Cohort-based Federated Learning Services for Industrial Collaboration on the Edge

Thomas Hiessl, Safoura Rezapour Lakani, Jana Kemnitz, Daniel Schall, and Stefan Schulte, *Member, IEEE*

Abstract—Machine Learning (ML) is increasingly applied in industrial manufacturing, but often performance is limited due to insufficient training data. While ML models can benefit from collaboration, due to privacy concerns, individual manufacturers cannot share data directly. Federated Learning (FL) enables collaborative training of ML models without revealing raw data. However, current FL approaches fail to take the characteristics and requirements of industrial clients into account. In this work, we propose a FL system consisting of a process description and a software architecture to provide FL as Service (FLaaS) to industrial clients deployed to edge devices. Our approach deals with skewed data by organizing clients into cohorts with similar data distributions. We evaluated the system on two industrial datasets. We show how the FLaaS approach provides FL to client processes by considering their requests submitted to the Industrial Federated Learning (IFL) Services API. Experiments on both industrial datasets and different FL algorithms show that the proposed cohort building can increase the ML model performance notably.

Index Terms—Federated learning, edge computing, collaborative AI, industrial collaboration, service-based architecture

1 Introduction

In recent years, Machine Learning (ML) has improved industrial manufacturing and process automation significantly, e.g., in fault classification, quality estimation, and soft sensing [1]. For instance, value-added and ML-based services like condition monitoring for production machines can be used to facilitate timely and cost-efficient maintenance actions throughout the lifetime of a machine [2], [3].

To achieve these benefits, high-quality ML models require a significant amount of training and test data. This data is often considered privacy-sensitive and needs to be protected from outside parties [4]. In addition, there is often a lack of ground truth data, which is referred to as the missing labeled data problem [5]. This is especially the case in industrial scenarios, since some labels can only be assigned to a dataset if critical and potentially rare events (e.g., a machine breakdown) are observed [6].

Since these large and labeled datasets cannot be shared with centralized servers for ML, a privacy-preserving way of knowledge sharing between collaborating devices is desired. This is the goal of Federated Learning (FL), as introduced by McMahan et al. [7]. FL is a recently emerged approach for transferring knowledge as model parameters (e.g., weights of neural networks) between edge devices without revealing raw data. For this, models are trained locally on edge devices and are then uploaded to an aggregation server that fuses model parameters, e.g., by averaging. After aggregation, the model is returned to the clients for evaluation. This process is executed repeatedly either until a pre-defined number of communication rounds or a provided level of quality (e.g., classification accuracy) is reached.

To optimize collective model training, various FL algorithms, e.g. [7], [8], [9], [10], have been researched focusing on model aggregation and client selection. Applying this concept for industrial machines that are distributed over multiple factories and facing heterogeneous environmental and operational conditions is referred to as Industrial Federated Learning (IFL) [11].

At this time, still many challenges [12] exist in FL which apply especially to industrial clients [11]. First, FL training processes typically start with selecting clients that are sampled by a server acting as central authority [7], [12]. Hence, single servers are used for orchestrating and monitoring the distributed training process managing connected edge devices [13]. So, in the majority of FL approaches the central authority defines the learning task by deciding e.g., on the used ML model, hyperparameters, and the FL algorithm [12]. However, in industrial applications, users (e.g., machine operators in production lines) have individual requirements regarding business partnerships when it comes to collaborations with other companies on improving and maintaining machine performance [11]. Hence, clients dependent on the learning task definition provided by the server are restricted to centrally managed ML models, therefore applications, and have no influence on selecting partners for FL. For example, pump manufacturers provide their products to customers (clients) from various industries like manufacturing, food and beverage, pharmaceutical engineering or water supply. Different use cases and therefore operational conditions are present in these domains, which may require adaptations of used ML models and hyperparameters for groups of clients. Moreover, even clients from the same field have restrictions to collaborate only with selected partners or with at least a defined minimum

- *T. Hiessl is a PhD student at TU Wien and is a research scientist at Siemens Technology in Vienna. Email: hiessl.thomas@siemens.com*
- *S. Rezapour Lakani and J. Kemnitz are data scientists at the Distributed AI Systems research group at Siemens Technology in Vienna.*
- *D. Schall is the head of the Distributed AI Systems research group at Siemens Technology in Vienna.*
- *S. Schulte is the head of the Christian Doppler Laboratory Blockchain Technologies for the Internet of Things at TU Wien.*

number of partners to increase the chance for actual ML model improvements. Hence, there is a need for a service-based system, empowering independent clients to create and submit ML models to the server and thereby enable FL as a Service (FLaaS) [14]. Using a FLaaS approach, allows a group of collaborating and independent clients to apply FL on these models and subsequently use it on their machines. Furthermore, considering client restrictions for collaboration partners, and the applied FL algorithm on the server provides flexibility to the clients.

Second, model updates for operating machines often needs to be triggered explicitly in industry under human supervision. This can be relevant in industry when ML models are used to assist human users [15]. Potential use cases are, e.g., condition classification of factory machines, failure detection, or even optimization of production processes. To update the used ML model with FL, a client may want to explicitly request participation in FL rounds, instead of automatically getting invoked in periodical execution plans by the server. This enables manual testing before deciding whether to use the model in production. Although, explicit participation in FL is not unique to IFL, this is significant for industrial applications.

Third, a prominent challenge in FL is the problem of heterogeneous data distributions [12], which is especially present in industrial domains when machines operate with different configurations under varying environmental and operational conditions. For instance, considering different liquids that are pumped in industrial processes, one can observe different error cases occurring over time detected by models using vibration patterns as input data [16]. Typically, one can observe that input data (e.g., vibration data) is varying across clients which is referred to as *feature distribution skew* [17]. Additionally, varying labels (e.g., error cases) can often be observed as well, which is referred to as *label distribution skew*. This phenomena corresponds to the non-IID (identically and independently distributed) data problem, which is a general issue in FL [12]. In non-IID settings, poor model quality can be observed by individual clients as the model is validated on their local data after FL was applied for all clients [12]. Therefore, FL systems need to provide mitigation strategies in implemented FL algorithms.

This paper addresses the aforementioned challenges of (i) enabling clients to individually and independently select used ML models and to define client criteria for collaboration in FL, (ii) enabling clients to explicitly participate in FL on an on-demand basis, and (iii) non-IID data distributions by proposing a FL system. Notably, the contributions are motivated and evaluated using scenarios from the industrial domain, but as discussed above, similar problems also occur in FL in general.

The contributions of this work can be summarized as follows:

- We present a FL process, the IFL process, for industrial clients including two algorithms dealing with individual and on-demand requests, and non-IID data.
- We design and implement a service-based system, the *IFL system*, covering the IFL process and providing FLaaS.
- We provide an evaluation of the IFL system using two time series-based industrial datasets. We present results on the model performance after FL and compare IID and non-IID scenarios.

We evaluated the IFL system on two industrial time series datasets, both providing several clients. We show that the IFL system considers on-demand participation of clients and yields significant improvements in classification accuracy applying FL on cohorts of similar clients rather than on the overall population.

The remainder of this paper is organized as follows: We describe the overall IFL system design in Section 2, and the system architecture in Section 3. In Section 4, we demonstrate the results of the FL experiments, and we review related work in Section 5. We conclude and outline future work in Section 6.

2 System Design

In order to present the design of our system, we first introduce the basic notation used in our work in Section 2.1. In Section 2.2, we introduce the IFL process and the algorithms, which are central to our approach of applying FL on cohorts of similar clients.

2.1 Basic Notation

To describe our IFL system model formally, we introduce the notation presented in Table 1. For this, we consider that a client $c \in C$ manages an asset $a \in A$ (e.g., a concrete heating pump) of a given asset type $type(a)$ (e.g, a centrifugal pump). Every asset type defines a data scheme $u(type(a))$ that needs to match with the data scheme $v(m)$ required by the ML model $m \in M$ that client c wants to train.

To participate in IFL, the i-th client c_i submits a task t_i to the IFL Server. For this, the client needs to specify asset a^{t_i}, model m^{t_i}, the individually selected FL algorithm $alg^{t_i} \in ALG$ for aggregating model weights, a cohort building algorithm $cb^{t_i} \in CB$, and federation criteria $crit^{t_i} \in CRIT$ before submitting t_i to the server.

The federation criteria $crit^{t_i}$ correspond to a set of requirements, where each need to be fulfilled by the system to consider t_i for upcoming FL rounds. So, we consider $t_i \in T$ with $T \subseteq A \times M \times CRIT \times CB \times ALG$.

A population p is a set of tasks that refer to the same asset type $type^p$, model m^p, FL algorithm alg^p and cohort building approach cb^p, where $p \subseteq T$ with $type^p = type(a^{t_i})$, $m^p = m^{t_i}$, $alg^p = alg^{t_i}$, and $cb^p = cb^{t_i}$ holds for all $t_i \in p$.

