KU LEUVEN

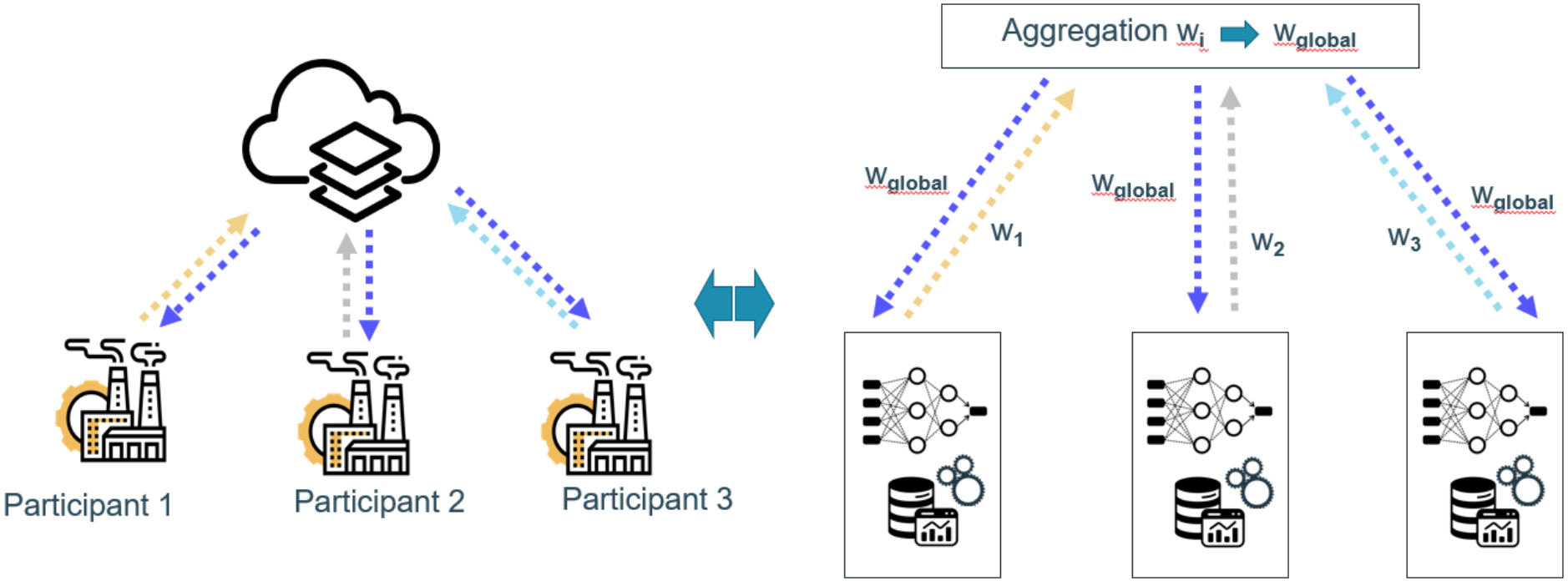
A federated learning approach to a fault diagnosis bearing problem

Fabrizio De Fabritiis (fabrizio.defabritiis@kuleuven.be)
KU Leuven - LMSD
September 1^{st} 2023

0 Outline

1 Outline

1 Federated learning scenario



1 Mathematical formulation

We consider a supervised learning task with features $\mathbf{x}$ in a sample space $\mathcal{X}$ and labels y in a label space $\mathcal{Y}$. We assume that we have K available clients, $K \in \mathbb{N}_{>1}$, with

$$D_k \coloneqq D_{\mathcal{X},k} \times D_{\mathcal{Y},k} \subseteq \mathcal{X} \times \mathcal{Y}$$

let $\mathcal{P}_k$ denote the data distribution of client k, we wish to minimize $F(w)$

$$F_k(w) \coloneqq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{P}_k}[f(x,y,w)], \quad F(w) = \frac{1}{K}\sum_{k=1}^{K} F_k(w)$$

