EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

EmoFedProto: Privacy-Preserving Vietnamese Speech Emotion Recognition via Prototype-Based Federated Learning

Quang-Anh Nguyen-Duc, Duc Minh Pham, Thai Dinh Kim *, Thao Phuong Pham, Minh-Anh Nguyen, Xuan-Hai Le, Van-Ninh Nguyen

International School, Vietnam National University, Hanoi, Vietnam

Abstract

Speech Emotion Recognition (SER) plays a fundamental role in affective computing by enabling machines to infer human emotional states from vocal expressions. However, most existing SER systems rely on centralized training paradigms, which raise serious privacy concerns due to the sensitive nature of speech data. Federated Learning (FL) offers a privacy-preserving alternative by allowing collaborative model training without sharing raw data, yet its performance often degrades significantly under non-IID data distributions, a common characteristic of speech emotion datasets caused by speaker variability and emotion imbalance. To address these challenges, we propose *EmoFedProto*, a prototype-based federated learning framework with clustering-enhanced prototype aggregation tailored for Vietnamese speech emotion recognition in low-resource settings. Instead of exchanging full model parameters, EmoFedProto communicates class-level feature prototypes, enabling more robust alignment across heterogeneous clients. Experiments conducted on the VNEMOS dataset under realistic non-IID and few-shot conditions demonstrate that EmoFedProto achieves an accuracy of 0.875, outperforming the baseline FedProto (0.825), while reducing performance variability by 44%. These results indicate that clustering-based prototype federated learning is an effective and communication-efficient solution for privacy-preserving speech emotion recognition, particularly in low-resource languages and real-world federated environments.

Received on 16 January 2026; accepted on 19 April 2026; published on 23 April 2026

Keywords: Speech Emotion Recognition, Federated Learning, Prototype-Based Learning, Non-IID Data, Low-Resource Languages, Vietnamese Speech

doi:10.4108/airo.11595

1. Introduction

Speech Emotion Recognition (SER) is a fundamental component of affective computing, enabling machines to infer human emotional states from vocal expressions. It supports a wide range of applications, including emotionally aware virtual assistants, mental health assessment systems, and personalized human–computer interaction [1]. Despite recent advances, most SER models rely on centralized training paradigms that require large-scale annotated speech datasets. Such approaches pose serious privacy and confidentiality risks, as speech signals inherently contain sensitive personal and biometric information.

Despite significant progress in SER for high-resource languages such as English, the development of robust emotion recognition systems for low-resource languages remains a substantial challenge. Low-resource languages are characterized by the scarcity of publicly available annotated data. Vietnamese, spoken by approximately 90 million people worldwide, presents unique difficulties for speech processing tasks due to its inherent linguistic characteristics [2]. As a monosyllabic and tonal language with six lexical tones, Vietnamese exhibits complex acoustic properties where pitch variations serve both linguistic

*Corresponding author. Email: thaikd@vnu.edu.vn.
Author emails: anhnd@vnuis.edu.vn, pmduc2808@gmail.com, 22070265@vnu.edu.vn, 22071104@vnu.edu.vn, hailx@vnu.edu.vn ninhnv@vnuis.edu.vn

and emotional functions simultaneously [3, 4]. This dual role of tonal features creates significant ambiguity in distinguishing emotional states from lexical tones, particularly between acoustically similar emotions such as anger and panic, which often exhibit overlapping patterns in pitch and energy [1].

Federated Learning (FL) has emerged as a promising decentralized learning paradigm in which multiple clients collaboratively train a global model without sharing raw data [5]. This privacy-preserving property makes FL particularly attractive for speech-based emotion recognition. However, the effectiveness of FL is often compromised under non-IID data distributions, which are common in SER due to speaker-dependent characteristics, diverse recording environments, and highly imbalanced emotional class distributions.

Recent studies have explored various strategies to mitigate these challenges, including semi-supervised federated learning [6], multimodal emotion recognition frameworks [7], and privacy-aware models based on physiological signals [8]. In parallel, the broader field of emotion and pattern recognition has witnessed significant breakthroughs in visual domains. For instance, Facial Emotion Recognition (FER) has been substantially improved through novel feature extraction techniques such as scattering wavelets [9], as well as comprehensive evaluations comparing modern Vision-Language Models against traditional deep learning architectures [10]. Furthermore, the critical need for deploying models in resource-constrained or real-time environments has driven the development of highly efficient, lightweight multi-scale architectures [11], providing valuable architectural insights for low-resource tasks. While these methods demonstrate encouraging results, research on federated learning for speech-only emotion recognition, especially in low-resource languages, remains limited. In particular, existing FL approaches struggle to maintain robust performance when confronted with severe data heterogeneity across distributed clients.

To address this gap, we propose *EmoFedProto*, a prototype-based federated learning framework adapted from FedProto [12] for speech emotion recognition. Instead of exchanging full model parameters, EmoFedProto communicates class-level feature prototypes that capture semantic representations of emotional categories. This design improves global alignment across heterogeneous clients and enhances generalization under non-IID conditions. We evaluate the proposed framework on the VNEMOS dataset [13], a Vietnamese emotional speech corpus representative of low-resource language scenarios, as well as the widely recognized German EmoDB dataset to validate its cross-lingual generalizability. The objective of this study is to develop a privacy-preserving SER model that effectively handles data heterogeneity and communication constraints while remaining suitable for real-world cross-silo federated deployments.

Contributions. The main contributions of this paper are summarized as follows:

(1) We propose *EmoFedProto*, a privacy-preserving prototype-based federated learning framework specifically designed for speech emotion recognition (SER) in low-resource and data-heterogeneous settings.