A cohort coh is a set of tasks with similar data distributions with $coh \subseteq p$. For this, the cohort building approach cb^p is used to assign every task within a population to a cohort. Furthermore, we consider COH^p as the set of all cohorts of population p. Hence, for $coh_1^p, \ldots, coh_{|COH^p|}^p \in COH^p$ it holds that $coh_j^p \subseteq p$ for all $j \in \{1, \ldots, |COH^p|\}$ with $coh_j^p \cap coh_k^p = \{\}$ and $j \neq k$.

Finally, FL is applied on population p using the FL algorithm alg^p to train one model per cohort $coh_j \in COH^p$. For this, we consider m^{coh_j} with $m^{coh_j} = alg^p(coh_j)$ for all $coh_j \in COH^p$.

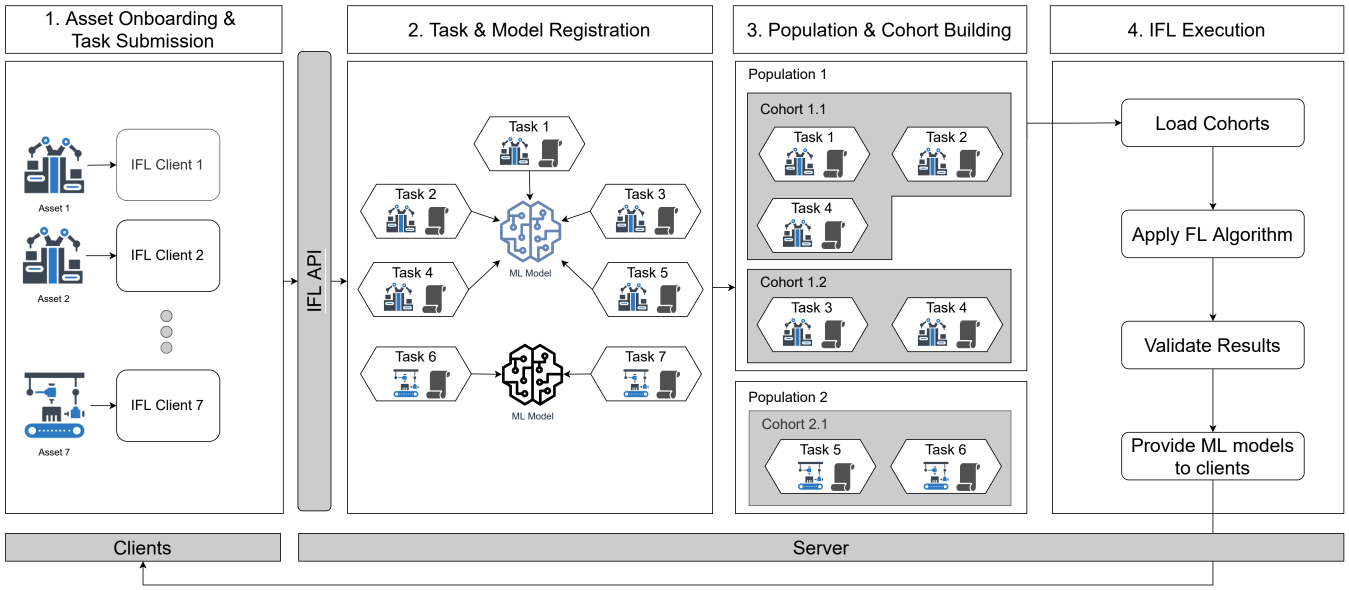


Fig. 1. IFL Process with 4 phases: 1. Clients are connected to their assets and submit tasks using the IFL API to participate in FL. 2. Submitted tasks are registered on the server referring to ML models used as base for FL. 3. Populations of tasks with same asset types are created. Cohorts further split populations in clusters of tasks with similar data distributions. 4. FL is executed for each cohort by applying the algorithm selected by the clients. Finally, validating results and providing the ML model to clients.

TABLE 1
IFL system entities

Notation	Description
C	Set of clients $c_i \in C$ participating in IFL
A	Set of assets $a \in A$
M	Set of ML models $m \in M$
$CRIT$	Set of federation criteria $crit \in CRIT$
CB	Set of cohort building approaches $cb \in CB$
ALG	Set of FL algorithms $alg \in ALG$
T	Set of tasks $t_i \in T$ submitted by clients
P	Set of populations $p \in P$
COH^p	Set of cohorts $coh \in COH^p$ of a given population p

2.2 IFL Process

In our solution, we address the discussed challenges and propose the *IFL process* depicted in Figure 1. The process is executed by the IFL system consisting of the *IFL Client* and the *IFL Services*. The client can be deployed on edge devices to train and operate ML models based on data generated by connected machines. The IFL Services offer an API to the clients providing knowledge aggregation and distribution on a central server.

To support FL for edge-based industrial clients, the IFL Client and services provide a four-step process depicted in Figure 1. As a prerequisite, we consider deployed client applications that invoke the IFL Client to establish connectivity to the IFL Services. Data is recorded from the asset and stored on the device. For this, we assume a classification problem with input data provided as matrix $X \subseteq R^{N_S \times N_V}$ with N_V variables and N_S samples, and a N_S-dimensional target vector $y \subseteq R^{N_S}$.

2.2.1 *Asset Onboarding and Task Submission*

The first step involves the IFL Client that needs to specify metadata that is referenced in a task. This metadata contains the used asset with the corresponding asset type, whereas the asset type can be reused, if other clients have already published this to the server. Similarly, a ML model is created or selected from the server to be applied to the data. This model is created upfront, based on the asset types' corresponding data structure and the ML task that needs to be solved.

Based on that, the client selects a cohort building approach. For this, the IFL Services provide two approaches. The first approach applies a cluster algorithm based on input data X to address potential feature distribution skew. The second approach clusters based on target data y to consider label distribution skew. In both cases, the respective cohort building approach is selected to reduce skewness within cohorts and to improve performance of models that are trained in a cohort by applying FL.

Furthermore, clients specify the knowledge aggregation algorithm, e.g., Federated Averaging [7], and individual federation criteria, e.g., the minimum number of clients in a population or minimum dataset size. This enables the client to control the knowledge aggregation process, e.g., to avoid that knowledge is transferred only between a small number of clients, which may lead to insufficient training results. Furthermore, this prevents that knowledge from an individual model is transferred to only a few other clients that may not contribute to the global model. To participate in IFL, clients then submit a task with the mentioned specifications to the server using the IFL API.

2.2.2 *Task and Model Registration*

In the second step, the server stores the defined assets, uploaded ML models and the tasks. Subsequently, the server verifies that the data scheme of the asset fits to the data scheme required by the referenced ML model, i.e., $u(type(a)) = v(m)$. Hence, it is ensured that clients can participate in FL rounds when the model is trained on their local data.

Algorithm 1 `PopulationCohortBuilding`

```
Input: new task $t$ received from client, populations $P$ stored on the server
    Update populations:
 1: $added_t \leftarrow False$
 2: for $p \in P$ do
 3:     if $type^p = type(a^{t_i})$, $m^p = m^{t_i}$, $alg^p = alg^{t_i}$, and $cb^p = cb^{t_i}$ then
 4:         $p \leftarrow p \cup t_i$
 5:         $added_t \leftarrow True$
 6:         if $crit^{t_i}$ holds for every $t_i \in p$ then
 7:             $COH^p \leftarrow CreateCohorts(p)$
 8:             break
 9:         end if
10:     end if
11: end for
12: if not $added_t$ then
13:     $p_{new} \leftarrow \{t_i\}$
14:     $P \leftarrow P \cup p_{new}$
15: end if

    CreateCohorts(p):
16: Initialize $F$ as $n \times m$ matrix for $n$ tasks and $m$ features
17: for $t_i \in p$ do
18:     if $cb^p$ = Target Distribution then
19:         $r \leftarrow \{mean(y_i^t), var(y_i^t), skew(y_i^t), kurt(y_i^t)\}$
20:     end if
21:     if $cb^p$ = Input Distribution then
22:         $r \leftarrow \{mean(X_i^t), var(X_i^t), skew(X_i^t), kurt(X_i^t)\}$
23:     end if
24:     add row $r$ to $F$
25: end for
26: for feature $f \in F$ do
27:     if $std(f) < \epsilon$ then
28:         remove $f$ from $F$
29:     end if
30: end for
31: $k \leftarrow elbow(F)$
32: $COH^p \leftarrow kMeans(F, k)$
33: return $COH^p$
```

2.2.3 *Population and Cohort Building*

The third step assigns tasks to populations and further splits populations into cohorts. This facilitates FL to be executed within small groups of similar clients. For this, we consider the server to execute Algorithm 1 to assign tasks accordingly as they are submitted by clients.

