

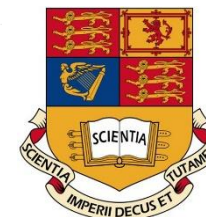
# Federated Intelligent Terminals Facilitate Stuttering Monitoring

Yongzi Yu<sup>1†</sup>, **Wanyong Qiu<sup>1†</sup>**, Chen Quan<sup>1</sup>, Kun Qian<sup>1\*</sup>, Zhihua Wang<sup>2</sup>,  
Yu Ma<sup>1</sup>, Bin Hu<sup>1\*</sup>, Bjoern W. Schuller<sup>3</sup> and Yoshiharu Yamamoto<sup>2</sup>

1 School of Medical Technology, Beijing Institute of Technology, Beijing, China

2 Educational Physiology Laboratory, The University of Tokyo, Japan

3 GLAM -- Group on Language, Audio, & Music, Imperial College London, UK



# CONTENT



01

Overview

02

Motivation

03

Data & Method

04

Result

05

Conclusion

# Overview

## Federated intelligent terminals for automatic monitoring of stuttering speech in different contexts

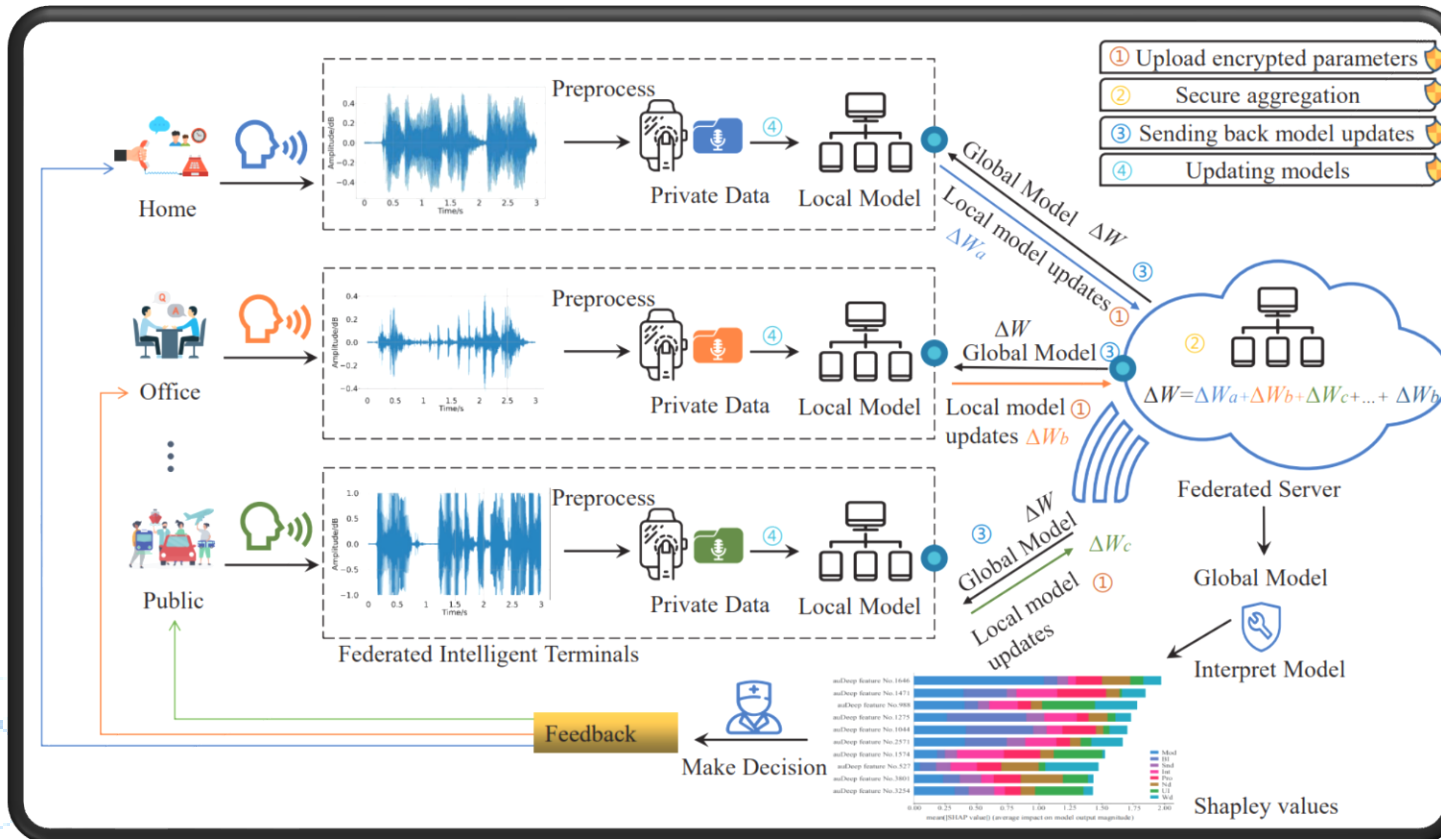


Fig.1 The framework of federated intelligent terminals

## Contribution

- The **first time** that FL<sup>[1]</sup> has been applied to stuttering scenarios
- Verify that XGBoost-based FL has **comparable performance** with centralised learning for stuttering classification
- Introduce **Shapley values** to measure changes in feature importance

[1] FL(Federated Learning)

# Motivation

- Monitoring of stuttering is **crucial** to speech therapy.
- Evaluation of stuttering by speech therapists can be **influenced** by too much manual subjective intervention
  - Comprehensive evaluation in **various contexts** is required.
  - The therapist's evaluation might be **influenced by many factors**
    - ❑ communication situation
    - ❑ psychological factors
    - ❑ linguistic complexity
    - ❑ personal subjectivity
- Problem **of data security**.

So we propose **the federated intelligent terminals** for automatic monitoring of stuttering speech in different contexts!