(2) We introduce a clustering-enhanced server-side prototype aggregation strategy that maintains multiple global prototypes per emotion class, enabling more effective modeling of diverse emotional expressions across heterogeneous (non-IID) clients.

(3) We conduct comprehensive experiments on the VNEMOS and EmoDB datasets under realistic non-IID and few-shot federated conditions. We demonstrate that EmoFedProto consistently outperforms standard federated baselines (FedAvg, FedProx, SCAFFOLD) and the original FedProto, improving recognition accuracy on VNEMOS to 0.875 while reducing performance variability across runs and clients by 44%.

(4) We provide an empirical comparison across multiple backbone architectures, including audio-native models (Wav2Vec2, HuBERT, AST), ResNet, Swin Transformer, MobileNetV3, and Vision Transformer (ViT), to justify the effectiveness of the proposed design choices for federated SER under resource-constrained environments.

The remainder of this paper is organized as follows: Section 2 describes the methodology of the proposed EmoFedProto framework and the clustering-enhanced aggregation strategy. Section 3 details the experimental datasets, pre-processing pipeline, and environment. Section 4 presents the experimental results and extensive ablation studies. Finally, Section 5 concludes the paper and discusses future research directions.

2. Methodology

2.1. Federated Learning

Federated Learning (FL), introduced by McMahan et al. [5], enables multiple clients to collaboratively train a global model without sharing raw data, making it particularly suitable for privacy-sensitive applications such as speech emotion recognition. In the canonical FedAvg algorithm [5], each client performs local stochastic gradient descent on its private data and

transmits only model parameter updates to a central server, which aggregates them via weighted averaging. While effective under IID conditions, FedAvg suffers significant performance degradation under non-IID data distributions, where heterogeneous local objectives cause client models to drift apart during local training. To address this, FedProx [14] introduces a proximal regularization term that penalizes the deviation of local model parameters from the global model, limiting client drift and improving convergence stability under heterogeneous data. SCAFFOLD [15] takes a different approach by introducing control variates to correct for client drift at the gradient level, providing variance reduction across clients and faster convergence guarantees under non-IID conditions.

Federated Prototype Learning (FedProto), proposed by Tan *et al.* [16], addresses data heterogeneity through a fundamentally different mechanism by enabling clients to communicate class-level prototypes—defined as the mean feature embeddings of samples belonging to each class—rather than exchanging full model gradients or parameters. These local prototypes are aggregated at a central server to form global prototypes, which are redistributed to clients as regularization signals during local training to encourage semantic alignment across heterogeneous data distributions and model architectures. Building on this foundation, FedPCL [17] incorporates contrastive learning at the prototype level, using local prototypes as anchor points to pull together same-class embeddings and push apart different-class embeddings across clients, yielding more discriminative global representations. FedNH [18] further addresses prototype instability through a normalized Hadamard-based initialization that ensures prototype vectors are uniformly distributed and mutually orthogonal in the embedding space. While both methods maintain a single global prototype per class, EmoFedProto preserves multiple cluster-level prototypes per emotion class via server-side K-means clustering, allowing each client to align toward its nearest prototype mode rather than a potentially misleading global average. This design is particularly suited to speech emotion recognition, where speaker-dependent acoustic variability causes local prototypes of the same class to form distinct clusters in the feature space.

2.2. Proposed EmoFedProto Framework

EmoFedProto is built upon the FedProto framework [16] and is designed for federated speech emotion recognition under low-resource and non-IID conditions. We consider a federated setting with $K = 4$ clients, where all clients participate in every communication round ($C = 1.0$). The training process runs for $R = 30$ communication rounds, and each client performs $E = 10$ local training epochs per round with a batch size of 8.

The local optimization objective consists of two components: a supervised classification loss $\mathcal{L}_S$, implemented as the negative log-likelihood loss, and a prototype regularization loss $\mathcal{L}_R$, defined as the mean squared error between local and global prototypes. The overall loss function is formulated as

$$\mathcal{L}(\mathbf{D}_i, \omega_i) = \mathcal{L}_S\Big(\mathcal{F}_i(\omega_i; \mathbf{x}), \mathbf{y}\Big) + \lambda\, \mathcal{L}_R\Big(\mathbf{C}_i, \overline{\mathbf{C}}_i\Big) \tag{1}$$

where $\mathbf{D}_i$ denotes the local dataset of client i, ω_i represents the model parameters, $\mathcal{F}_i(\cdot)$ represents the local model's prediction function, $\mathbf{C}_i$ are the local class prototypes, and $\overline{\mathbf{C}}_i$ are the corresponding global prototypes.

Local model updates are optimized using the AdamW optimizer with an initial learning rate of 1×10^{-4}, a weight decay of 1×10^{-4}, and a numerical stability parameter $\epsilon = 10^{-8}$. The learning rate is decayed by a factor of 0.95 every 10 communication rounds, with a minimum threshold of 1×10^{-6}. A cosine annealing schedule is applied across local epochs, and gradient clipping is employed to prevent exploding gradients.

Local class prototypes are computed as the mean feature embeddings of samples belonging to each class:

$$\mathbf{C}_i^{(j)} = \frac{1}{|\mathbf{D}_{i,j}|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathbf{D}_{i,j}} f_i(\boldsymbol{\phi}_i; \mathbf{x}) \tag{2}$$

where $\mathbf{D}_{i,j}$ denotes the subset of local samples belonging to class j, and $f_i(\boldsymbol{\phi}_i; \mathbf{x})$ is the feature embedding function parameterized by $\boldsymbol{\phi}_i$.