2 Outline

2 non-identically distributed settings

In general we cannot assume that the data is identically distributed over the clients, that is

$$\mathcal{P}_k \neq \mathcal{P}_l \qquad \text{for } k, l = 1, \ldots, K$$

In real-world problems, the data D_k on a given client k depends on individual conditions, thus this local data set is not necessarily representative of the data set D_l of client l

Therefore, in practice, F_k might be an arbitrary bad approximation of F

2 non-identically distributed settings 2

Assuming we have an IoT sensor based anomaly detection classification task in an industrial context, we want to analyse different non-identically distributed settings

$$\mathcal{P}^k_{\mathcal{X},\mathcal{Y}}(\mathbf{x}, y) = \mathcal{P}^k_{\mathcal{Y}|\mathcal{X}}(y|\mathbf{x})\mathcal{P}^k_{\mathcal{X}}(\mathbf{x}) = \mathcal{P}^k_{\mathcal{X}|\mathcal{Y}}(\mathbf{x}|y)\mathcal{P}^k_{\mathcal{Y}}(y)$$

This allows us to characterize different settings of non-identically distributed data

2 Feature distribution skew

We assume that $\mathcal{P}^k_{\mathcal{Y}|\mathcal{X}} = \mathcal{P}^l_{\mathcal{Y}|\mathcal{X}}$ for all clients

However, $\mathcal{P}^k_{\mathcal{X}} = \mathcal{P}^l_{\mathcal{X}}$ possibly fails to hold for all clients

2 Label distribution skew

We assume that $\mathcal{P}^k_{\mathcal{X}|\mathcal{Y}} = \mathcal{P}^l_{\mathcal{X}|\mathcal{Y}}$ for all clients

However, $\mathcal{P}^k_{\mathcal{Y}} = \mathcal{P}^l_{\mathcal{Y}}$ possibly fails to hold for all clients

2 Same label, different features

We assume that $\mathcal{P}_{\mathcal{Y}}^{k} = \mathcal{P}_{\mathcal{Y}}^{l}$ for all clients

However, $\mathcal{P}_{\mathcal{X}|\mathcal{Y}}^{k} = \mathcal{P}_{\mathcal{X}|\mathcal{Y}}^{l}$ possibly fails to hold for all clients

2 Same feature, different labels

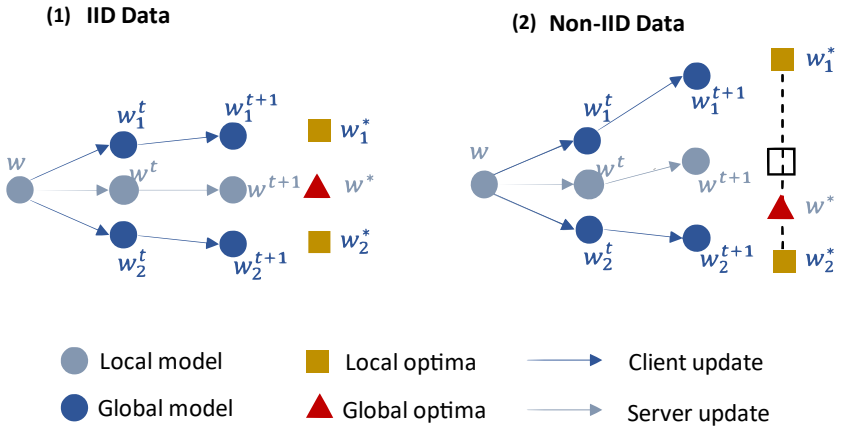
We assume that $\mathcal{P}^k_{\mathcal{X}} = \mathcal{P}^l_{\mathcal{X}}$ for all clients

However, $\mathcal{P}^k_{\mathcal{Y}|\mathcal{X}} = \mathcal{P}^l_{\mathcal{Y}|\mathcal{X}}$ possibly fails to hold for all clients