First, to build populations, we consider tasks with equal configurations. Particularly, we search for a population p with the same asset type, ML model, cohort building algorithm and FL algorithm as referred to in task t_i. This is necessary to consider a valid FL setting, whereas the same algorithms and model need to be applied on a common data scheme. If any population matches the task configuration as checked in line 3, the task is added. Otherwise, a new population p_{new} is created and added to the populations store P on the server (lines 13 and 14).

Furthermore, it is required to reach a consensus regarding the federation criteria, i.e., all criteria $crit^{t_i}$ have to hold for all tasks $t_i \in p$ as verified in line 6. For instance, we consider a criterion for requiring a minimum number of tasks in a population as a precondition before starting FL. To formally describe this exemplary federation criterion, let $q(t_i)$ be a function with $q(t_i) < |p|$ for all $t_i \in p$, where $q(t_i)$ returns the minimum number of tasks that is required by t_i. Based on that, the server eventually starts cohort creation in line 7 to improve model accuracy when FL is executed.

For this, we consider that data of individual clients can be non-IID. As we identified in [11], non-IID data can be observed when assets operate in heterogeneous industrial environments. Hence, in `CreateCohorts(p)`, the server makes use of aggregated data that describe the clients data distribution.

For this, we consider two cohort building approaches, i.e., *Target Distribution* and *Input Distribution*. Applying the former one, the server requests the client to compute the statistical moments mean, variance, skewness and kurtosis of the target data y^t in line 19. These measures provide information on the shape of the data's underlying distribution function, i.e., location (mean), dispersion (variance), asymmetry (skewness) and the form of tails (kurtosis). We used these features to capture the target distribution of every client's dataset as precise as possible, which is the basis for accurately assigning the corresponding task to a cohort of tasks with similar target distributions. Therefore, we can reduce label distribution skew in a cohort. In line 24, the features are returned to the server and added as a new row to the feature matrix F.

The *Input Distribution* approach computes the mentioned moments based on all n variables of the input matrix X^t (line 22) and adds them to F. This feature matrix F consists of $|p|$ rows and u columns, where u is the number of features. Hence, using *Target Distribution* we have $u = 4$, while for *Input Distribution* we consider $u = 4 \times n$. Analagous to *Target Distribution*, using *Input Distribution* we can reduce feature distribution skew.

Since F may contain many features with limited information, we remove all features f from F with $std(f) \leq \epsilon$ in line 28, where $std(f)$ computes the standard deviation and ϵ is a pre-configured parameter on the server.

To eventually create the cohorts COH^p, we apply the *k-Means* [18] cluster algorithm based on the reduced feature matrix F. The *k-Means* cluster algorithm is an iterative approach that considers a fixed number of k clusters, whereas data points are assigned to cluster centers that minimize variance within clusters. Optimally, we want clusters where all the data in a cluster are close to each other, and the distance between two clusters is as large as possible. To identify k, we apply the *elbow* method [19], [20] in line 31 to find the optimal value with respect to the *silhouette* [21] values. In line 32 we finally apply k-means with the identified k to assign all tasks $t_i \in p$ to cohorts COH^p. We used the combination of k-means and the elbow-method since it can autmatically be applied by the IFL Service without the need for human validation of the number of built cohorts and still avoiding under- and over-fitting of the trained cluster model.

2.2.4 IFL Execution

The fourth and final step in the IFL process applies the selected FL algorithm alg^p on all created cohorts COH^p as described in Algorithm 2. To start the execution, the server loops through all $coh \in COH^p$ and initializes the models $m_0^{t_i}$ for all tasks $t_i \in coh_j$ in line 2. For this, the model architecture of the underlying neural network is created by building all layers as defined in the base model m^p of the population.

To actually share knowledge between the involved clients, FL algorithms can be invoked by the server. In this work, we evaluate the integration of two FL algorithms in the IFL Services to be applied on cohorts, i.e., *Federated Averaging* (*FedAvg*) [7] and *Sequential FL* (*SeqFL*) [22] We selected the two algorithms since they are well-established and often considered as a benchmark for FL.

Both algorithms invoke clients by iterating over their issued tasks t_i of the processed cohort coh_j, and calling the function `ClientUpdate(t,m)` on the client (lines 6 and 12). The clients individually train a model that is passed to them by the server in `ClientUpdate(t,m)` as defined in line 18. For this, in lines 19-24, the dataset X^{t_i} is split into batches B and is iterated several times as defined in the number of epochs E. In line 22, the model is updated using one step of gradient descent optimization, considering a loss function l and a learning rate $alpha$. After iterating the defined number of epochs, the optimized model is returned to the server in line 25, to exchange the gained knowledge with other clients.

In FedAvg, this is achieved by summing up model weights from all clients of a given round and dividing it by the number of tasks in the cohort $|coh_j|$ as presented in line 8. In our approach, we integrated a simplified equally weighted averaging, whereas in the original approach client models are weighted with an additional factor expressing the proportion of the number of examples N_{S_i} of client c_i with respect to the total number of examples. In the next round, the aggregated model is again distributed to all clients.

In SeqFL, the model is passed sequentially from one client to another, whereas the subsequent client optimizes the model that the previous client has optimized before. The result of the last client in a given round r is used as input for the first client of round $r+1$ (line 14).

After training using either of the aforementioned algorithms, the resulting model $m_R^{coh_j}$ is validated by involved clients on their local test datasets as stated in line 17. This yields validation metrics (i.e., test set accuracies) that are passed to the clients along with the model, which concludes the IFL execution process. As a result, the client is enabled to decide whether to operate the IFL-based model or an individually trained model on the edge device.

Algorithm 2 `IFL Execution`

```
Input: $COH^p$ received from population and cohort building step
   Server execution:
 1: for $coh_j \in COH^p$ do
 2:   initialize $m_0^{t_i}$ for all tasks $t_i \in coh_j$
 3:   for round $r = 1, ..., R$ do
 4:     if $alg^p$ = FedAvg then
 5:       for task $t_i \in coh_j$ do
 6:         $m_{r+1}^{t_i} \leftarrow ClientUpdate(t_i, m_r^{t_i})$
 7:       end for
 8:       $m_{r+1}^{coh_j} \leftarrow \frac{1}{|coh_j|} \sum_{i=1}^{|coh_j|} m_{r+1}^{t_i}$
 9:     end if
10:     if $alg^p$ = SeqFL then
11:       for task $t_i \in coh_j$ do
12:         $m_r^{t_{i+1}} \leftarrow ClientUpdate(t_i, m_r^{t_i})$
13:       end for
14:       $m_{r+1}^{coh_j} \leftarrow m_{r+1}^{t_0} \leftarrow m_r^{t_{|coh_j|}}$
15:     end if
16:   end for
17:   validate and send $m_R^{coh_j}$ to all clients $c_i$ of tasks $t_i \in coh_j$
18: end for

   ClientUpdate(t,m) // Run task $t$ on client
19: $B \leftarrow (X^t$ split into batches of size $B)$
20: for epoch $e \in 1..E$ do
21:   for batch $b \in B$ do
22:     $m \leftarrow m - \alpha \nabla l(m, b)$
23:   end for
24: end for
25: return $m$ to server
```

3 System Architecture

To run the IFL process described in Section 2.2, the IFL system provides the service-based architecture depicted in Figure 2. In this FLaaS approach, we consider two main locations, i.e., the shop floor and the IFL Server.

On the shop floor, assets are operated and generated data (e.g., measured vibrations) are sent to an *Edge Device* to train a ML model that can be used for, e.g., condition monitoring. The IFL Server consists of the *IFL Services* that provide an interface to onboard assets, submit tasks, and apply FL on cohorts of clients. This architecture allows multiple clients from different locations (i.e., shop floors) to use the IFL Services and participate in FL. This can be useful to provide FL to multiple sites within a company or even to multiple companies that aim for collaborating on common ML problems.

3.1 IFL Client on Edge Devices

To support the asset onboarding and tasks submission step in the IFL process as described in Section 2.2.1, the *IFL Client* is deployed to Edge Devices. In particular, the IFL Client is invoked by the *Industry Application* to participate in FL. The Industry Application can include arbitrary business logic that makes use of FLaaS to train ML models on recorded asset data. To support this, the IFL Client connects to the IFL Services and creates an IFL task, which triggers subsequent IFL process steps yielding a trained model at the end.

As described in the IFL process, the IFL client needs to specify relevant metadata as the used asset, ML model etc. beforehand. This metadata is stored along the used dataset on a local data storage to ensure transparency on the history of trained ML models and involved assets.

