be attributed to the insufficient amount of data per client in the local datasets, which hindered the creation of a robust embedding capable of accurately representing the data instances across the 100 classes of the dataset.

In the second set of experiments, the focus shifted to integrating EfficientNet and MobileNetV3 models alongside ResNet models, referred to as $FedHARM_{rem}$. These architectures are recognized for their efficiency and performance on mobile

**Table 1:** Accuracy (%) comparison of FL Methods on CIFAR-10 and CIFAR-100 datasets in IID settings. The validation was performed on the test set of each database, and the resulting number is the mean accuracy across the clients.

| Method | CIFAR-10 | | | CIFAR-100 | | |
|---|---|---|---|---|---|---|
| | 5 Clients | 10 Clients | 20 Clients | 5 Clients | 10 Clients | 20 Clients |
| $FedCon$ [24] | - | 81.47 | - | - | - | - |
| $FedSimCLR$ | 68.1 | - | - | 39.75 | - | - |
| $FedCA$ [45] | 71.25 | - | - | 43.30 | - | - |
| $FedSimSiam$ [2] | 76.27 | - | - | 49.79 | - | - |
| $FedHARM_{\text{res}}$(ours) | 82.88 | 81.51 | 78.01 | 52.38 | 52.54 | 38.89 |
| $FedHARM_{\text{rem}}$(ours) | **83.87** | **82.03** | **79.76** | **53.9** | 51.49 | 40.03 |

devices, making them an intriguing choice for FL scenarios. Within the context of the CIFAR-10 dataset, the $FedHARM_{rem}$ variant outperformed all other methodologies evaluated, including $FedHARM_{res}$, showcasing superior efficacy across various client settings. This improved performance is supported by the fact that architectures such as EfficientNet and MobileNetV3 generally exhibit better performance compared to ResNet variants. The enhancement in the quality of representations learned at the local dataset level indicates that the overall descriptors of the shared subset have facilitated the creation of a more significant embedding. Consequently, as shown in Table 1, this has led to notable accuracy rates of 83.87%, 82.03%, and 79.76% in settings with 5, 10, and 20 clients, respectively. For the non-IID setup experiments with $\alpha = 0.1$, only $FedHARM_{rem}$ was utilized due to its superior speed and efficiency. This model demonstrated the best performance in the IID setting, thus it was tested on non-IID data as well. Similarly, the proposed methodology maintained robust performance, achieving significant accuracy results on CIFAR-10 and CIFAR-100 with 5, 10, and 20 clients, respectively, highlighting its adaptability to varied data distributions. This outcome underscores the effectiveness of integrating advanced neural network architectures to enhance the robustness and representational capacity of embeddings in distributed learning environments. The results are illustrated in table2

**Table 2:** Accuracy (%) comparison of FL Methods on CIFAR-10 and CIFAR-100 datasets in non-IID setup. The validation was performed on the test set of each database, and the resulting number is the mean accuracy across the clients.

| Method | CIFAR-10 | | | CIFAR-100 | | |
|---|---|---|---|---|---|---|
| | 5 Clients | 10 Clients | 20 Clients | 5 Clients | 10 Clients | 20 Clients |
| $FedCon$ [24] | - | **81.96** | - | - | - | - |
| $PerFCL$ [47] | - | 75.5 | 75.1 | - | **61.2** | **58.6** |
| $MOON$ [21] | - | 69.1 | 73.6 | - | 60 | 57.5 |
| $FedSimCLR$ | 64.06 | - | - | 38.70 | - | - |
| $FedCA$ [45] | 68.01 | - | - | 42.34 | - | - |
| $FedHARM_{\text{rem}}$(ours) | **82.91** | 81.97 | **76.84** | **43.78** | 38.98 | 37.42 |

**Table 3:** Accuracy (%) Results of the Proposed Method on the MNIST Dataset in IID and non-IID setup, against the FedCon system

| Method | IID | | | Non-IID | | |
|---|---|---|---|---|---|---|
| | 5 Clients | 10 Clients | 20 Clients | 5 Clients | 10 Clients | 20 Clients |
| $FedCon$ [24] | - | 98.08 | - | - | **98.22** | - |
| $FedHARM_{res}$ | 98.79 | 98.42 | 97.93 | 97.76 | 97.21 | 96.02 |
| $FedHARM_{rem}$ | **98.94** | **98.65** | **98.54** | **98.61** | 97.98 | **96.92** |