2 Quantity skew

We cannot assume that different clients hold the same amount of data, that is $n_k = n_l$ for all $k, l = 1, \ldots, K$

Some clients will generate more data than others resulting in different amounts of local data



2 Other sources of heterogeneity in FL

In FL, heterogeneity does not exclusively refer to a non-identical data distribution, but also addresses violations of independence assumptions on the distribution $\mathcal{Q}$

- Due to limited, slow and unreliable communication on a client, the availability of a client is not guaranteed for all communication rounds.
- An active client can drop out of training at a given communication round
- A new client fulfilling the respective technical requirements can participate in training at a given communication round

2 Communication bottleneck

In each communication round, the participating clients send a full model update w back to the central server

However, in a typical FL setting the clients are usually limited in terms of communication bandwidth.

It is crucial to minimize the communication costs by reducing the number of communication rounds or using compressed communication schemes for the model updates to the central server on each client

2 Performance evaluation of a FL approach

In traditional approaches In traditional approaches, the performance of a machine learning model is most commonly defined as the model accuracy

Since communication in federated learning is much more expensive than computation, it is crucial to minimize communication. Therefore, we define performance as the highest classification accuracy achieved after a given amount of communication.

3 Outline

- We consider an industrial classification problem. The task is the classification of the healthy and anomalous conditions of an industrial asset.
- The data is measured by a sensor that is attached to the asset.
- We expect to find differences in the measurements due to variations in sensor and asset type, and therefore differences in the feature distribution.
- In reality, the anomalous conditions differ from asset to asset.
  Therefore, we expect to find differences in the label distribution.

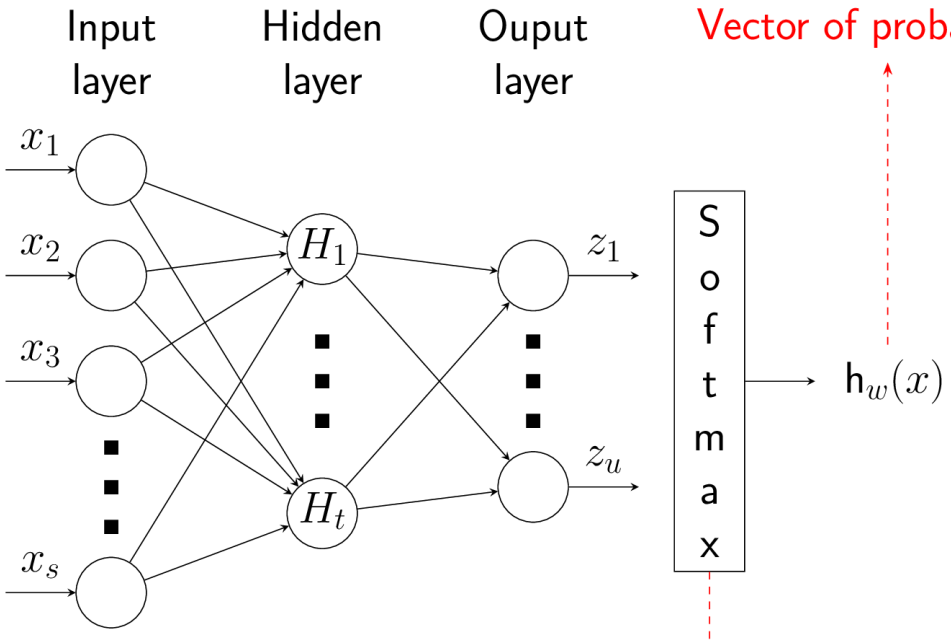
We consider an industrial classification problem