# Method-Data and Explainable

## Data Preparation

- The experimental data are taken from the **Kassel State of Fluency** (KSoF) corpus.<sup>[1]</sup>
  - Train: 23 speakers
  - Devel: 6 speakers
  - Sample number: 3,471
  - Length of each audio: 3-second
  - Classes: 8
  - Feature: 4,096 dimensions extracted by auDeep.

## Shapley[2] value Tool

Fairly **evaluate feature contributions** by assigning each feature a numerical value to represent its impact.

**Table.1** The Distribution of annotations in KSoF dataset

Stuttering Labels	KSoF [%]
Block (Bl)	20.74
Prolongation (Pro)	12.02
Sound Repetition (Snd)	14.76
Word/Phrase Repetition (Wd)	3.88
Modified Speech Technique (Mod)	24.75
Interjection (Int)	24.44
No Dysfluencies (Nd)	12.97
Unintelligible (Ul)	5.77

[1] The data can be accessed by request from the Kassel State of Fluency (KSoF) dataset at <https://zenodo.org/record/6801844>

[2] SHAP (SHapley Additive exPlanations) is a game-theoretic method to explain the output of ML models. <https://shap.readthedocs.io>.

# Method-Centralised model

## XGBoost Ensemble Learning Model

### Positive:

- ✓ Good at **parallel** computing
- ✓ Highly **scalable**
- ✓ Uses **minimal resources** for algorithmic optimization
- ✓ Has **flexible** portability and precise libraries

### Object Function:

$$Obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^t \Omega(f_i)$$
$$\approx \sum_{i=1}^n [g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)] + \Omega(f_i)$$

i refers to the  $i^{\text{th}}$  sample,  $\hat{y}_i = \sum_{k=1}^K f_k(x_i)$ .

$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$  and  $h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}^{(t-1)})$

the first order gradient

the second order gradient

# Method-Federated model

The framework is based on **FATE**[1].

