Federated Intelligent Terminals Facilitate Stuttering Monitoring

Yongzi Yu[1†], **Wanyong Qiu**[1†], Chen Quan[1], Kun Qian[1*], Zhihua Wang[2], Yu Ma[1], Bin Hu[1*], Bjoern W. Schuller[3] and Yoshiharu Yamamoto[2]

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

Overview

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

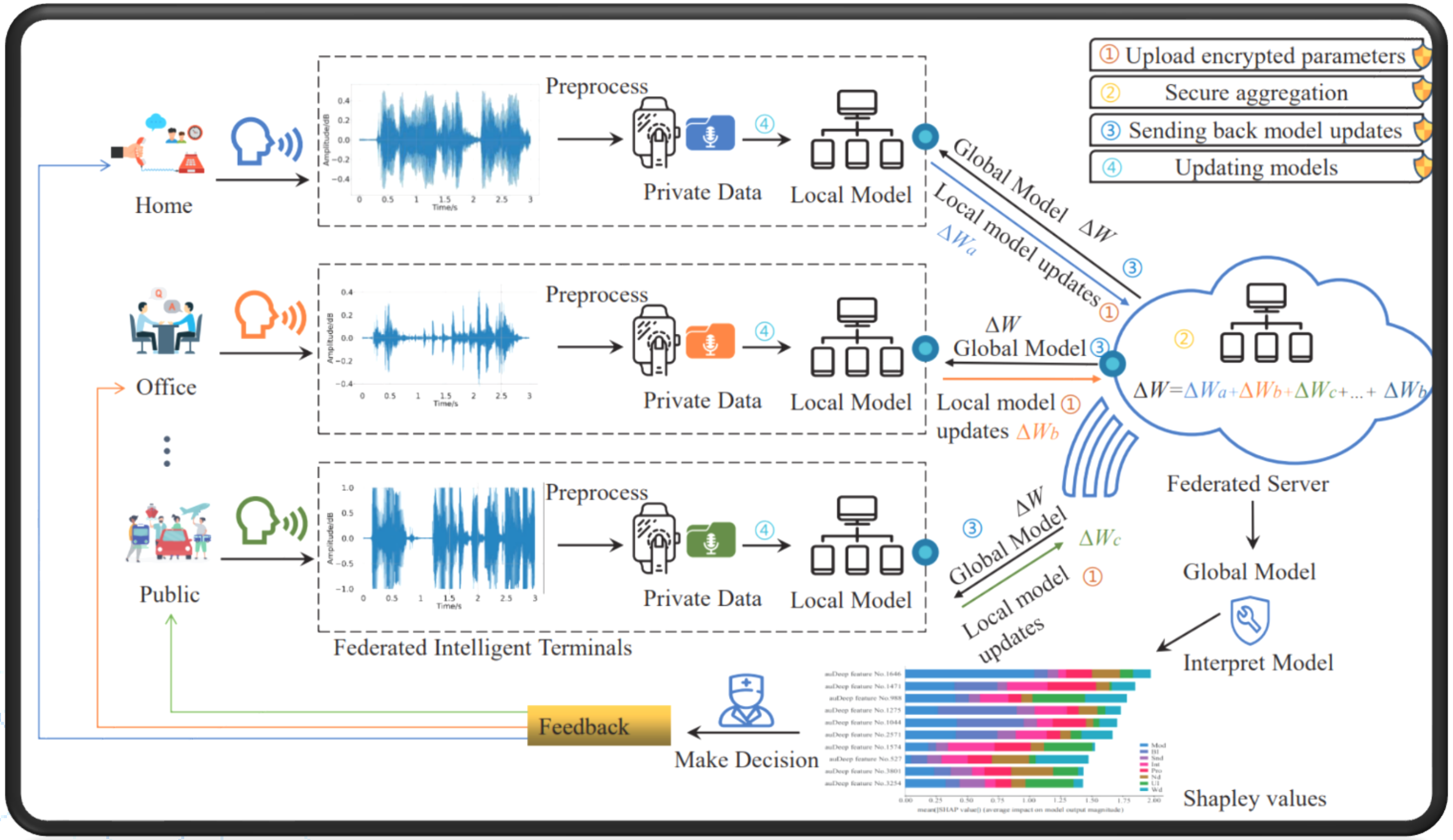


Fig.1 The framework of federated intelligent terminals

Contribution

- The **first time** that FL[1] 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.[1]
  - ➢ 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:

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i)$$

$$\approx \sum_{i=1}^{n} \left[g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_i)$$

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

$\underbrace{\boldsymbol{g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)}}_{\text{the first order gradient}}$ and $\underbrace{\boldsymbol{h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)}}_{\text{the second order gradient}}$