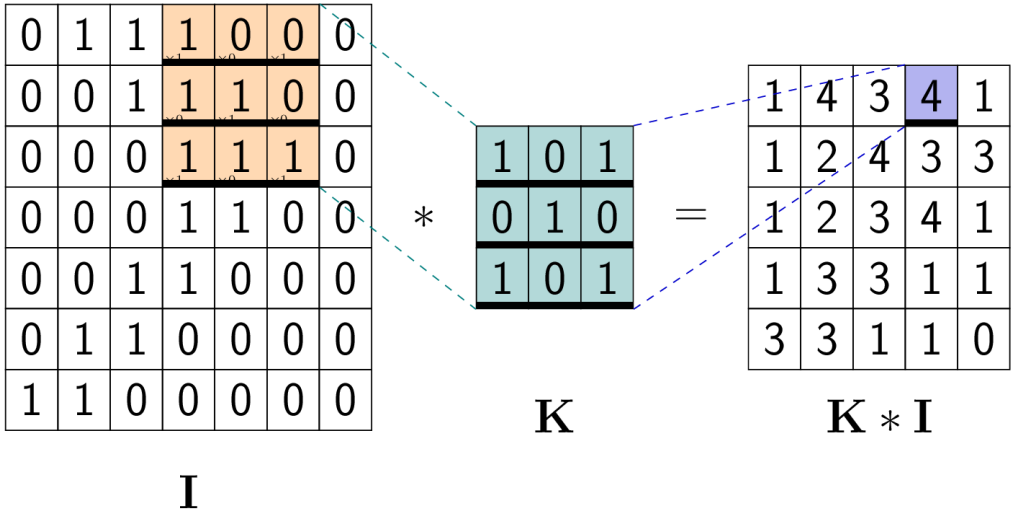
Let $\mathcal{X} \subset \mathbb{R}^n$ denote the sample space and let $\mathcal{Y} = \{1, \ldots, C\} \subset \mathbb{N}$ denote the label space

The cross-entropy loss for a multiclass classification problem is defined as:

$$E(w) \coloneqq -\frac{1}{N} \sum_{c=1}^{C} \sum_{n=1}^{N} y_n^c \log(h_w(x_n))$$

3 softmax activation





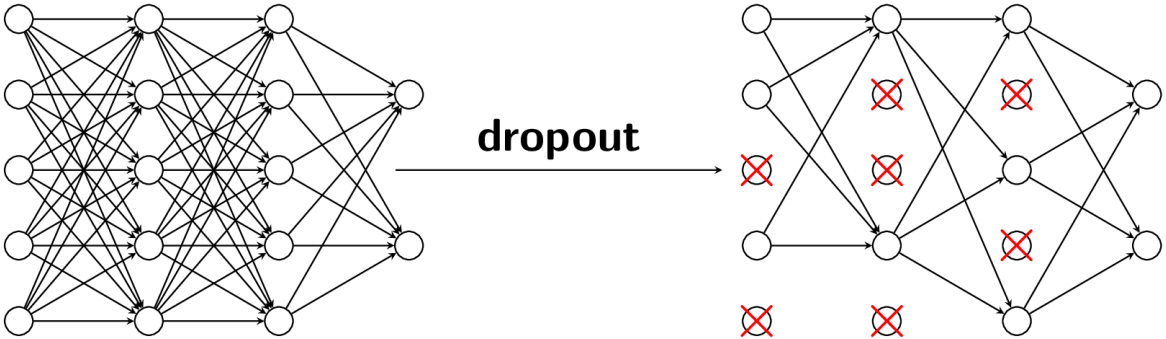
$$(K * I)(i,j) = \sum_{m,n} K(m,n) I(i+n, j+m)$$

For a 2D input with C^i channels, the shape of a kernel is (k, k, C^i, C^o)

$$(K_l * I)(i,j) = \sum_{c=0}^{C_i-1} \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} K_l(m,n,c) I(i+n, j+m, c) \quad l = 1, \ldots, C^o$$