2.3. Clustering-Based Prototype Aggregation

To further enhance robustness under severe data heterogeneity, we introduce a clustering-based prototype aggregation strategy at the server side. The key motivation stems from the observation that, in speech emotion recognition, local prototypes for the same emotion class often exhibit a *multi-modal* structure across clients—for instance, male and female speakers typically produce acoustically distinct realizations of the same emotion (*e.g.*, anger expressed through low-pitched shouting vs. high-pitched exclamation). When such multi-modal prototypes are naively averaged, the resulting global prototype falls in a low-density region of the feature space between the modes, providing a misleading regularization signal to all clients. We refer to this failure mode as *prototype collapse*, and formalize it below to justify the clustering-based alternative.

Formalizing prototype collapse. Consider a federated setting with K clients, each contributing a local prototype $\mathbf{C}_i^{(j)}$ for emotion class j.

Definition 1 (Multi-Modal Prototype Distribution). The local prototypes $\{\mathbf{C}_i^{(j)}\}_{i=1}^K$ exhibit an M-modal structure if there exist M disjoint subsets $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M$ of clients, with $\bigcup_{m=1}^M \mathcal{S}_m = \{1, \ldots, K\}$, such that the intra-cluster variance is substantially smaller than the inter-cluster variance:

$$\begin{aligned} \sigma_{\text{intra}}^2 &\triangleq \sum_{m=1}^{M} \alpha_m \sum_{i \in \mathcal{S}_m} \frac{w_i}{W_m} \|\mathbf{C}_i^{(j)} - \boldsymbol{\mu}_m\|^2 \\ &\ll \sigma_{\text{inter}}^2 \triangleq \sum_{m=1}^{M} \alpha_m \|\boldsymbol{\mu}_m - \boldsymbol{\mu}\|^2 \end{aligned} \tag{3}$$

where $\boldsymbol{\mu}_m = \sum_{i \in \mathcal{S}_m} \frac{w_i}{W_m} \mathbf{C}_i^{(j)}$ is the weighted centroid of cluster m, $W_m = \sum_{i \in \mathcal{S}_m} w_i$, $\alpha_m = W_m / W$ are the cluster weight proportions, $W = \sum_{i=1}^K w_i$, and $\boldsymbol{\mu} = \sum_{i=1}^K \frac{w_i}{W} \mathbf{C}_i^{(j)}$ is the overall weighted mean.

Under this structure, standard FedProto's weighted averaging incurs an irreducible alignment error:

Proposition 1 (Alignment Error of Weighted Averaging). Let $\bar{\mathbf{C}}^{(j)} = \sum_{i=1}^K \frac{w_i}{W} \mathbf{C}_i^{(j)}$ be the weighted average prototype used in standard FedProto. The expected squared alignment error satisfies:

$$\mathcal{E}_{\text{avg}} \triangleq \sum_{i=1}^{K} \frac{w_i}{W} \|\mathbf{C}_i^{(j)} - \bar{\mathbf{C}}^{(j)}\|^2 = \sigma_{\text{intra}}^2 + \sigma_{\text{inter}}^2 \tag{4}$$

In particular, when $M = 2$ and $\alpha_1 \approx \alpha_2 \approx 0.5$, the averaged prototype lies at a distance of $\approx \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|/2$ from each mode.

Proof. By the law of total variance, the weighted mean squared deviation can be decomposed as follows:

$$\begin{aligned} \sum_{i=1}^{K} \frac{w_i}{W} \|\mathbf{C}_i^{(j)} - \bar{\mathbf{C}}^{(j)}\|^2 = &\underbrace{\sum_{m=1}^{M} \alpha_m \sum_{i \in \mathcal{S}_m} \frac{w_i}{W_m} \|\mathbf{C}_i^{(j)} - \boldsymbol{\mu}_m\|^2}_{\sigma_{\text{intra}}^2} \\ &+ \underbrace{\sum_{m=1}^{M} \alpha_m \|\boldsymbol{\mu}_m - \bar{\mathbf{C}}^{(j)}\|^2}_{\sigma_{\text{inter}}^2} \end{aligned} \tag{5}$$

This decomposition directly yields the expression in (4). Since $\bar{\mathbf{C}}^{(j)} = \sum_{m=1}^M \alpha_m \boldsymbol{\mu}_m$ (when intra-cluster deviation is negligible), the global prototype is a convex combination of the centroids. For $M = 2$, this gives $\|\bar{\mathbf{C}}^{(j)} - \boldsymbol{\mu}_1\| = \alpha_2 \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ and $\|\bar{\mathbf{C}}^{(j)} - \boldsymbol{\mu}_2\| = \alpha_1 \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$, both yielding $\approx \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|/2$ when $\alpha_1 \approx \alpha_2$. □

Proposition 1 shows that the inter-cluster variance σ_{inter}^2 constitutes an irreducible component of the alignment error under weighted averaging, as it cannot be reduced by collecting more samples or adding more clients. This mathematically motivates the necessity of partitioning the prototypes before aggregation.

Aggregation procedure. For each emotion class j, local prototypes received from participating clients, along with their corresponding sample counts, are collected. K-means clustering is applied to partition these prototypes into at most $k = 2$ clusters; when the number of available prototypes is insufficient, k is automatically reduced to ensure numerical stability. Within each cluster c, a global prototype is computed as a weighted average of local prototypes, where the weights are proportional to the number of samples contributing to each prototype:

$$\tilde{\mathbf{C}}_c^{(j)} = \frac{\sum_{i \in \mathcal{N}_{j,c}} w_i \mathbf{C}_i^{(j)}}{\sum_{i \in \mathcal{N}_{j,c}} w_i} \tag{6}$$

where $\mathcal{N}_{j,c}$ denotes the set of clients whose prototypes belong to cluster c for class j, $w_i = |\mathbf{D}_{i,j}|$ is the sample count of client i for class j, and $\mathbf{C}_i^{(j)}$ is the corresponding local prototype. Up to k global prototypes are maintained for each class, and each client is regularized toward its nearest cluster centroid $\tilde{\mathbf{C}}_{c^*(i)}^{(j)}$ where $c^*(i) = \arg\min_c \|\mathbf{C}_i^{(j)} - \tilde{\mathbf{C}}_c^{(j)}\|$. These global prototypes are broadcast back to clients to regularize subsequent local training rounds.