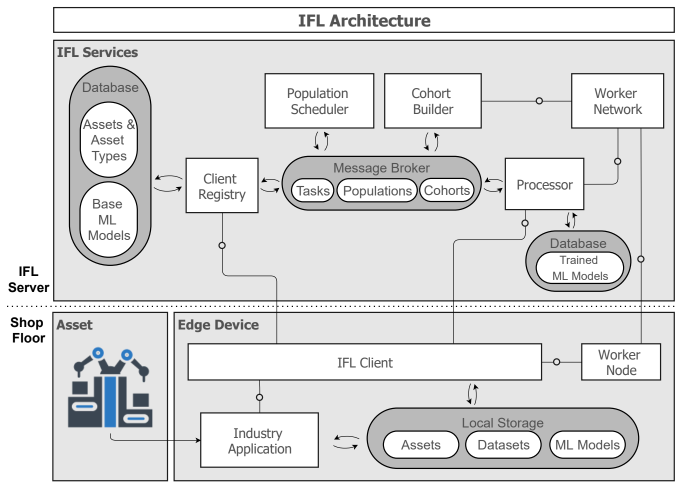


Fig. 2. IFL Architecture: Edge-based IFL Client and IFL Services

After submitting a task, the IFL Client starts a local *Worker Node* that can run e.g., the `ClientUpdate(t,m)` function (see line 18 in Algorithm 2). This function is called by the server in the IFL Execution step as described in Section 2.2.4. For this, the IFL Client provides the training and test dataset to the Worker Node to optimize the model m that is received from the server in the respective communication round. As a prerequisite, the Worker Node needs to register at the Worker Network that is located on the IFL Server. This enables that Worker Nodes can be found, as a lookup on the server is initiated to invoke clients. Worker Nodes are relevant in this architecture, since they enable isolated training per task. This training can even be outsourced to trusted Edge Devices by spawning Worker Nodes on remote locations. This can be useful if resources (e.g., CPU, Memory, GPU) are fully utilized on a given Edge Device, but still further training is planned.

3.2 IFL Services

To provide FLaaS, the IFL process is supported on the server side, which considers *Client Registry*, *Population Scheduler*, *Cohort Builder*, and the *Processor* as independent services. We further consider a database for assets, associated asset types, and ML models. The latter are considered as base ML models, reflecting a deep learning neural network architecture with untrained weights.

Using the Client Registry, the ML models can be created by clients and used by other clients by referring to them in the submitted tasks. For this, the Client Registry provides an API that accepts assets, asset types, ML models and tasks. The latter are validated with respect to a matching data scheme of referred assets and the data scheme that is required by the ML model before eventually saving it to the *Message Broker*. This broker supports a publish/subscribe messaging architecture to share processed output (i.e., validated tasks, built populations, and cohorts) between the services. This asynchronous communication enables loosely-coupled services, whose instances can be scaled up and down considering varying loads of task submissions.

To support the population and cohort building as described in Section 2.2.3, the population scheduler consumes tasks from the Message Broker and assigns them to a population. If provided federation criteria hold for all clients, the population is published to the Message Broker for cohort building. Hence, the Cohort Builder is triggered to create cohorts that are again published to the Message Broker. However, to create cohorts, the statistical moments of the underlying data distributions are needed, as presented in Algorithm 1. For this, the Cohort Builder makes use of the Worker Network to access registered Worker Nodes assigned to the clients. In this approach, the Cohort Builder submits a query to the Worker Network to retrieve references to the Worker Nodes. The query is used to retrieve only Worker Nodes of clients that have submitted tasks to the currently processed population. This is useful since the selective approach does not invoke clients of other populations blocking their resources.

To address the IFL Execution process step, as described in Section 2.2.4, the Processor service subscribes to created cohorts on the Message Broker. Hence, the cohort is loaded and the seleted FL algorithm is applied by again accessing the Worker Network. For this, the Processor sends the ML model to respective Worker Nodes and retrieves the updated model after training. After the last communication round, the trained model is validated on test data by the Worker Nodes. The validation metrics are stored along with the trained ML model in a database. To deploy the model to the shop floor, participating clients can download the model as provided by the Processor API.

4 Evaluation

In this section, we show the applicability of the IFL process on different datasets including industrial data. We evaluate our IFL system approach on IID and non-IID datasets.

In the following, we explain our evaluation setup (Section 4.1). The characteristics of the datasets used for our evaluation and the scenarios designed from these datasets are described in Section 4.2. Our experimental design is explained in Section 4.3. Finally, the evaluation results based on the designed scenarios are reported in Section 4.4.

4.1 Experimental Setup

All the experiments are run on a Windows machine with 3.0 GHz Intel Xeon Scalable Processor with 8 CPU cores and 32 GiB RAM, hosting the IFL Services. IFL Clients run on a Windows machine with 3.3 GHz Intel Xeon Scalable Processor with 2 CPU cores and 4 GiB RAM.

The implementation of the IFL system uses PySyft[1], an open source framework for private deep learning to operate on data that resides on remote locations, i.e., the server invokes PySyft to connect to the clients and to train ML models on their local datasets. The Worker Nodes and Worker Network communication, as desribed in Section 3, is implemented using PyGrid[2]. PyGrid is based on PySyft and adds the functionality for registering nodes and searching for nodes in a network to eventually establish connection between the IFL services and the Worker Nodes on the Edge Devices.

1. https://github.com/OpenMined/PySyft
2. https://github.com/OpenMined/PyGrid

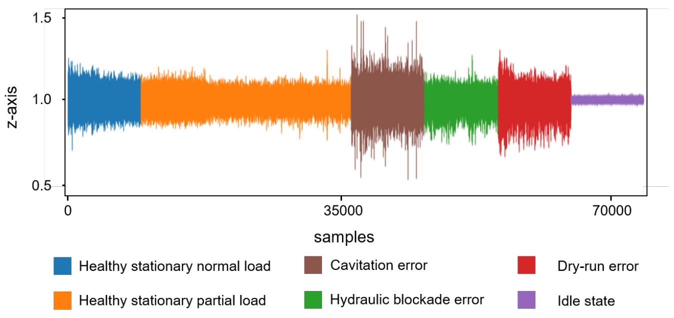


Fig. 3. Pump classification dataset: a schematic view of samples from one sensor. Only the virtual Z-axis of the sensor is shown here. Each color indicates an anomalous (hydraulic blockade failure, dry-run error, and cavitation error) or a healthy (healthy stationary normal load, healthy stationary partial load, and idle state) pump condition.

4.2 Datasets

We evaluate our proposed approach using two real-world datasets from industry. In this section, we explain the characteristics of each dataset and elaborate on the ML approaches used on these datasets.

4.2.1 Pump Condition Classification Dataset

This dataset contains acceleration data from four pumps. The data is obtained using an Industrial Internet of Things (IIoT) sensor, namely the *SITRANS multi sensor* [23]. This sensor provides 512 acceleration samples at three dimensions every minute, thus providing a time series representation. In order to ensure that the model is not affected by the orientation of the sensor that is mounted on the pumps, a virtual sensor alignment based on the Kabsch algorithm [24] is performed.

The time series data is labeled based on the conditions of the pumps. Figure 3 shows a schematic view of the virtual Z-axis from 70,000 samples obtained from one of the sensors. As it can be seen, there are six classes indicating anomalous or healthy conditions. The healthy conditions are healthy stationary normal load (the load of pumped water at the range of $50\frac{m^3}{h}$), healthy stationary partial load (the load is $[12.5\frac{m^3}{h}, 37.5\frac{m^3}{h}]$), and idle state (the pump is turned off). The anomalous conditions are hydraulic blockade failure (nothing was pumped at the start of pumping), dry-run error (nothing was pumped at the end of pumping), and cavitation error (the water in a pump turns to a vapor at low pressure).

We follow a sliding window approach for collecting data with a window size of 512 and a step size of 256. We then extract Mel-frequency Cepstral Coefficients (MFCC) from the samples in each window. MFCC coefficients are widely used as features for audio classification. We employ it here for acceleration data as both having strong relation due to air-borne and structure-borne sound transmission [25]. The MFCC represent the power spectrum of the short-term Fourier transform on a nonlinear frequency scale inspired by human biology uniformly spaced below 1 kHz and a logarithmic scale above 1 kHz. The resulting 20 MFCC coefficients are then considered as our features.

This dataset is suitable to demonstrate how clients with real-world data, generated by connected assets (i.e., pumps), and a condition monitoring ML model can be integrated into the IFL process to support collaboration within cohorts of distributed but similar clients.

Hence, we consider two evaluation scenarios for this dataset: 1) *pump classification IID* where the distribution of target data (i.e., pump conditions) is IID, the data is shuffled and partitioned into 9 clients each receiving on average 26,554 samples and 2) *pump classification non-IID* where the data is shuffled and partitioned into 10 clients each receiving on average 212,738 samples but with a non-IID target distribution. In both cases, we have imbalanced data.