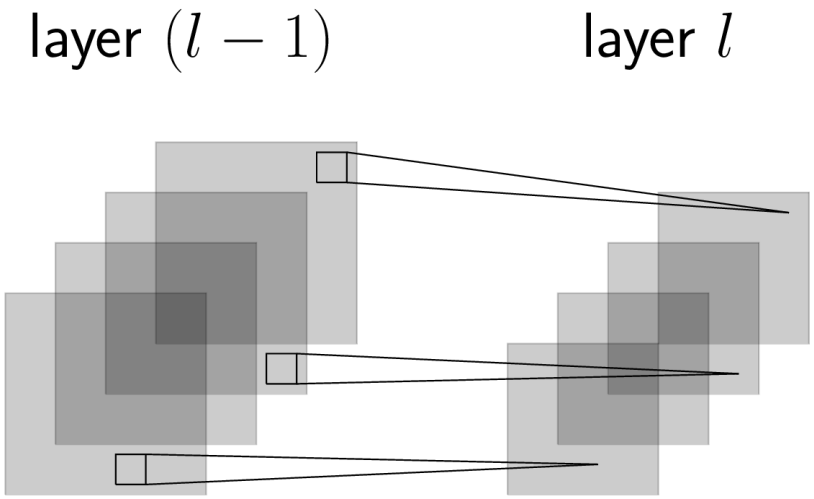
$$W_o = \frac{W_i - k + 2p}{s} + 1, \qquad H_o = \frac{H_i - k + 2p}{s} + 1$$

3 Dropout

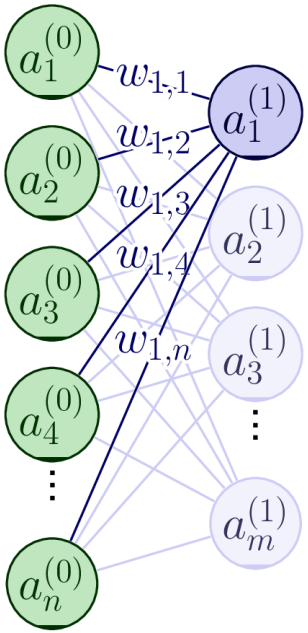


The output of a randomly chosen subset of the neurons are temporarily set to zero during the training of a given mini-batch. This makes it so that the neurons cannot overly adapt to the output from prior layers as these are not always present.

3 Pooling layer



3 Activation

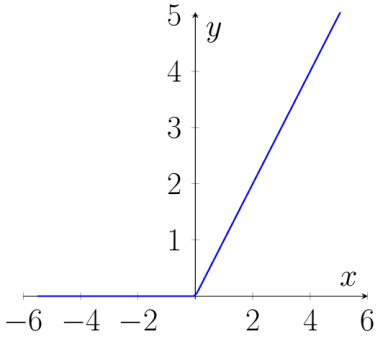


$$= \sigma\left(w_{1,0}a_0^{(0)} + w_{1,1}a_1^{(0)} + \ldots + w_{1,n}a_n^{(0)} + b_1^{(0)}\right)$$

$$= \sigma\left(\sum_{i=1}^{n} w_{1,i}a_i^{(0)} + b_1^{(0)}\right)$$

$$\begin{pmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \vdots \\ a_m^{(1)} \end{pmatrix} = \sigma\left[\begin{pmatrix} w_{1,0} & w_{1,1} & \ldots & w_{1,n} \\ w_{2,0} & w_{2,1} & \ldots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,0} & w_{m,1} & \ldots & w_{m,n} \end{pmatrix}\begin{pmatrix} a_1^{(0)} \\ a_2^{(0)} \\ \vdots \\ a_n^{(0)} \end{pmatrix} + \begin{pmatrix} b_1^{(0)} \\ b_2^{(0)} \\ \vdots \\ b_m^{(0)} \end{pmatrix}\right]$$

$$a^{(1)} = \sigma\left(\mathbf{W}^{(0)}a^{(0)} + \mathbf{b}^{(0)}\right)$$



4 Outline

Local SGD and global aggregation following $w_r = \sum_{k=1}^{K} \frac{n_k}{n} w_r^k$

to tackle the minimization problem $\min_{w \in \mathbb{R}^d} \frac{n_k}{n} \sum_{k=1}^{K} F_k(w)$

where $F_k(w) \coloneqq \frac{1}{n_k} \sum_{i=1}^{n_k} f_i(w)$ with $f_i(w) \coloneqq \mathcal{L}(x_i, y_i, w)$