Proposition 2 (Error Reduction via Clustering). Under the same M-modal prototype distribution, if the server applies K-means clustering with $k \geq M$ and each client i is regularized toward its nearest cluster centroid, then the expected alignment error satisfies

$$\mathcal{E}_{\text{cluster}} \triangleq \sum_{i=1}^{K} \frac{w_i}{W} \|\mathbf{C}_i^{(j)} - \tilde{\mathbf{C}}_{c^*(i)}^{(j)}\|^2 \leq \sigma_{\text{intra}}^2 \tag{7}$$

yielding an error reduction of $\Delta\mathcal{E} \geq \sigma_{\text{inter}}^2$ compared to standard weighted averaging.

Proof. By definition, the K-means algorithm partitions the local prototypes to minimize the within-cluster sum of squared distances. The ground-truth modal partition $\{\mathcal{S}_m\}_{m=1}^M$ represents one valid grouping configuration. Therefore, the alignment error achieved by the optimal K-means clustering is strictly upper-bounded by the error of this ground-truth partition:

$$\mathcal{E}_{\text{cluster}} \leq \sum_{m=1}^{M} \alpha_m \sum_{i \in \mathcal{S}_m} \frac{w_i}{W_m} \|\mathbf{C}_i^{(j)} - \boldsymbol{\mu}_m\|^2 = \sigma_{\text{intra}}^2 \tag{8}$$

Comparing this with Eq. (4), the clustering strategy entirely eliminates the inter-cluster variance σ^2_{inter}. When the multi-modal structure is well-separated across clients (*i.e.*, $\sigma^2_{\text{inter}} \gg \sigma^2_{\text{intra}}$), this error reduction is substantial, thereby providing a significantly more accurate regularization signal. □

2.4. Model Architecture

EmoFedProto adopts a ViT backbone to extract high-level representations from speech signals. Specifically, we employ the pre-trained `vit_l_16` model from the `torchvision` library [19, 20], which produces 1024-dimensional feature embeddings. The choice of ViT over conventional CNN architectures is motivated by the nature of our input representation. Since raw speech utterances are converted into MFCC spectrograms and treated as 2D images, ViT's patch-based self-attention mechanism is particularly well-suited to this setting. Unlike CNNs, which rely on local receptive fields and capture only spatially adjacent features, ViT partitions the spectrogram into non-overlapping patches and models global dependencies across all patches simultaneously through multi-head self-attention. This allows the model to capture long-range temporal dynamics in the MFCC representation, such as pitch contour evolution and energy distribution shifts across the full utterance, which are critical cues for distinguishing emotional states. The classification head consists of a linear projection layer followed by a ReLU activation, a dropout layer for regularization, and a final linear layer that outputs log-probabilities for the target emotion classes. To reduce computational overhead in federated settings, the ViT backbone can be optionally frozen during training. In all cases, class prototypes are extracted from the 1024-dimensional embeddings prior to the classification head.

Input Representation and Patch Embedding. In the context of speech emotion recognition, input utterances are first converted into Mel-Frequency Cepstral Coefficient (MFCC) spectrograms, which are treated as 2D inputs analogous to images. Let $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ denote an MFCC spectrogram, where H, W, and C represent its height, width, and number of channels, respectively. The spectrogram is partitioned into $N = \frac{HW}{P^2}$ non-overlapping patches of size $P \times P$. Each patch is flattened into a vector $\mathbf{x}_p \in \mathbb{R}^{P^2C}$ and projected into a D-dimensional embedding space via a learnable linear projection matrix $\mathbf{E} \in \mathbb{R}^{(P^2C)\times D}$.

To enable global context modeling, a learnable class token $\mathbf{z}_{\text{cls}} \in \mathbb{R}^D$ is prepended to the sequence of patch embeddings. Positional embeddings $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$ are then added to preserve spatial information:

$$\mathbf{Z}_0 = [\mathbf{z}_{\text{cls}}; \mathbf{x}_1\mathbf{E}; \ldots; \mathbf{x}_N\mathbf{E}] + \mathbf{E}_{\text{pos}} \tag{9}$$

Transformer Encoder. The embedded sequence $\mathbf{Z}_0$ is processed by L stacked transformer encoder layers. At layer ℓ (where $\ell \in \{1, \ldots, L\}$), Multi-Head Self-Attention (MHSA) is first applied. For each attention head $i \in \{1, \ldots, h\}$, the query, key, and value matrices are computed as

$$\mathbf{Q}_i = \mathbf{Z}_{\ell-1}\mathbf{W}_i^Q, \quad \mathbf{K}_i = \mathbf{Z}_{\ell-1}\mathbf{W}_i^K, \quad \mathbf{V}_i = \mathbf{Z}_{\ell-1}\mathbf{W}_i^V \tag{10}$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{D\times d_h}$ are learnable projection matrices. The attention output for head i is computed as

$$\text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_h}}\right)\mathbf{V}_i \tag{11}$$

where $d_h = D/h$ denotes the dimensionality of each attention head.