As a learning approach, we use an Artificial Neural Network (ANN) with two dense layers with 64 units each using ReLu activations and followed by 40% dropout, and a last output layer without any activation function (5,894 total parameters). We use this dropout-rate to avoid that the model is biased towards a single client. We consider just two dense layers to limit the number of parameters, which affect training time and would require larger datasets. However, to facilitate classification (i.e., linear separability), the number of units per layer (64) is chosen to be significantly greater than the input features (20). This model is then considered as our base model m^p that is uploaded to the server and used by tasks of population p for evaluation.

The data is normalized using a Gaussian distribution before being fed into the model. Therefore, the client data is adjusted to a common scale, which is necessary to train a common model. The transformation can be described as:

$$z = \frac{x - \mu}{\sigma}, \tag{1}$$

where μ and σ are the mean and standard deviation of training samples, x stands for input sample (i.e., MFCC features), and z indicates the normalized samples. The normalization parameters μ and σ are computed for each client training data separately and have been applied on the respective client test data.

4.2.2 Material Classification Dataset

This dataset contains vibration data and material hardness labels during machining metals from two machine tools. The vibration data is obtained at a frequency of 1Hz, thus providing 60 samples per minute. The objective here is to classify material hardness based on vibration data. As it can be seen in Figure 4, there are six material hardness classes in this dataset.

It can be observed that some materials are only machined from one of the machine tools, thus the distribution of hardness labels between the machines is not uniform. Furthermore, the dataset is imbalanced, for example, the material with hardness label A has been observed more than the other material hardnesses. FL is reasonable and applicable for this dataset because 1) both machines are similar in construction and 2) non-IID distribution of material hardness labels makes transferring one model for both machines not applicable.

In this use case, we demonstrate how IFL can be applied on a small scale, considering only two assets (i.e., machine tools) generating non-IID data, and still enabling ML model improvements.

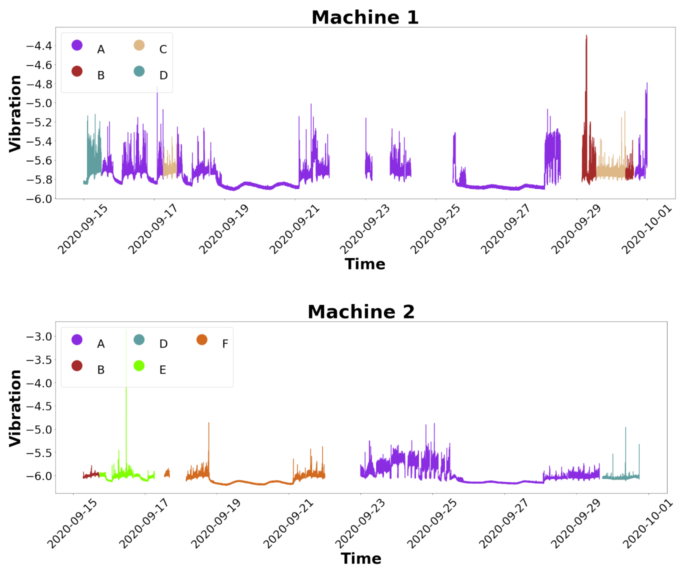


Fig. 4. Material hardness labels for two machine tools. Each row shows vibration time series data for one machine, colors indicate different material hardness.

The input vibration time series data is split into sequences of maximum two continuous hours during the operation time of the machines. These sequences are then divided into training and test data, keeping 70% of sequence data for training. Our data is obtained from these sequences following a sliding window approach, with a window size of 120 (i.e., samples of every two minutes). For training sequences, we have overlapping windows to cover the entire data with a stepsize of 60, whereas for testing sequences, we do not use any overlapping. We then compute Fast Fourier Transform (FFT) [26] on the samples and consider the magnitude of the computed FFTs as our features. This yields a 120 dimensional feature vector for each training data.

Although, ANNs can be used as feature learners as well, we used FFT features, to reduce the number of parameters that need to be trained according to Lin et al. [27]. In general, we used ANNs in all scenarios since models can easily be aggregated, e.g., by averaging parameters [7].

We consider one evaluation scenario for this dataset using two clients, one for each machine tool. The target data distribution (i.e., material hardness) is non-IID and the data is imbalanced. We have a total number of 5,985 samples for the first client and 7,136 samples for the second client.

As a learning approach, we employ an ANN with two dense layers each with 256 units, each using ReLu activations and followed by 40% dropout, and a last output layer without any activation function (98,310 total parameters). This model is then considered as our base model in FL.

As in the pump classification dataset, we choose this configuration to avoid biased models, and to facilitate linear separability. Likewise, we normalize data to consider a common scale for the two clients.

4.3 Experimental Design

All the scenarios as discussed in Section 4.2 are provided in a JavaScript Object Notation (JSON) format. The scenario JSON contains configuration for clients, tasks, assets, and the used model.

The client configuration provides settings for the client's name, the path to a client dataset, the tasks associated with a client, and additional organizational descriptions.

The task configuration has settings for the selected FL algorithm alg^{t_i} (FedAvg or SeqFL), cohort strategy cb^{t_i} (if cohorts of similar clients should be built for federation and the algorithm to be used for building cohorts), and federation criteria $crit^{t_i}$ (e.g., the minimum number of required clients that need to join a population to start the training).

The asset configuration gives settings for name, description, location, and environment description of an asset.

The model configuration provides settings for the path of the base model to be used for federation, and the training parameters (e.g., number of communication rounds, number of epochs per communication round, learning rate, batch size, etc).

Based on the aforementioned configuration parsed from the scenario JSON, the individual client processes independently create their assets, refer to a common ML model, and eventually submit tasks to the Client Registry API to join for FL.

After FL is finished, resulting ML models are validated using the classification accuracy. If applicable, we consider classification accuracy on cohort test data (i.e., test data from all the clients in a cohort) to ensure comparability between models trained only on client data, models trained on central data, and models trained with FL on cohorts. To further compare cohort-based FL with FL not using cohort building (i.e., FL algorithm is applied on the overall population), we consider classification accuracy on the overall population. Finally, the clients download the model using the Processor API to conclude the IFL process which terminates the evaluation scenario.

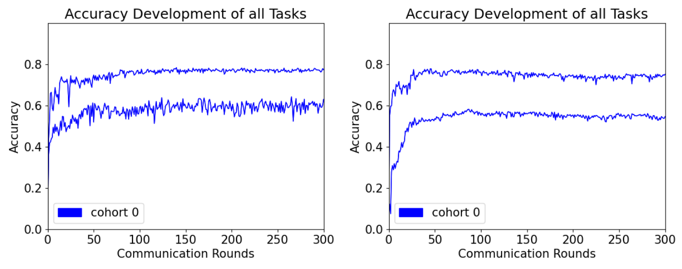
4.4 Results

In this section, we report on experiments evaluating the performance of our IFL system on multiple scenarios from three datasets (as described in Section 4.2) using FedAvg and SeqFL algorithms.

In all the experiments, our base model for training is an ANN model. We use Pytorch [28] for training and prediction. We compute loss using cross entropy loss function [29]. Since we deal with imbalanced data, we use a weighted cross entropy loss function. The loss l is computed for each class c separately and can be written as:

$$\begin{aligned} l(x,c) &= weight[c] \times -\log \frac{\exp x[c]}{\sum_j \exp x[j]} \\ &= weight[c](-x[c] + \log \sum_j \exp x[j]) \end{aligned} \tag{2}$$

where x is an observation (i.e., the output logit of an ANN for a sample whose size is equal to the number of classes), and $weight[c]$ is the class weight computed for each client independently considering a balanced heuristic inspired by [30].



(a) FedAvg (b) SeqFL

Fig. 5. Accuracy of federated model after each communication round on the material classification dataset.

The class weight can be described as:

$$weight[c] = \begin{cases} \frac{N_S}{N_C \times N_{S_c}}, & \text{if } N_{S_c} > 0, \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where N_S is the total number of samples for each client, N_C is the total number of classes provided as a training parameter to each client because there can be classes which are not present for a particular client, and N_{S_c} is the number of samples for this class in a client dataset.

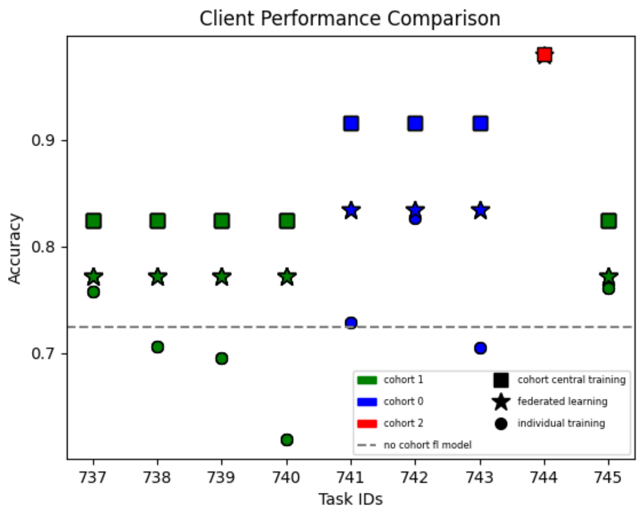
We perform batch optimization for training where training loss is computed on batches of data instead of the entire data due to memory efficiency and overfitting problem. For optimization, we use Adam optimizer [31] and a batch size of 128. In order to compromise between time and performance, we keep the learning rate $1e-3$ in our experiment. A lower learning rate makes training much slower and a higher learning rate might deteriorate the performance.