4 FedAvg

Algorithm 1 Federated Averaging

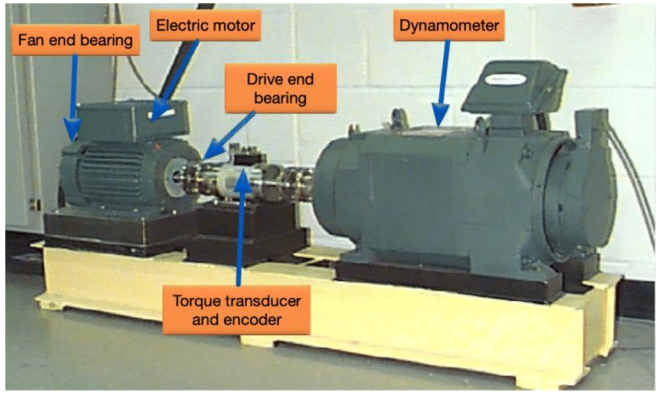
Server executes:
Initialize w_0
for each round n= $1, \ldots, N$ **do**
  $S_r \coloneqq$ random set of C$\cdot K$ clients
  for each client $k = 1 \in S_r$ in parallel **do**
    $w_r^k \leftarrow \text{ClientUpdate}(k, w_{r-1})$
  end for
  $w_r \leftarrow \Sigma_{k=1}^{K} \frac{n_k}{n} w_r^k$
end for
Client executes:
ClientUpdate(k, w):
$\mathcal{B} \leftarrow$ (split D_k into batches of size B)
for each epoch $t = 1, \ldots, E$ **do**
  for all each batch $b \in \mathcal{B}$ **do**
    $w \leftarrow w - \nabla \mathcal{L}(b, w)$
  end for
end for
return w to server

5 Outline

5 CWRU dataset

- Measurements from accelerometer installed at the Drive end (DE) of the system
- Sampling frequency: 12 kHz
- time series for each defect located in 1 of 3 parts of the bearing: ball, inner race, outer race

Motor Load	Normal	Inner Fault (inch)			Outer Fault (inch)			Bearing Fault (inch)		
0		0.007	0.014	0.021	0.007	0.014	0.021	0.007	0.014	0.021
1										
2										
3										