Outputs from all heads are concatenated and linearly projected to form the MHSA output:

$$\text{MHSA}(\mathbf{Z}_{\ell-1}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \tag{12}$$

A residual connection and layer normalization are then applied, followed by a position-wise Feed-Forward Network (FFN):

$$\mathbf{Z}'_\ell = \text{LayerNorm}\Big(\mathbf{Z}_{\ell-1} + \text{MHSA}(\mathbf{Z}_{\ell-1})\Big) \tag{13}$$

$$\text{FFN}(\mathbf{x}) = \text{ReLU}(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{14}$$

$$\mathbf{Z}_\ell = \text{LayerNorm}\Big(\mathbf{Z}'_\ell + \text{FFN}(\mathbf{Z}'_\ell)\Big) \tag{15}$$

After the final encoder layer L, the representation corresponding to the class token from $\mathbf{Z}_L$ is used as the global embedding for the input utterance. This embedding serves as the feature representation for both emotion classification and prototype computation in the EmoFedProto framework.

3. Environment & Dataset

3.1. VNEMOS Dataset

In this research, we employ the Vietnamese Emotion Speech Dataset (VNEMOS) [13]. VNEMOS features a hybrid collection of both acted and natural speech. The acted data was sourced from professional actors in films and television series, while the natural speech was captured from spontaneous interactions in live broadcasts. In total, the dataset contains 250 audio segments (exactly 50 samples per emotion class) from 27 distinct media sources, amounting to approximately 30 minutes of content; furthermore, Fig. 2 illustrates the gender distribution of the dataset. It is carefully balanced across five fundamental emotions of anger, happiness, sadness, neutral, and anxiety, establishing it

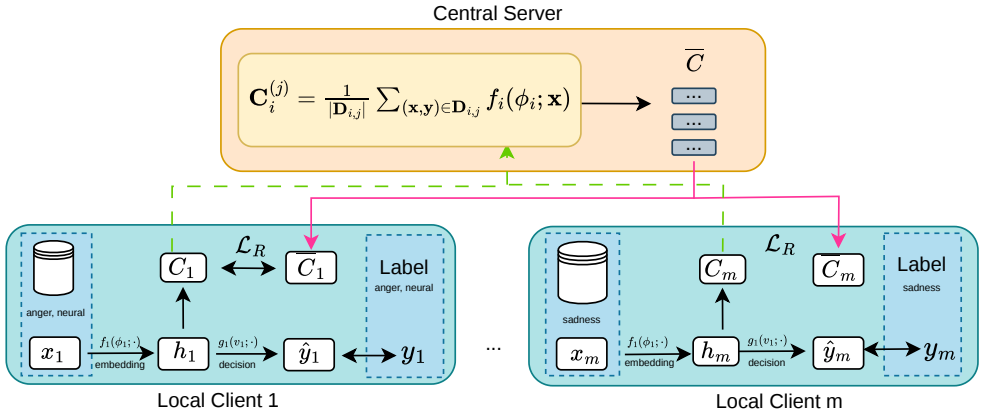

Figure 1. Overview of the proposed EmoFedProto framework, illustrating local prototype computation at clients, clustering-based prototype aggregation at the server, and global prototype redistribution for regularized federated training

as a reliable benchmark for speech emotion recognition models in Vietnamese. Fig. 3 shows the soundwave shape overtimes of each emotion classes. The full dataset is publicly available at https://bit.ly/VNEMOS_data.

Crucially, to prevent potential data leakage arising from the mixture of 27 distinct media sources, we implemented a strict source-aware data partitioning protocol. We ensured that audio segments originating from the same media source do not appear simultaneously in the training and testing sets, nor do they overlap across the local datasets of different federated clients. This rigorous isolation guarantees that the model learns generalized emotional representations rather than merely memorizing source-specific acoustic signatures, totaling 50 samples per category to ensure a perfectly balanced distribution throughout the evaluation.

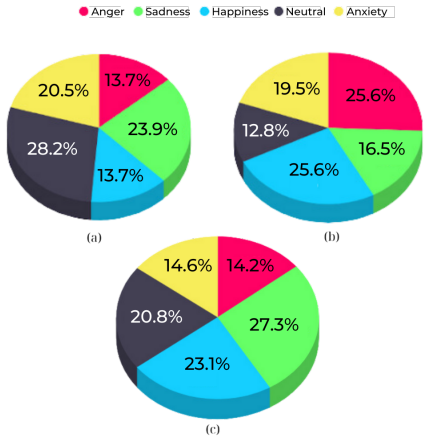

Figure 2. VNEMOS Emotions Data length Distribution (a) Male, (b) Female, (c) VNEMOS

3.2. EmoDB Dataset

To assess the generalizability, we also conduct additional experiments on the Berlin Database of Emotional Speech (EmoDB) [22], a widely adopted benchmark for speech emotion recognition in German. EmoDB comprises 535 audio recordings from 10 professional actors, covering seven emotion categories with anger, boredom, disgust, fear, happiness, neutral, and sadness. Unlike VNEMOS, EmoDB exhibits a naturally imbalanced class distribution, with anger being the most represented category and disgust the least. This imbalance, combined with the increased number of emotion categories, makes EmoDB a more challenging and complementary benchmark for evaluating federated speech emotion recognition under heterogeneous conditions.

3.3. Data Partitioning and Federated Simulation

Following the task-heterogeneous non-IID partitioning protocol of FedProto [12], the dataset is distributed across $K = 4$ clients such that each client is assigned a distinct subset of emotion classes with a limited number of training samples per class. Specifically, the number of classes per client n_i is sampled from $\{2, 3\}$ under a 3-way ($W = 3$, $\delta = 1$) configuration, and the number of training samples per class k_i is sampled from $\{15, 16\}$ shots. The union of assigned classes across all clients covers the full emotion taxonomy, while each individual client observes only a partial label space, reflecting realistic institutional data silos. Within each client, the assigned samples are divided further into a local training set (80%) and a local evaluation set (20%) using a stratified split, ensuring that at least one sample per assigned class appears in each partition.