The XGBoost-based horizontal FL steps:

- a) Clients hold different training samples and train the ensemble tree model.
- b) For each feature, the client accumulates the **gradient** of its samples' loss.
- c) Clients send the gradient to the server.
- d) The server aggregates the gradients from the clients and finds out the best weights.
- e) The server broadcasts the best weights to clients.

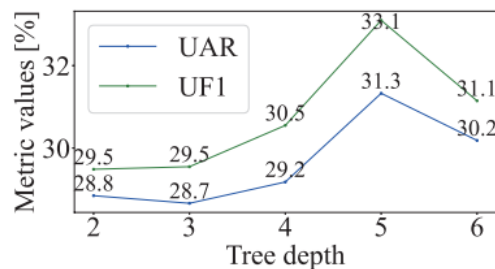
[1] FATE (Federated AI Technology Enabler) supports the FL architecture, as well as the secure computation and development of various ML algorithms. <https://github.com/FederatedAI/FATE>

**Algorithm 1:** Implementation of XGBoost-based horizontal FL

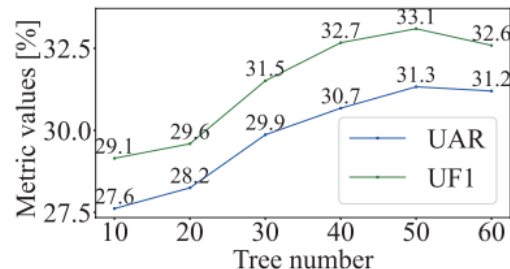
**Input:**  $N$ , the number of the clients, where the  $i^{th}$  client holds  $n_i$  instance spaces  
**Input:**  $d$ , feature dimension  
**Input:**  $x$ , the dataset matrix  
**Output:** the best split point for the current instance space

```
1 /*On clients*/
2 for each client  $i = 1$  to  $N - 1$  do
3     Propose each feature's values by percentiles to form feature bins
4     for each feature bin do
5         Accumulate the  $g, h$  of all sample spaces in this feature bin to get  $G, H$ 
6     end
7 end
8 /*On federated server */
9 for each client  $i = 1$  to  $N - 1$  do
10    for each feature  $m = 1$  to  $d - 1$  do
11         $g_l = g_l + \text{Decrypt}(G \text{ feature bins})$ 
12         $h_l = h_l + \text{Decrypt}(H \text{ feature bins})$ 
13         $g_r = g - g_l, h_r = h - h_l$ 
14        Score =
             $\text{Max}(\text{Score}, \frac{1}{2} [\frac{g_l^2}{h_l + \lambda} + \frac{g_r^2}{h_r + \lambda} - \frac{g^2}{h + \lambda}] - \gamma)$ 
15    end
16 end
17 Broadcast the  $m_{opt}$  and the corresponding threshold value to all clients to split
```

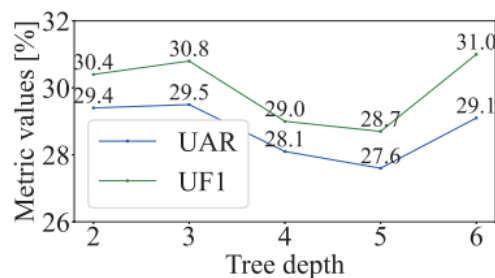
# Result



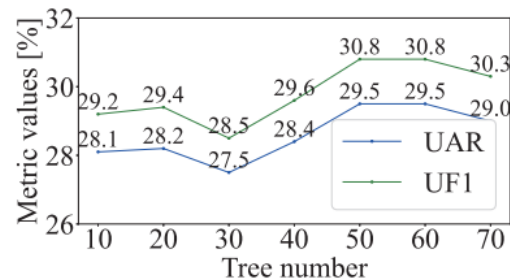
(a) Tree depth for XGBoost



(b) Tree number for XGBoost



(c) Tree depth for FL



(d) Tree number for FL

**Fig.2** Model performance variation (UAR and  $UF_1$  in [%]) between centralised learning and federated learning

$$UF_1 = \frac{2 * TPc}{2 * TPc + FPc + FNc}$$

$$UAR = \frac{\sum_{i=1}^{N_c} Recall_i}{N_c}$$

Evaluation matrix: **UAR** and **UF1**

- XGBoost is optimal with 50 trees and depth 5
- FL is optimal with 50 trees and depth 3