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

```
/*On clients*/
for each client i = 1 to N − 1 do
    Propose each feature's values by percentiles to form feature bins
    for each feature bin do
        Accumulate the g, h of all sample spaces in this feature bin to get G, H
    end
end
/*On federated server */
for each client i = 1 to N − 1 do
    for each feature m = 1 to d − 1 do
        g_l = g_l + Decrypt(G feature bins)
        h_l = h_l + Decrypt(H feature bins)
        g_r = g − g_l,  h_r = h − h_l
        Score = Max(Score, 1/2 [g_l^2/(h_l+λ) + g_r^2/(h_r+λ) − g^2/(h+λ)] − γ)
    end
end
Broadcast the m_opt and the corresponding threshold value to all clients to split
```

$$\text{Score} = \text{Max}\left(\text{Score}, \frac{1}{2}\left[\frac{g_l^2}{h_l+\lambda} + \frac{g_r^2}{h_r+\lambda} - \frac{g^2}{h+\lambda}\right] - \gamma\right)$$

Result

$$UF_1 = \frac{2 * TPc}{2 * TPc + FPc + FNc} \qquad UAR = \frac{\sum_{i=1}^{N_c} Recall_i}{N_c}$$

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

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

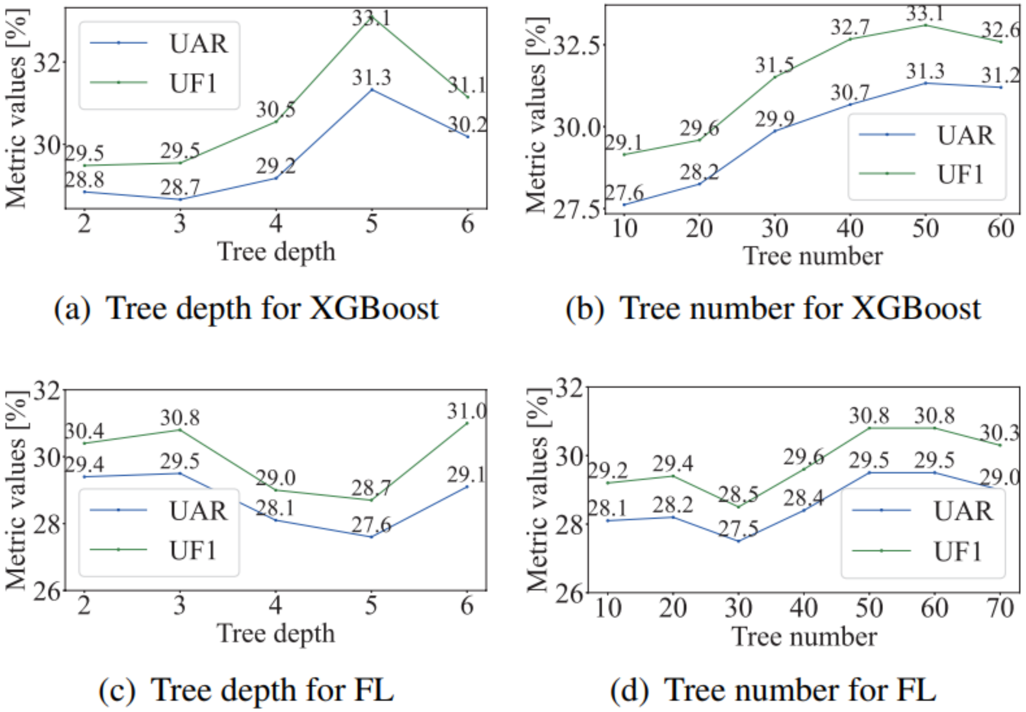


Fig.2 Model performance variation (UAR and UF_1 in [\%]) between centralised learning and federated learning