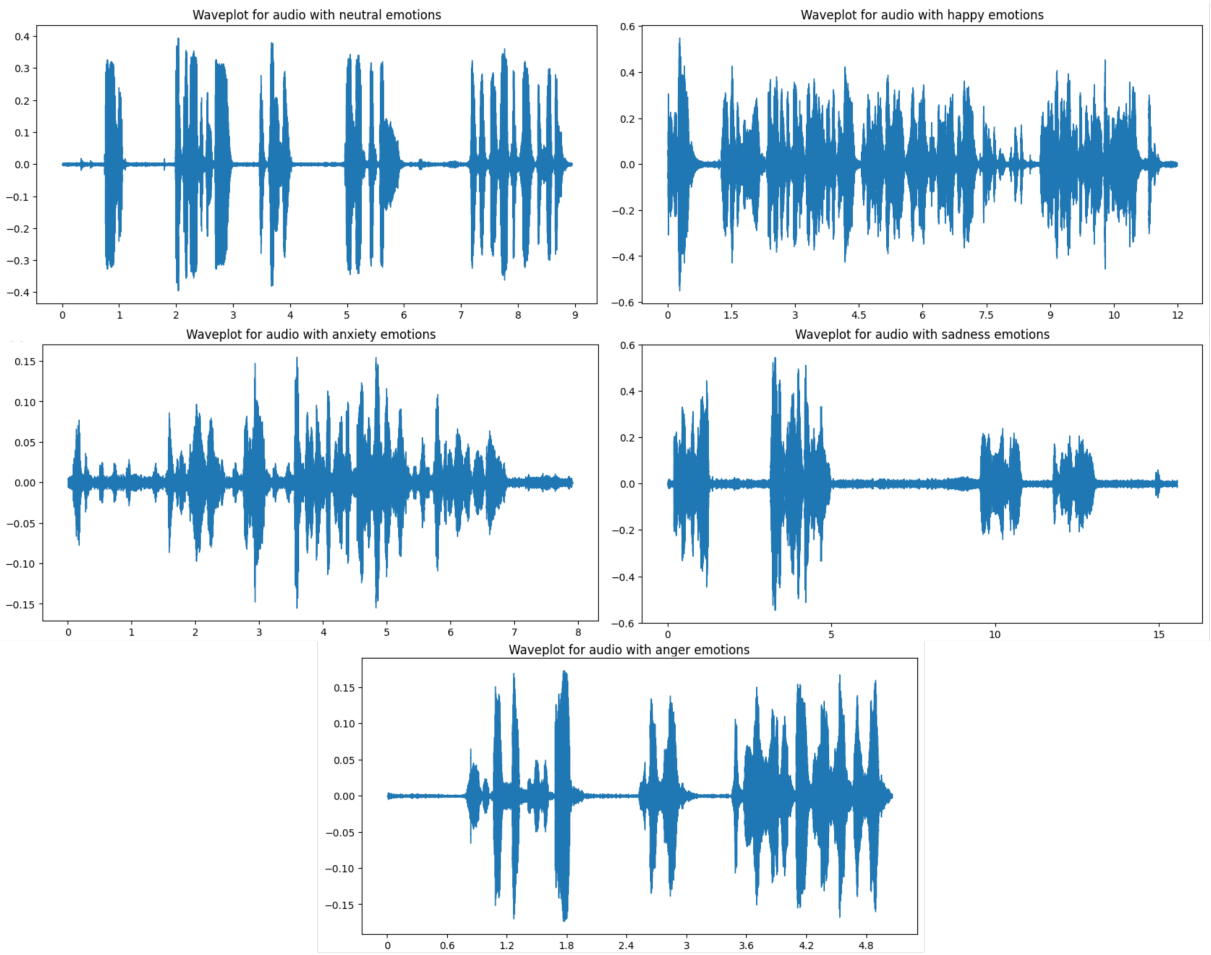

Figure 3. Speech Emotions Sound Wave from Five Classes in VNEMOS with x Axis for Times and y Axis for Amplitude

3.4. Pre-processing

The raw acoustic waveforms of the VNEMOS corpus were transformed into a feature space optimized using a multistage pipeline. Initially, each time-domain signal underwent sampling rate conversion to a canonical frequency of 16kHz, this step ensures spectral consistency across the entire dataset and optimizes computational overhead. Afterwards, the amplitudes of the signals were peak-normalized in order to mitigate energy-level variations that are not related to emotional expression. The continuous signal was segmented into 25 ms-long frames with a frame shift of 10 ms in order to analyze the temporal evolution of spectral content. The frame shift was then applied to reduce spectral leakage artifacts during the subsequent Fourier analysis by multiplying each frame by a Hamming window function.

We extracted Mel-Frequency Cepstral Coefficients (MFCCs), a representation of cepstral features that is perceptually weighted. This extraction process extracted by following

1. Applying the Short-Time Fourier Transform (STFT) to map the signal from the time domain to the time-frequency domain, yielding a power spectrogram

2. Convolving the spectrogram with a Mel filterbank to emulate the non-linear frequency response of the human cochlea

3. Applying logarithmic compression to the Mel-band energies

4. Performing a Discrete Cosine Transform (DCT) to decorrelate the filterbank energies and produce a compact feature representation

Given the heterogeneous durations of the data samples, a fixed-size input tensor was required for the neural network. We standardized the input length to 5 seconds. Shorter sequences were extended using symmetrical padding. This method reflects the signal at its boundaries to fulfill the required temporal length.

3.5. Environment

All experiments were conducted on a workstation running Windows 11 Pro, equipped with an Intel Core i5-12400F CPU, 32 GB of RAM, and an NVIDIA GeForce RTX 4070 Ti GPU. The implementation was based on Python 3.11.9 and PyTorch 2.3.0, with GPU acceleration enabled via the CUDA 11.8 toolkit.