Pro-	6.0	50.7	6.0	22.4	10.4	4.5	0.0	0.0
Nd-	4.1	65.8	4.1	5.4	16.5	2.2	0.9	0.9
Int-	10.5	39.5	18.4	7.9	22.4	1.3	0.0	0.0
Bl-	5.1	48.0	0.0	34.7	6.1	4.1	0.0	2.0
Mod-	1.0	19.1	2.6	9.3	64.9	3.1	0.0	0.0
Snd-	4.0	50.0	6.0	6.0	6.0	24.0	4.0	0.0
Wd-	11.8	41.2	5.9	5.9	11.8	11.8	11.8	0.0
UI-	0.0	55.0	5.0	5.0	5.0	5.0	0.0	25.0
Pro-	6.0	50.7	6.0	22.4	10.4	4.5	0.0	0.0
Nd-	4.1	65.8	4.1	5.4	16.5	2.2	0.9	0.9
Int-	10.5	39.5	18.4	7.9	22.4	1.3	0.0	0.0
Bl-	5.1	48.0	0.0	34.7	6.1	4.1	0.0	2.0
Mod-	1.0	19.1	2.6	9.3	64.9	3.1	0.0	0.0
Snd-	4.0	50.0	6.0	6.0	6.0	24.0	4.0	0.0
Wd-	11.8	41.2	5.9	5.9	11.8	11.8	11.8	0.0
UI-	0.0	55.0	5.0	5.0	5.0	5.0	0.0	25.0

(a) XGBoost confusion matrix

Pro-	5.2	50.0	5.2	20.7	8.6	10.3	0.0	0.0
Nd-	3.4	60.3	5.4	5.8	20.3	3.1	0.7	1.0
Int-	7.6	37.9	28.8	4.5	19.7	1.5	0.0	0.0
Bl-	4.3	45.7	4.3	32.6	5.4	7.6	0.0	0.0
Mod-	0.6	26.9	2.3	8.8	58.5	2.3	0.6	0.0
Snd-	4.3	52.2	6.5	4.3	6.5	23.9	2.2	0.0
Wd-	11.8	64.7	0.0	0.0	11.8	11.8	0.0	0.0
UI-	0.0	57.9	5.3	10.5	0.0	0.0	0.0	26.3
Pro-	5.2	50.0	5.2	20.7	8.6	10.3	0.0	0.0
Nd-	3.4	60.3	5.4	5.8	20.3	3.1	0.7	1.0
Int-	7.6	37.9	28.8	4.5	19.7	1.5	0.0	0.0
Bl-	4.3	45.7	4.3	32.6	5.4	7.6	0.0	0.0
Mod-	0.6	26.9	2.3	8.8	58.5	2.3	0.6	0.0
Snd-	4.3	52.2	6.5	4.3	6.5	23.9	2.2	0.0
Wd-	11.8	64.7	0.0	0.0	11.8	11.8	0.0	0.0
UI-	0.0	57.9	5.3	10.5	0.0	0.0	0.0	26.3