4.4.1 *Impact of IFL on Non-IID Data*

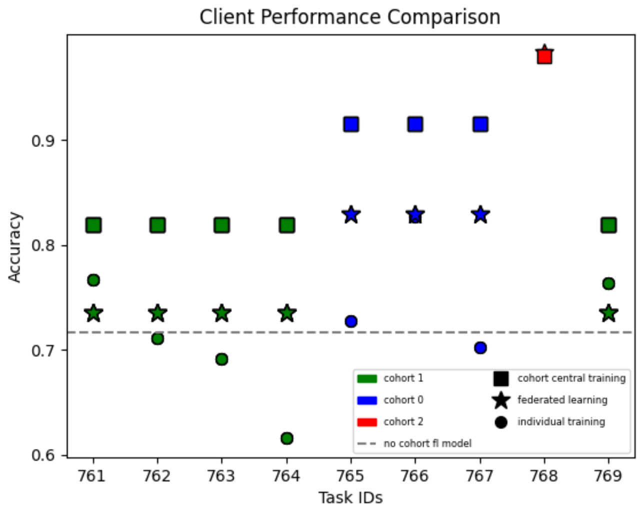
The material classification dataset as discussed in Section 4.2.2 has non-IID data and target distribution. We provided two scenarios for each FL algorithm with a default cohort strategy (when no cohort is built) in JSON format. We trained a federated model with two clients (one for each machine tool). The training is done for 300 communication rounds and one epoch per communication round for each client. The average accuracy on the test data of this dataset is shown in Figure 5. As it can be seen, the accuracy of the federated model improves after each communication round from an initial accuracy of 5%. We obtain for both clients a slightly higher accuracy using the FedAvg algorithm. But in both cases, we get the best accuracy of 76% in spite of non-IID data distribution which underlines the usability of our IFL system.

4.4.2 *Impact of Cohort-based IFL on IID Data*

We have shown the applicability of our IFL system on industrial data with two clients. Next, we evaluate our approach on *pump classification IID* dataset (Section 4.2.1) with 9 clients where the data has an IID target distribution. We made two scenarios in JSON format per FL algorithm using cohort strategy based on input data distribution. Our federation criteria in both scenarios is the minimum number of clients for starting federation which we set as 9 (the total number of clients).



(a) FedAvg



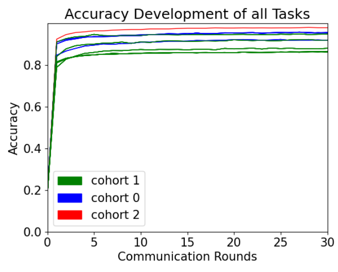
(b) SeqFL

Fig. 6. Client performance comparison on the pump classification dataset with IID target distribution. Accuracy of FL model is compared with individual training, cohort central training, and global FL on cohort test data.

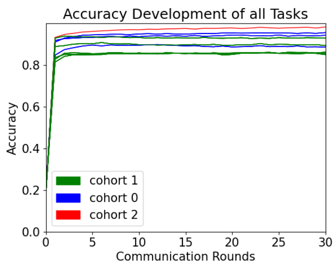
Figure 6 shows evaluation results of this scenario after 30 communication rounds and five epochs per communication round. The performance of the federated model is compared with three other evaluation approaches:

- No cohort IFL model: A federated model is trained on the training data of all the clients and tested on the test data of all the clients.
- Cohort central training: A model is trained on the training data of all the clients in the same cohort and is tested on cohort test data.
- Individual training: A model is trained on each client and transferred to the other clients (i.e., testing on cohort test data).

It can be seen in Fig. 6 that the accuracy of the cohort-based FL model is higher than the global FL model. The cohort building approach in this experiment is based on input data. The result as depicted in Fig. 6 indicates that, however the target distribution is IID, but the input data



(a) FedAvg



(b) SeqFL

Fig. 7. Accuracy of federated model after each communication round on the pump classification dataset for 9 clients with IID target distribution.

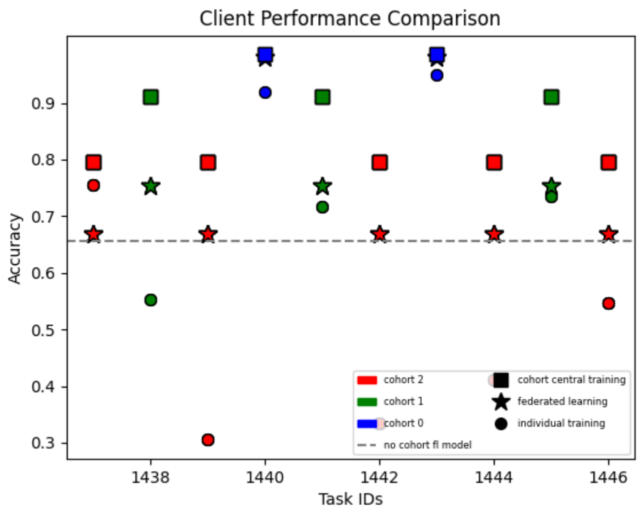
of pump conditions does not have IID distribution. More precisely, the input data of some pump conditions are more similar to each other than other pump conditions. Therefore, we obtained a higher performance by learning separate FL models on similar groups of clients. This result highlights our contribution using a cohort-based IFL approach with either FedAvg or SeqFL algorithms.

The best result achieved using federation on each cohort is shown in Fig. 6 as cohort central performance. It can be observed that we can achieve an accuracy close to the central cohort performance using our IFL system. In order to show the importance of using FL in this scenario, we compared the accuracy of federated model on each cohort with individual training. We can observe that our federated model on average, improves the accuracy of each client compared to the individual training. Using FedAvg, this improvement can be seen more clearly than for SeqFL. This emphasizes the applicability of FL for this experiment.

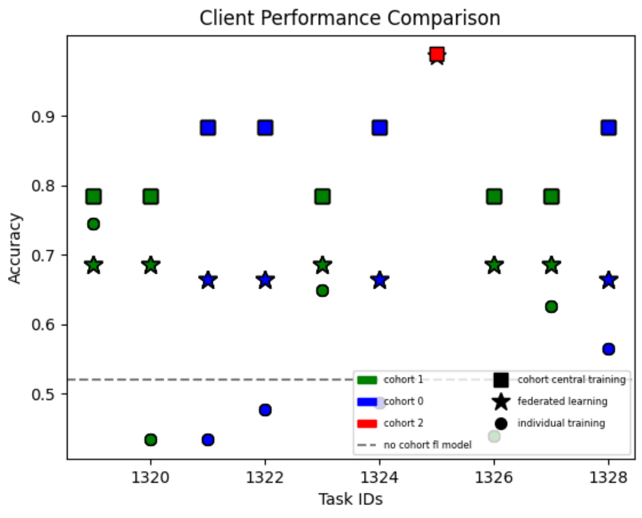
The average accuracy on the test data after each communication round is shown in Figure 7. It can be seen that the federated model reaches to an accuracy above 80% just after a few communication rounds. We can see that cohort 3 has the highest accuracy than the other cohorts. The reason is that this cohort has only one client, therefore FL performs like a central model (without federation) for this model. The accuracy of FL for the clients in cohort 1 is slightly lower than others. As it can be seen in Fig. 6, the central cohort accuracy for this cohort is also lower than the other cohorts. Therefore, FL cannot achieve a higher accuracy. One reason for the lower performance of cohort 1 is that, it has more clients compared to the other cohorts which training a suitable model more difficult.