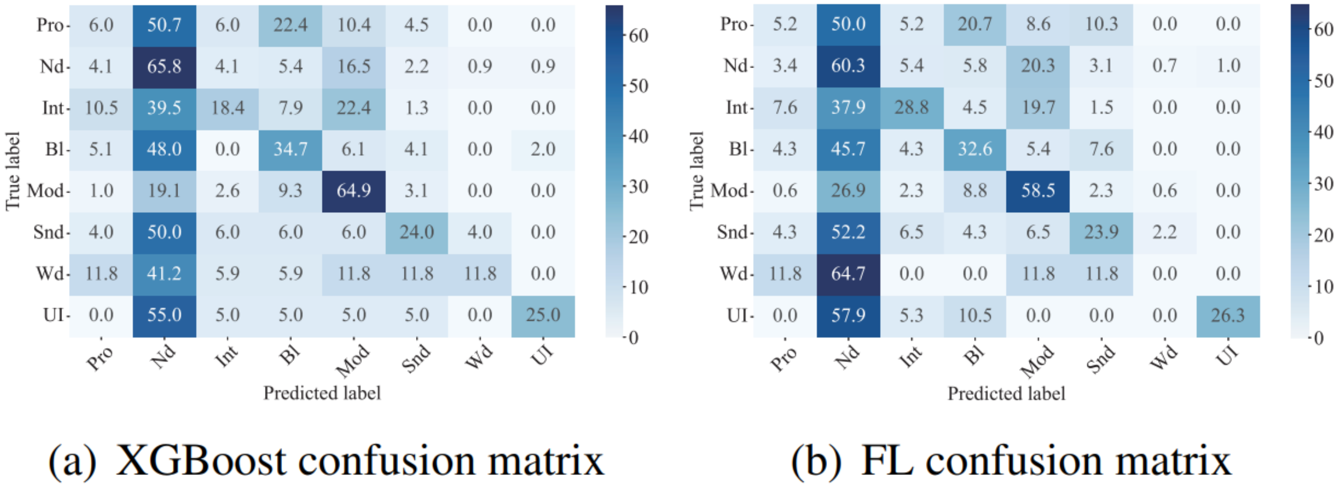
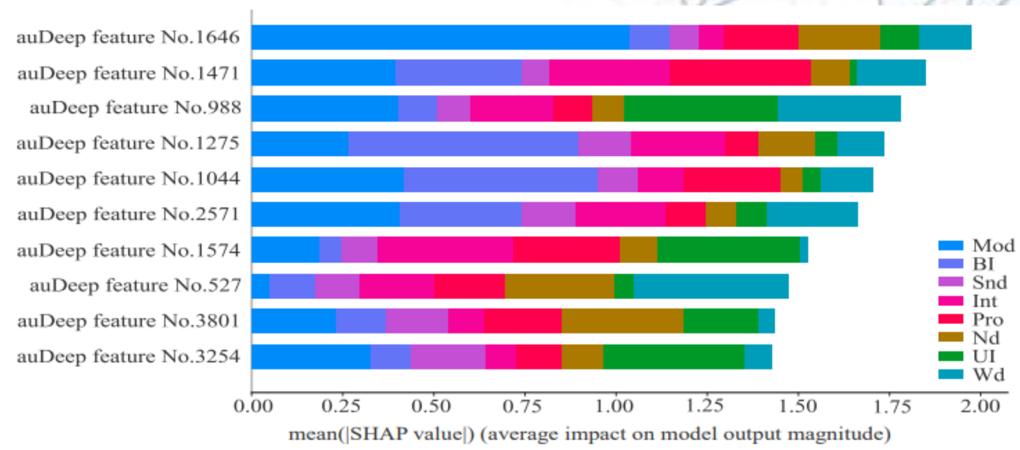


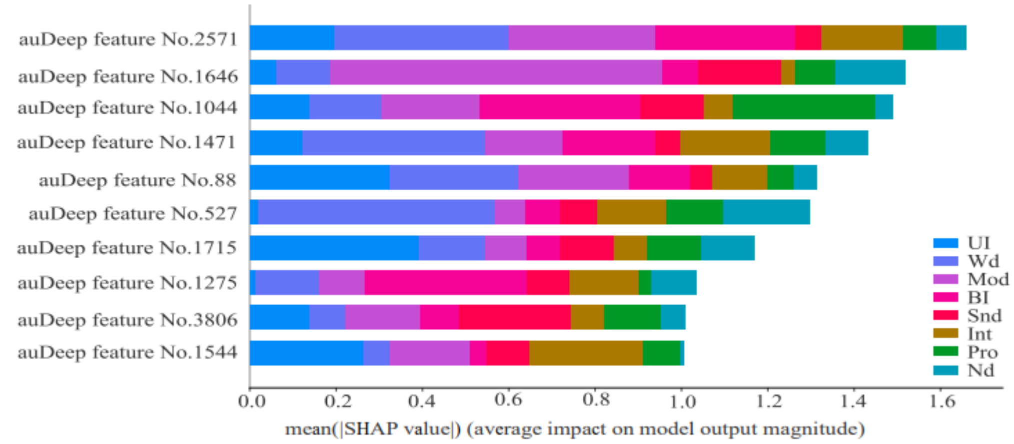
Fig.3 Normalised confusion matrix (in [\%]) of true labels and predicted labels between centralised learning and federated learning.

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[1], Wanyong Qiu[1], Chen Quan[1], Kun Qian[1*], Zhihua Wang[2], Yu Ma[1], Bin Hu[1*], Bjoern W. Schuller[3] and Yoshiharu Yamamoto[2]

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