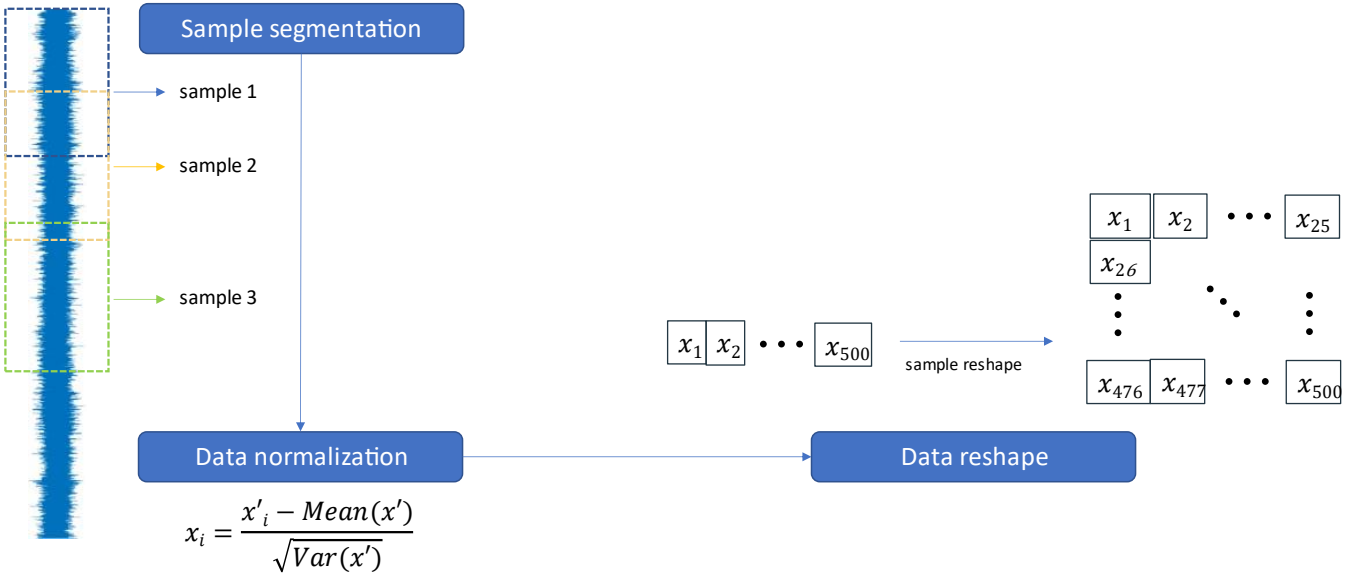


Label	Fault class	Severity (diameter, depth)	Working condition
c_0	-	-	0 HP
c_1	Inner race	0.18 mm, 0.28 mm	0 HP
c_2	Ball	0.18 mm, 0.28 mm	0 HP
c_3	Outer race	0.18 mm, 0.28 mm	0 HP
c_4	Inner race	0.36 mm, 0.28 mm	0 HP
c_5	Ball	0.36 mm, 0.28 mm	0 HP
c_6	Outer race	0.36 mm, 0.28 mm	0 HP
c_7	Inner race	0.53 mm, 0.28 mm	0 HP
c_8	Ball	0.53 mm, 0.28 mm	0 HP
c_9	Outer race	0.53 mm, 0.28 mm	0 HP

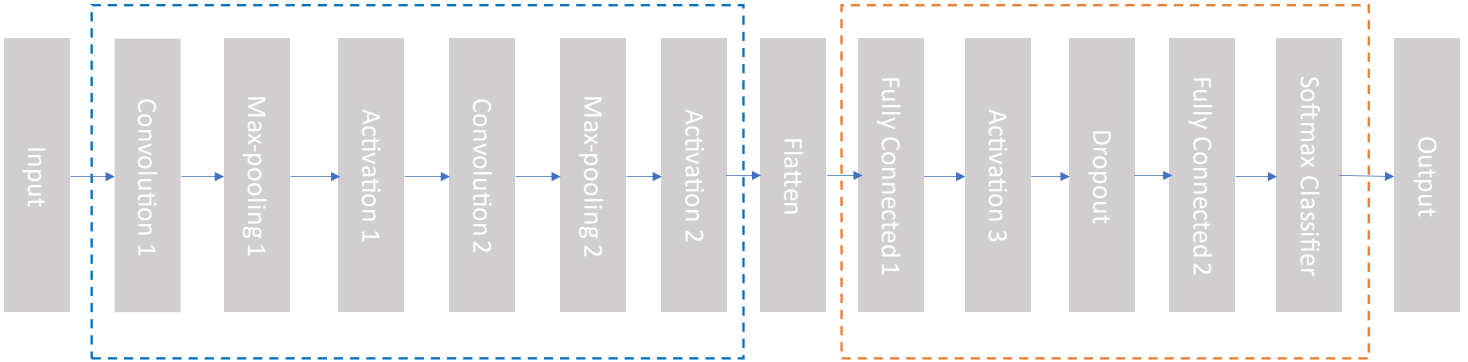
Table: Class labels and related fault description