For optimization, we employed the AdamW optimizer, the prototype regularization weight is set to $\lambda = 0.01$, consistent with the training configuration described in Section III, using an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. A cosine annealing learning rate schedule was applied across local training epochs, with the learning rate decayed by a factor of 0.95 every 10 communication rounds and bounded below by 1×10^{-6}. Gradient clipping was additionally applied to improve training stability.

3.6. Evaluation Metrics

To evaluate the performance of EmoFedProto and the baseline models, we employ Accuracy and F1-score as the primary metrics.

Accuracy (Acc) measures the proportion of correctly predicted emotion instances among the total number of samples, defined as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

While Accuracy provides a general overview of model performance, the F1-score is utilized to assess the balance between precision and recall, especially for the imbalanced EmoDB dataset. The F1-score is the harmonic mean of Precision and Recall, calculated as:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{17}$$

In this study, we report the macro-averaged F1-score to ensure that each emotion class contributes equally to the final metric, preventing minority classes from being overshadowed by dominant ones.

4. Experiments and Discussion

4.1. Experimental Results

The proposed EmoFedProto framework was evaluated on the VNEMOS dataset for speech emotion recognition under a federated learning setting. Its performance was compared with two representative centralized approaches reported in the literature [13, 23], serving as reference baselines. Table 1 summarizes the accuracy results.

EmoFedProto achieved an accuracy of 0.87, outperforming Work 2 (0.86) [23] and Work 1 (0.83) [13]. These results indicate that the proposed clustering-based prototype aggregation strategy can effectively mitigate the impact of non-IID data distributions in a federated environment with four clients, each configured under a 3-way, 15-shot learning setting. While the compared methods are centralized and therefore not directly optimized for federated learning, the competitive performance of EmoFedProto demonstrates its potential for robust speech emotion recognition in distributed and privacy-sensitive scenarios. A more comprehensive evaluation with additional federated baselines and metrics is provided in subsequent analyses.

Table 1. Accuracy comparison on the VNEMOS dataset for speech emotion recognition

Model	Work 1 [13]	Work 2 [23]	**EmoFedProto**
Accuracy	0.83	0.86	**0.87**

Comparison of Baseline and Modification. Table 2 presents the performance comparison among federated learning baselines under the same non-IID federated configuration. A mong t he b aselines, F edAvg [5] achieves the lowest accuracy of 0.758 with the highest variability of 0.147, reflecting i ts w ell-known sensitivity to heterogeneous data distributions. FedProx [14] and SCAFFOLD [15] achieve accuracies of 0.800 and 0.833 respectively, with progressively reduced standard deviations. Meanwhile, FedProto [16] attains 0.825 but exhibits a high variability (0.147) comparable to that of FedAvg [5].

Table 2. Performance comparison between methods under a federated non-IID setting on the VNEMOS dataset

Method	Accuracy	± Accuracy
FedAvg [5]	0.758	0.147
FedProx [14]	0.800	0.124
SCAFFOLD [15]	0.833	0.100
FedProto [16]	0.825	0.147
EmoFedProto	**0.875**	**0.083**

EmoFedProto achieves the highest accuracy of 0.875, showing an absolute improvement of 5.0 percentage points over FedProto [16] and 11.7 percentage points over FedAvg [5]. Beyond accuracy, EmoFedProto demonstrates substantially improved training stability, reducing the standard deviation from 0.147 to 0.083—a 44% reduction in performance variability compared to FedProto [16]. These results indicate that the proposed clustering-based prototype aggregation strategy effectively mitigates the adverse effects of non-IID data distributions.

Evaluation on the EmoDB Dataset. Table 3 compares EmoFedProto against federated baselines on the EmoDB dataset [22] under non-IID settings. EmoFedProto achieves the highest F1-score of 0.865 and ties with SCAFFOLD [15] for the best accuracy at 0.866, outperforming the original FedProto [16] by 3.3% in accuracy and 2.3% in F1-score. Among the baselines, FedProx [14] achieves the lowest variability (±0.059 for accuracy and ±0.051 for F1-score) through proximal regularization, but it sacrifices p eak a ccuracy (0.858). SCAFFOLD [15] matches EmoFedProto in accuracy yet exhibits high variability (±0.122), indicating uneven corrections across clients. EmoFedProto improves upon standard FedProto [16] by replacing naive prototype averaging with weighted K-Means clustering, which aggregates cluster centers weighted by sample counts to produce more robust global prototypes under non-IID distributions.

Table 3. Performance comparison between methods under a federated non-IID setting on the EmoDB dataset

Method	Acc	± Acc	F1	± F1
FedAvg [5]	0.824	0.102	0.833	0.122
FedProx [14]	0.858	**0.059**	0.830	**0.051**
SCAFFOLD [15]	**0.866**	0.122	0.852	0.145
FedProto [16]	0.833	0.122	0.842	0.101
EmoFedProto	**0.866**	0.122	**0.865**	0.124

Furthermore, EmoFedProto surpasses SCAFFOLD [15] in F1-score (0.865 versus 0.852) despite achieving an equal overall accuracy. This indicates a more balanced per-class performance, as prototype-based methods explicitly maintain per-class representations that prevent minority emotion classes from being absorbed by dominant classes during global aggregation.