Due to the noted lack of comparable outcomes for the MNIST database, an assessment was carried out, as shown in Table 3, in order to compare the framework's performance with FedCon's in a 10-client setting. In this analysis, the $FedHARM_{rem}$ variation demonstrated superior performance, achieving an accuracy of 98.65% and 97.98% in IID and non-IID settings, respectively. Furthermore, in the IID setup, configurations involving 5 and 20 clients, $FedHARM_{res}$ variation recorded accuracies of 98.79% and 97.93%, respectively. In contrast, $FedHARM_{rem}$ exhibited even greater heterogeneity in its performance, achieving remarkable accuracies of 98.94% in the 5-client setup and 98.54% in the 20-client configuration. In non-IID setting, $FedHARM_{res}$ attained 97.76 and 96.02 for 5 and 20 clients, while $FedHARM_{rem}$ reached 98.61 and 96.92. This detailed comparison underscores the robustness and adaptability of the $FedHARM$ variations across different benchmark datasets.

## 5 Conclusion and Future Work

The study presents a comprehensive exploration of model architecture variability within the FL framework, introducing innovative strategies to incorporate diverse CNNs, specifically *ResNet*, $EfficientNet$, and *MobileNet*. It highlights the challenges and proposes solutions for efficient model aggregation and communication, emphasizing representation learning and model-agnostic frameworks. The proposed $FedHARM$ approach, especially $FedHARM_{rem}$, significantly outperforms existing methods in $CIFAR-10$, $CIFAR-100$ and $MNIST$ IID datasets evaluations, demonstrating the effectiveness of representation-centric and model agnostic aggregation across different architectures. This research paves the way for more adaptable, efficient, and privacy-preserving FL systems, capable of leveraging the strengths of different architectures to improve learning outcomes across decentralized networks. The study sets the stage for extensive future research, with numerous potential experiments to further enhance FL, involving an in-depth analysis into data heterogeneity among clients, and integrating a wider array of network architectures. Nevertheless, in practical FL scenarios, sampling shared data with a distribution similar to local data is challenging due to privacy protections and it necessitates several security checks to safeguard data integrity on local nodes.

**Acknowledgements.** This research was supported by the European Commission under contract No. 101121329 ARIEN.

## References

1. Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L.: Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. Journal of big Data **8**, 1–74 (2021)
2. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
3. Chowdhury, D., Banerjee, S., Sannigrahi, M., Chakraborty, A., Das, A., Dey, A., Dwivedi, A.D.: Federated learning based covid-19 detection. Expert Systems **40**(5), e13173 (2023)
4. Collins, L., Hassani, H., Mokhtari, A., Shakkottai, S.: Exploiting shared representations for personalized federated learning. In: International conference on machine learning. pp. 2089–2099. PMLR (2021)
5. Dave, I.R., Chen, C., Shah, M.: Spact: Self-supervised privacy preservation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20164–20173 (2022)
6. Doshi, K., Yilmaz, Y.: Federated learning-based driver activity recognition for edge devices. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 3338–3346 (June 2022)
7. Gkelios, S., Kastellos, A., Boutalis, Y., Chatzichristofis, S.A.: Universal image embedding: Retaining and expanding knowledge with multi-domain fine-tuning. IEEE Access (2023)
8. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006)
9. Hammam, A., Bonarens, F., Ghobadi, S.E., Stiller, C.: Identifying out-of-domain objects with dirichlet deep neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4560–4569 (2023)
10. Han, S., Park, S., Wu, F., Kim, S., Wu, C., Xie, X., Cha, M.: Fedx: Unsupervised federated learning with cross knowledge distillation. In: European Conference on Computer Vision. pp. 691–707. Springer (2022)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
13. Huang, W., Ye, M., Du, B.: Learn from others and be yourself in heterogeneous federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10143–10153 (2022)
14. Jang, J., Ha, H., Jung, D., Yoon, S.: Fedclassavg: Local representation learning for personalized federated learning on heterogeneous neural networks. In: Proceedings of the 51st International Conference on Parallel Processing. pp. 1–10 (2022)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)

17. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
18. Lee, S., Seong, H., Lee, S., Kim, E.: Correlation verification for image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5374–5384 (June 2022)
19. Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., et al.: A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering (2021)
20. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10713–10722 (June 2021)
21. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10713–10722 (2021)
22. Lin, T., Kong, L., Stich, S.U., Jaggi, M.: Ensemble distillation for robust model fusion in federated learning. Advances in Neural Information Processing Systems **33**, 2351–2363 (2020)
23. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
24. Long, Z., Wang, J., Wang, Y., Xiao, H., Ma, F.: Fedcon: A contrastive framework for federated semi-supervised learning. arXiv preprint arXiv:2109.04533 (2021)
25. Luo, M., Chen, F., Hu, D., Zhang, Y., Liang, J., Feng, J.: No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. Advances in Neural Information Processing Systems **34**, 5972–5984 (2021)
26. Mammen, P.M.: Federated learning: Opportunities and challenges (2021)
27. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
28. Meng, Q., Zhou, F., Ren, H., Feng, T., Liu, G., Lin, Y.: Improving federated learning face recognition via privacy-agnostic clusters. arXiv preprint arXiv:2201.12467 (2022)
29. Psaltis, A., Chatzikonstantinou, C., Patrikakis, C.Z., Daras, P.: Fedrcil: Federated knowledge distillation for representation based contrastive incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3463–3472 (2023)
30. Psaltis, A., Kastellos, A., Patrikakis, C.Z., Daras, P.: Fedlid: Self-supervised federated learning for leveraging limited image data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1039–1048 (2023)
31. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Advances in Neural Information Processing Systems **34**, 12116–12128 (2021)
32. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
33. Shahid, O., Pouriyeh, S., Parizi, R.M., Sheng, Q.Z., Srivastava, G., Zhao, L.: Communication efficiency in federated learning: Achievements and challenges (2021)
34. Shen, Y., Zhou, Y., Yu, L.: Cd2-pfed: Cyclic distillation-guided channel decoupling for model personalization in federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10041–10050 (2022)

35. Smith, L.N.: A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820 (2018)
36. Stäcker, L., Fei, J., Heidenreich, P., Bonarens, F., Rambach, J., Stricker, D., Stiller, C.: Deployment of deep neural networks for object detection on edge ai devices with runtime optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 1015–1022 (October 2021)
37. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning. pp. 6105–6114 (2019)
38. Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., Zhang, C.: Fedproto: Federated prototype learning across heterogeneous clients. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 8432–8440 (2022)
39. Tan, Y., Long, G., Ma, J., Liu, L., Zhou, T., Jiang, J.: Federated learning from pretrained models: A contrastive learning approach. Advances in Neural Information Processing Systems **35**, 19332–19344 (2022)
40. Tastan, N., Nandakumar, K.: Capride learning: Confidential and private decentralized learning based on encryption-friendly distillation loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8084–8092 (2023)
41. Wang, J., Guo, S., Xie, X., Qi, H.: Federated unlearning via class-discriminative pruning. In: Proceedings of the ACM Web Conference 2022. pp. 622–632 (2022)
42. Xia, H., Li, K., Ding, Z.: Personalized semantics excitation for federated image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19301–19310 (2023)
43. Yao, C.H., Gong, B., Qi, H., Cui, Y., Zhu, Y., Yang, M.H.: Federated multi-target domain adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1424–1433 (2022)
44. Yin, H., Mallya, A., Vahdat, A., Alvarez, J.M., Kautz, J., Molchanov, P.: See through gradients: Image batch recovery via gradinversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16337–16346 (2021)
45. Zhang, F., Kuang, K., Chen, L., You, Z., Shen, T., Xiao, J., Zhang, Y., Wu, C., Wu, F., Zhuang, Y., et al.: Federated unsupervised representation learning. Frontiers of Information Technology & Electronic Engineering **24**(8), 1181–1193 (2023)
46. Zhang, L., Shen, L., Ding, L., Tao, D., Duan, L.Y.: Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10174–10183 (June 2022)
47. Zhang, Y., Xu, Y., Wei, S., Wang, Y., Li, Y., Shang, X.: Doubly contrastive representation learning for federated image recognition. Pattern Recognition **139**, 109507 (2023)