- $D_1 = \{(\mathbf{x}_j^1, y_j^1)\}_{j=1}^{|D_1|}$ where $y_j^1 \in \{c_0, c_1, c_2, c_3, c_4\}$
- $D_2 = \{(\mathbf{x}_j^2, y_j^2)\}_{j=1}^{|D_2|}$ where $y_j^2 \in \{c_5, c_6, c_7\}$
- $D_3 = \{(\mathbf{x}_j^3, y_j^3)\}_{j=1}^{|D_3|}$ where $y_j^3 \in \{c_8, c_9\}$

5 Data pre-processing



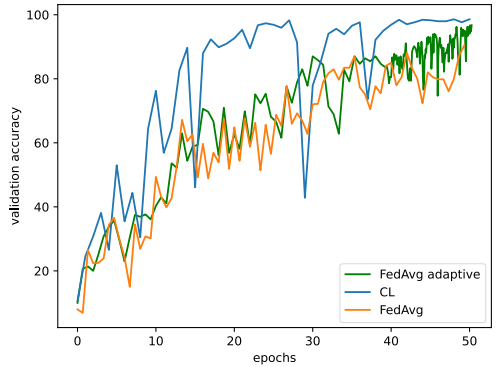
5 Network architecture

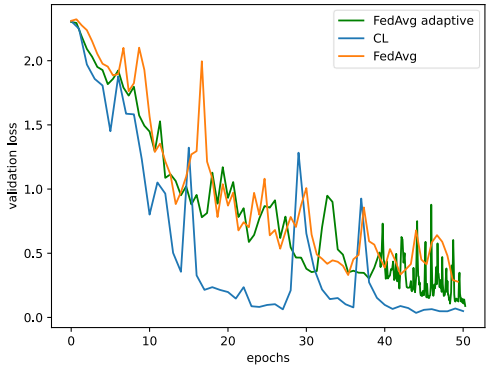


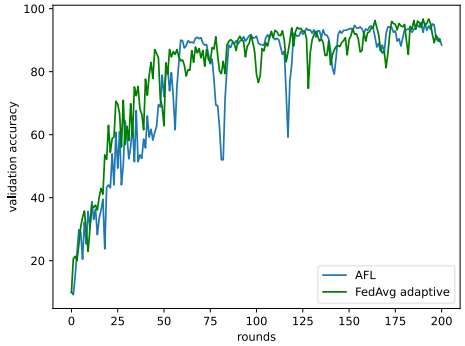
5 Results

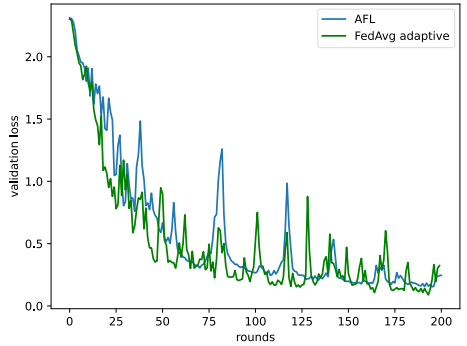
Results of the Federated Learning models on the test set						
Model	Precision	Recall	Accuracy	F1-Score	Round	Epochs
CL	0.979828	0.978125	0.978125	0.978168	-	50
FedAvg adaptive	0.971875	0.973255	0.971875	0.971860	193	~ 50
AFL	0.956630	0.9484375	0.9484375	0.9483477	196	~ 36
FedAvg($\tau = 10$)	0.907337	0.8890625	0.8890625	0.888795	75	50

Table: Performance on the test data set









Acknowledgment

The authors gratefully acknowledge the European Commission for its support of the Marie Sklodowska Curie program through the H2020 ETN MOIRA project (GA 955681)

fabrizio.defabritiis@kuleuven.be

www.mech.kuleuven.be/lmsd

www.mech.kuleuven.be/lmsd-joboffers

lmsd_kuleuven

lmsd-kuleuven

lmsd.kuleuven

lmsd-kuleuven

lmsd_kuleuven

KU Leuven Mecha(tro)nic System Dynamics (LMSD)