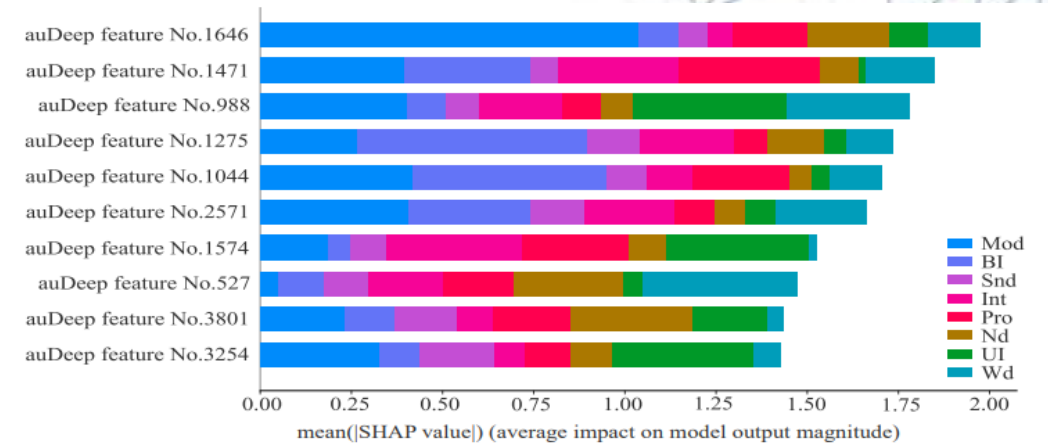
(b) FL confusion matrix

**Fig.3** Normalised confusion matrix (in [%]) of true labels and predicted labels between centralised learning and federated learning.

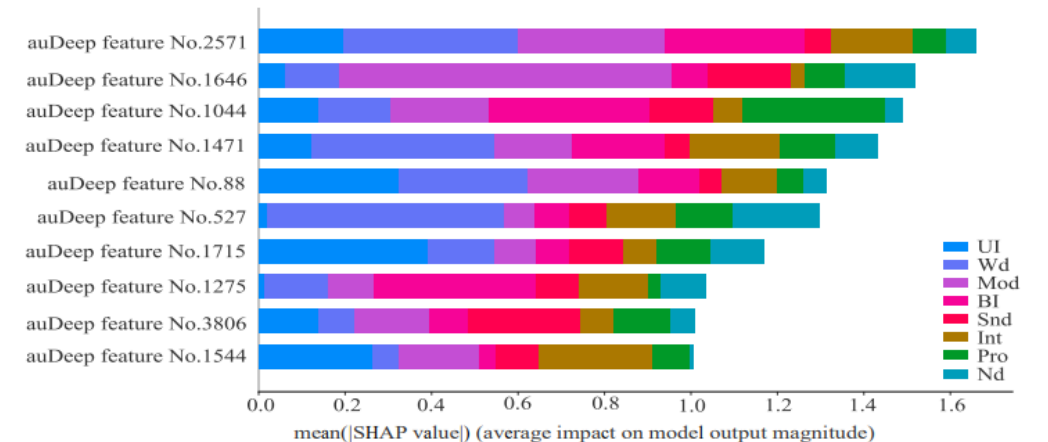


# Conclusion

- FL has considerable **privacy-preserving** advantages over centralised learning
- Offered a valid verification and basis for the **FL paradigm** on automatic monitoring of stuttering is provided
- Shapley values can fairly evaluate the **contribution** of features
- Future work: **lightweight models** and the deployment of FITs models on devices



(a) The contribution of significant auDeep\_features from all class predictions for the XGBoost model (average feature importance).



(b) The contribution of significant auDeep\_features from all class predictions for the FL model (average feature importance).

**Fig.4** The features sorted by the mean of Shapley values for all class predictions



# THANKS

Yongzi Yu<sup>1†</sup>, **Wanyong Qiu<sup>1†</sup>**, Chen Quan<sup>1</sup>, Kun Qian<sup>1\*</sup>, Zhihua Wang<sup>2</sup>,  
Yu Ma<sup>1</sup>, Bin Hu<sup>1\*</sup>, Bjoern W. Schuller<sup>3</sup> and Yoshiharu Yamamoto<sup>2</sup>

1 School of Medical Technology, Beijing Institute of Technology, Beijing, China

2 Educational Physiology Laboratory, The University of Tokyo, Japan

3 GLAM -- Group on Language, Audio, & Music, Imperial College London, UK