4.4.3 Impact of Cohort-based IFL on Non-IID Data

We showed that cohorts can improve the performance of IFL system on datasets with IID target distribution. We go one step further and evaluate the performance of a cohort-based IFL system on the *pump classification non-IID* (Section 4.2.1) with non-IID target distribution with 10 clients. For this experiment, we also provided two JSON scenarios, each scenario considers one FL algorithm. The cohort strategy in both scenarios is based on input data distribution. And the federation criteria in both scenarios is the minimum number of clients for starting federation which is set as 10 (the total number of clients). Figure 8 shows the evaluation results after 30 communication rounds and five epochs per



(a) FedAvg

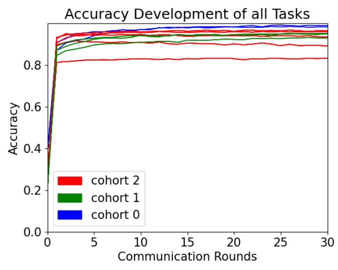


(b) SeqFL

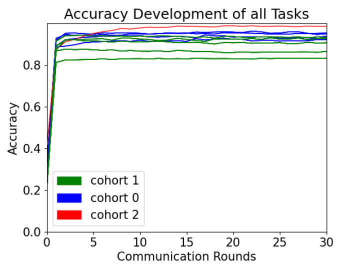
Fig. 8. Client performance comparison on the pump classification dataset with non-IID target distribution. Accuracy of FL model versus individual training, cohort central training, and global FL on cohort test data.

communication round. It can be seen that the federated models achieve a high accuracy compared to their respective central cohort models. Furthermore, the cohort-based approach gives on average a higher performance than the no cohort federated model. This result also emphasizes the impact of finding similarity between clients and using them in federation.

Figure 9 shows the average accuracy on the client test data of this dataset after each communication round. As it can be observed, the federated model reaches at least 80% accuracy just after a few communication rounds. It can be seen that in cohorts with more clients such as cohort 2 in Fig.9(a) or cohort 1 in Fig.9(b), the accuracy of some clients is lower than the other clients in the same cohort. The reason is that we used the individual client test datasets, a k-Means clustering approach for building cohorts. The clusters are initialized randomly and the clients get assigned to the clusters with the minimum mean squared distance. The random cluster initialization may result in outliers during



(a) FedAvg



(b) SeqFL

Fig. 9. Accuracy of federated model after each communication round on the pump classification dataset for ten clients with non-IID target distribution.

assignment.

4.5 Discussion

As we have demonstrated in the evaluation, the IFL process with the integrated cohort building and FL algorithms can be applied to improve ML models for industrial scenarios. Our FLaaS-based approach provides FL to client processes by considering their IFL tasks and defined metadata as described in 4.3. So, the clients explicitly submit the tasks using the Client Registry API, and download the resulting model using the Processor API. This invocation of the service-based architecture and the IFL process addresses the challenges of (i) individual, independent, and (ii) explicit participation in FL with (iii) non-IID data. The individual implications of the presented solutions are discussed in the following:

4.5.0.1 Impact of datasets on defining clients: The respective experiments are limited only to two industrial datasets with a small number of assets and finding may not be generalizable. For example, the pump condition classification dataset only contains acceleration data from four pumps, but resulted in 9-10 different clients. The reason is that those pumps were completely dismantled and rebuilt several times to achieve the most realistic possible conditions for creating 9-10 clients with regard to feature and label skews. Furthermore, the measurements and conditions were recorded by different experts and due to the asset complexity, partial load and error creation resulted in additional data skews.

4.5.0.2 Impact of cohort-based FL: Our experimental evaluation proves the applicability of our cohort-based IFL system especially on non-IID data. It can be seen that the accuracy of a cohort-based approach is on average higher than for FL models without considering cohorts. However, a limitation is as shown in Figures 9(a), 9(b), that the clusters are initialized randomly and the clients get assigned to the clusters with the minimum mean squared distance. The random cluster initialization may result in outliers during assignment. Providing a dynamic cohort-assignment approach, that enables changes in cohort assignment between communication rounds, is left as our future work. It is relevant to evaluate scenarios which consider clients adopting and rejecting FL models of the cohort, and switching cohorts on-demand. Similarly, it can be desirable for industrial clients to incorporate model updates from individually selected clients that are not necessarily in the same cohort, e.g., by defining a query to select industrial partners.

4.5.0.3 FL versus model transfer: The results reported on industrial datasets (i.e., pump classification and material classification datasets), indicate that FL is applicable and reasonable for these datasets. Based on our results, transferring a model from one client to the others is not applicable for these datasets and does not give a higher accuracy than FL. However, as it can be seen in Figure 6(b) the individual training result of cohort 1 for task 761 performs slightly better than the FL model. The potential reason might be that this client has either more data or more variety of data in this cohort. Making an adaptive approach, for deciding if FL or model transferring should be used in a cohort is left as our future work.

4.5.0.4 Impact of FL algorithm: FedAvg performed better than SeqFL for all the scenarios. It can be explained that in the FedAvg algorithm, the parameters from all the clients are aggregated (i.e., averaged), therefore the aggregated model fits better to all the clients. At the moment, both of the algorithms are implemented without parallelization. Parallelization would reduce training times significantly. FedAvg is the only suitable algorithm for parallelization among the both, thus more appropriate for our IFL system. Due to small number of assets in our experiments, the parallelization aspect would have neglectable impact on training time, therefore is not considered currently. But it will be considered in our future work with more assets.

4.5.0.5 Impact of resource-constrained edge-devices on FL: In our evaluation all the clients are running on machines with similar resource capabilities, therefore yet we did not consider resource constraints and availability of edge devices in the IFL process. Since our current focus was on the applicability of the IFL system on edge devices, we considered optimizing accuracy rather than optimzing resources. Our IFL services will be extended in the future to overcome this limitations.

5 Related Work

Despite being a very recent research topic, FL has been studied widely in various applications [32], [33], [34]. The basic idea of FL as proposed in [7] is to federate a global FL model among all the clients. However, one of the main challenges in FL is that the client's data especially in real scenarios are non-IID. Thus, a global FL model might not perform well for all the clients.

To overcome this problem, clustered FL approaches are proposed where clusters (or cohorts) of similar clients are identified and the model is personalized for each cluster. Clustered FL approches can be classified into two categories: centralized approaches where the server identifies client's assignment to a cluster [35], [36], and decentralized approaches where the clients choose and update the model parameters that best fit them [37], [38]. We followed a centralized cohort (clustered) FL approach by considering features extracted from the client's data. However, the IFL clients can influence the cohort building by defining the cohort building approach and by defining the federation criteria that affects prior population building. In our work, the server also identifies the number of required cohorts. Building cohorts (clusters) on the server in our work is efficient because only few features from clients are used. As it is shown in Section 4.3, a cohort-based FL approach yields a higher performance than a global FL especially for scenarios with non-IID data distribution.

The FedAvg algorithm aggregates model parameters by averaging the value of local parameters from clients at each communication round [7]. Clients often have different dataset sizes and data distributions, thus training an effective global model with a good convergence in FedAvg algorithm is challenging. Therefore, clients contribution in aggregation operation of FedAvg is weighted either proportional to their local dataset size [7] or using multiple criteria such as diversity of dataset, data quality, and dataset size of clients local dataset [39]. In our work, all clients are weighted equally, but we used a weighted loss function to deal with imbalanced data (Section 4.4). In this way, we do not use any information about clients data on the server.

Selecting edge-based clients for participating in FL rounds is an important aspect for reducing communication cost and potentially long-lasting training times. This can be due to heterogeneous compute capabilities provided on resource-constrained edge devices. Clients may be selected randomly [7], [40] or based on pre-defined criteria such as availability of their resources [41]. In our approach, we follow a clustered FL approach, using all the available clients for building a FL population. As discussed in Section 4.1, clients' resources are similar in our evaluation setup. Therefore, we did not need to include any resource-based criteria on the server for selecting the clients. However, in the IFL system, we consider client-defined federation criteria for building FL populations. This could be used as a basis for defining requirements on resources that are provided on edge devices of potential FL partners.

Focusing on edge-based FL, Feraudo et al. [42] provide an architecture that enables asynchronous FL for edge devices using a publish/subscribe pattern. For this, the client registers for FL by sending a message to a message broker, which notifies the server, that waits until the end of a predefined timespan to consider the start of FL rounds. Similarly, Bonawitz et al. [39] proposed a FL system enabling edge devices to declare their training intention on a central server that starts FL as a required number of clients have joined. Furthermore, the authors considers selection and rejection of clients on the server. Hence, both approaches address the challenge of a varying number of edge devices, particularly unavailable edge devices. Our IFL services enable clients to define, e.g., a minimum number of FL partners, using custom federation criteria. This democratic approach provides more power to the edge clients, e.g., by reducing the required minimum number of FL partners if too less clients joined in previous FL attempts due to potential unavailability.

To provide FLaaS to edge and mobile devices, Kourtellis et al. [14] provide high-level APIs that can be invoked by mobile apps to collaborate on common ML problems. The authors propose a hierarchical scheme for collecting model updates on local device services considering potentially multiple apps. The updates can be forwarded to e.g., edge nodes or central servers for aggregation. Furthermore, the authors address challenges like the design of collaboration accross (hierarchical) network topologies, permission and privacy management, and forms of providing FL-based ML models to clients. However, the approach does not provide a service for building cohorts based on the client data distribution, which is relevant for non-IID datasets.

6 Conclusion

In this paper, we have presented the IFL system, consisting of the IFL services and the IFL client that provide FLaaS for edge-based clients connected to industrial machines. The machine data is used by the IFL client to collaboratively train ML models together with other FL clients. For this, we introduced the IFL process that includes a four-step approach enabling cohort-based FL and therefore addressing the challenge of non-IID data. We proposed a service architecture that supports the IFL process by providing APIs to clients for participating in FL rounds.

As discussed, our IFL system has three main contributions, (i) explicit participation in FL on an on-demand basis, (ii) an architecture supporting individual and independent selection of ML models, and (iii) handling non-IID data distributions using cohorts. To the best of our knowledge, we provided the first FLaaS-based system that supports all three contributions.

Evaluation on diverse and real-world industrial data is missing in current FL literature. However, we showed the high potential of our IFL system by applying it on two completely diverse and large real-world industrial datasets, used for pump condition classification and material classification. Our results also show the importance of a cohort-based FL approach, yielding higher accuracies on average compared to executing FL algorithms on the overall population.

The proposed IFL system is an important step towards incorporating FL in industrial applications and opens future directions as discussed in Section 4.5 for FL research, e.g., adaptive cohort assignment considering resource optimization on edge devices.

Acknowledgments

We thank Stefan von Dosky and Thomas Bierweiler for the vibration sensors and the data on the pump classification use case, and Damian Jaeger, Stephanie Holly, and Michael Ungersboeck for technical contributions in the evaluation.

References

[1] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data mining and analytics in the process industry: The role of machine learning," *IEEE Access*, vol. 5, pp. 20 590–20 616, 2017.

[2] I. S. Candanedo, E. H. Nieves, S. R. González, M. T. S. Martín, and A. G. Briones, "Machine learning predictive model for industry 4.0," in *Knowledge Management in Organizations*, L. Uden, B. Hadzima, and I.-H. Ting, Eds. Cham: Springer International Publishing, 2018, pp. 501–510.

[3] S. Bagavathiappan, B. Lahiri, T. Saravanan, J. Philip, and T. Jayakumar, "Infrared thermography for condition monitoring–a review," *Infrared Physics & Technology*, vol. 60, pp. 35–55, 2013.

[4] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2015, p. 1310–1321.

[5] S. Abt and H. Baier, "Are we missing labels? a study of the availability of ground-truth in network security research," in *2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*. IEEE, 2014, pp. 40–55.

[6] A. Diez-Olivan, J. Del Ser, D. Galar, and B. Sierra, "Data fusion and machine learning for industrial prognosis: Trends and perspectives towards industry 4.0," *Information Fusion*, vol. 50, pp. 92–111, 2019.

[7] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2016, pp. 1273–1282.

[8] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *35th International Conference on Machine Learning (ICML)*, vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 5650–5659.

[9] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *arXiv preprint arXiv:1912.13445*, 2019.

[10] P. Yu, L. Wynter, and S. H. Lim, "Fed+: A family of fusion algorithms for federated learning," *arXiv preprint arXiv:2009.06303*, 2020.

[11] T. Hiessl, D. Schall, J. Kemnitz, and S. Schulte, "Industrial federated learning – requirements and system design," in *Highlights in Practical Applications of Agents, Multi-Agent Systems, and Trustworthiness. The PAAMS Collection*. Springer International Publishing, 2020, pp. 42–53.

[12] E. by: Peter Kairouz and H. B. McMahan, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1, 2021.

[13] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.

[14] N. Kourtellis, K. Katevas, and D. Perino, "Flaas: Federated learning as a service," in *Proceedings of the 1st Workshop on Distributed Machine Learning*, ser. DistributedML'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 7–13. [Online]. Available: https://doi.org/10.1145/3426745.3431337

[15] M. Kritzler, J. Hodges, D. Yu, K. Garcia, H. Shukla, and F. Michahelles, "Digital companion for industry," in *Companion Proceedings of The 2019 World Wide Web Conference*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 663–667. [Online]. Available: https://doi.org/10.1145/3308560.3316510

[16] X. M. Zhao, Q. H. Hu, Y. G. Lei, and M. J. Zuo, "Vibration-based fault diagnosis of slurry pump impellers using neighbourhood rough set models," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 224, no. 4, pp. 995–1006, 2010.

[17] H. Kevin, P. Amar, M. Onur, and P. G. B., "The non-IID data quagmire of decentralized machine learning," in *36th International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 4387–4398.

[18] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.

[19] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[20] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a "kneedle" in a haystack: Detecting knee points in system behavior," in *2011 31st International Conference on Distributed Computing Systems Workshops*. IEEE, 2011, pp. 166–171.

[21] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[22] K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D. L. Rubin, and J. Kalpathy-Cramer, "Distributed deep learning networks among institutions for medical imaging," *Journal of the American Medical Informatics Association*, vol. 25, no. 8, pp. 945–954, 03 2018. [Online]. Available: https://doi.org/10.1093/jamia/ocy017

[23] T. Bierweiler, H. Grieb, S. von Dosky, and M. Hartl, "Smart sensing environment – use cases and system for plant specific monitoring and optimization," in *Automation 2019*. VDI Verlag, 2019, pp. 155–158.

[24] F. L. Markley, "Attitude determination using vector observations and the singular value decomposition," *Journal of the Astronautical Sciences*, vol. 36, no. 3, pp. 245–258, 1988.

[25] L. Cremer and M. Heckl, *Structure-borne sound: structural vibrations and sound radiation at audio frequencies*. Springer Science & Business Media, 2013.

[26] R. N. Bracewell and R. N. Bracewell, *The Fourier transform and its applications*. McGraw-Hill New York, 1986.

[27] S. Lin, N. Liu, M. Nazemi, H. Li, C. Ding, Y. Wang, and M. Pedram, "Fft-based deep learning deployment in embedded systems," in *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2018, pp. 1045–1050.

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[29] C. M. Bishop, *Pattern recognition and machine learning*. Springer New York, NY, USA, 2007.

[30] G. King and L. Zeng, "Logistic regression in rare events data," *Political Analysis*, vol. 9, no. 2, pp. 137–163, 2001.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *International Journal of Medical Informatics*, vol. 112, pp. 59–67, 2018.

[33] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 92–104.

[34] F. Fioretto and P. Van Hentenryck, "Privacy-preserving federated data sharing," in *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 638–646.

[35] A. Ghosh, J. Hong, D. Yin, and K. Ramchandran, "Robust federated learning in a heterogeneous environment," *arXiv preprint arXiv: 1906.06629*, 2019.

[36] F. Sattler, K. R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2020.

[37] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *arXiv preprint arXiv: 2006.04088*, 2020.

[38] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, "Three approaches for personalization with applications to federated learning," *arXiv preprint arXiv: 2002.10619*, 2020.

[39] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecný, S. Mazzocchi, H. B. McMahan, T. V. Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," *arXiv preprint arXiv: 1902.01046*, 2019.

[40] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–7.

[41] S. A. Rahman, H. Tout, A. Mourad, and C. Talhi, "FedMCCS: Multi criteria client selection model for optimal iot federated learning," *IEEE Internet of Things Journal*, vol. 8, pp. 4723–4735, 2020.

[42] A. Feraudo, P. Yadav, V. Safronov, D. A. Popescu, R. Mortier, S. Wang, P. Bellavista, and J. Crowcroft, "Colearn: Enabling federated learning in mud-compliant iot edge networks," in *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, ser. EdgeSys '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 25–30. [Online]. Available: https://doi.org/10.1145/3378679.3394528

Thomas Hiessl received his diploma degree in computer science from TU Wien in 2017. He then joined Siemens Austria, working on the development of industrial Internet of Things technology and distributed artificial intelligence. He is currently pursuing the Ph.D. in computer science at TU Wien. His research interests include federated learning, cloud computing and edge computing in Internet of Things applications.

Safoura Rezapour Lakani is a data scientist by the Distributed AI System research group at Siemens Technology in Vienna. She received her PhD in computer science from the University of Innsbruck in 2018. Her research interests includes machine learning, in particular transfer and federated learning.

Jana Kemnitz is a senior data scientist and machine learning expert at the Distributed AI System research group at Siemens Technology in Vienna. She received her PhD as a Marie Curie fellow from the Paracelsus Medical University. Her main interests are image- and signal processing, machine learning and medical science.

Dr. Daniel Schall is head of the Distributed AI System research group at Siemens Technology in Vienna. His main interests are AI systems, platforms, and applications that have real-world impact. He published more than 80 journal and conference papers and 3 books on service-oriented computing. Daniel Schall received his PhD in computer science from TU Wien in 2009.

Stefan Schulte is Associate Professor and head of the Christian Doppler Laboratory Blockchain Technologies for the Internet of Things at the Faculty of Informatics at TU Wien. His research interests span the areas of cloud computing and Internet of Things, and the application and extension of blockchain technologies. Findings from his research have been published in more than 100 refereed scholarly publications.