4.2. Ablation Study

Comparison of Different Architectures. Table 4 presents the accuracy and cross-client variability of seven different backbone architectures evaluated within the EmoFedProto framework. The proposed ViT backbone [20] achieves the highest accuracy of 0.875 with a low variability of ±0.083. Notably, audio-specific models such as Wav2Vec2 [27], HuBERT [28], and AST [29] all underperform ViT [20], despite being pre-trained on large-scale audio corpora. We attribute this to the efficacy of MFCC preprocessing, which provides a compact, perceptually motivated representation that explicitly encodes the spectral envelope information relevant to emotion. Self-supervised models like Wav2Vec2 [27] and HuBERT [28] learn features from raw waveforms through objectives designed for speech content recognition; consequently, their representations tend to favor phonetic discrimination over the prosodic and timbral cues critical for emotion recognition. While AST [29] uses Mel spectrograms internally, it lacks the decorrelation and dimensionality reduction provided by the Discrete Cosine Transform (DCT) in MFCCs. Furthermore, all three audio-specific models exhibit higher performance variability (±0.144, ±0.175, and ±0.100, respectively), indicating that MFCC features yield more stable prototypes under non-IID distributions.

Table 4. Accuracy and performance variability of different backbone architectures under the EmoFedProto framework

Model	Accuracy	± Accuracy
CNN	0.750	0.086
ResNet [24]	0.830	0.070
Swin Transformer [25]	0.833	0.173
MobileNetV3 [26]	0.750	**0.055**
Wav2Vec [27]	0.775	0.144
HuBert [28]	0.825	0.175
AST [29]	0.833	0.100
ViT (Proposed) [20]	**0.875**	0.083

CNN-based models, including a standard CNN and MobileNetV3 [26] (both achieving 0.750) as well as ResNet [24] (0.830), are constrained by their local receptive fields, which limits their ability to model long-range dependencies across the spectrogram. ViT [20] overcomes this limitation through patch-based self-attention, modeling global relationships among all spectrogram patches simultaneously to capture temporal dynamics, such as pitch contour evolution and energy distribution shifts across the full utterance. The absolute improvement of 4.5 percentage points over ResNet [24] confirms that global information flow is critical for speech emotion recognition. Finally, the Swin Transformer [25] achieves a competitive accuracy of 0.833 but exhibits the highest variability at ±0.173, further demonstrating that local windowed attention produces less consistent prototypes for federated aggregation compared to the full self-attention mechanism of ViT [20].

Cluster Sensitivity Analysis. Table 5 reports the performance of EmoFedProto across varying numbers of clusters (k). Setting $k = 1$ recovers the original FedProto [16] aggregation scheme, achieving an accuracy of 0.825 with the highest variability of 0.147. This reflects the instability introduced by prototype collapse under multi-modal client distributions. Setting $k = 2$ yields the optimal accuracy of 0.875 and a substantially reduced standard deviation of 0.083, confirming that two clusters are sufficient to capture the dominant sources of acoustic variability across clients (e.g., gender-driven differences in emotional expression). Increasing the number of clusters to $k = 3$ reduces

the accuracy to 0.850 and increases variability to 0.128, suggesting that the feature space is over-partitioned relative to the available prototype count. At $k = 4$, performance degrades further to 0.816, as each client prototype essentially forms a singleton cluster, causing the cross-client alignment signal to effectively vanish.

Table 5. Performance comparison across number of k

k	Accuracy	± Accuracy
$k = 1$	0.825	0.147
$k = 2$	**0.875**	**0.083**
$k = 3$	0.850	0.128
$k = 4$	0.816	0.119

5. Conclusion and Future Work

In this paper, we proposed EmoFedProto, a prototype-based federated learning framework for privacy-preserving speech emotion recognition. By extending the FedProto paradigm to exchange class-level feature prototypes and incorporating a K-means clustering-based aggregation strategy at the server, EmoFedProto effectively mitigates the adverse effects of non-IID data distributions arising from speaker variability and emotional multi-modality.

Experimental results on the Vietnamese VNEMOS dataset demonstrate that EmoFedProto achieves an accuracy of 0.875, while exhibiting a 44% reduction in performance variability compared to the baseline FedProto framework. Furthermore, extensive evaluations on the larger German EmoDB corpus and comparisons against standard federated baselines (FedAvg, FedProx, and SCAFFOLD) confirm the framework's superior generalizability, robust class-imbalance handling (F1-score of 0.865), and training stability. These findings establish EmoFedProto as an effective and scalable solution for privacy-preserving **cross-silo** federated collaborations, particularly suited for institutions dealing with highly sensitive and low-resource speech datasets.

Future work will explore several directions to further enhance the framework. First, we plan to investigate adaptive weighting mechanisms for prototype alignment to handle extreme cases of client data skewness. Second, we aim to integrate client-specific prototype personalization to further optimize local performance. Finally, deploying and analyzing the framework under formal differential privacy (DP) constraints will be considered to provide strict mathematical privacy guarantees alongside the empirical robustness demonstrated in this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Nguyen, T., Tran, T., & Truong, B. (2026). Human-Guided Reasoning with Large Language Models for Vietnamese Speech Emotion Recognition. *arXiv preprint* arXiv:2604.01711.

[2] Tran, L. T. T., Kim, H. G., La, H. M., & Van Pham, S. (2024). Automatic speech recognition of vietnamese for a new large-scale corpus. *Electronics*, 13(5), 977.

[3] Luong, H. T., & Vu, H. Q. (2016, December). A non-expert Kaldi recipe for Vietnamese speech recognition system. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)* (pp. 51-55).

[4] Thanh, P. V., Huyen, N. T. T., Quan, P. N., & Trang, N. T. T. (2024, April). A robust pitch-fusion model for speech emotion recognition in tonal languages. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 12386-12390). IEEE.

[5] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2021). Advances and open problems in federated learning. *Proceedings of the IEEE*, *109*(1), 40–108.

[6] Tsouvalas, V., Ozcelebi, T., & Meratnia, N. (2022). Privacy-preserving speech emotion recognition through semi-supervised federated learning. *arXiv preprint*.

[7] Nandi, A., & Xhafa, F. (2022). A federated learning method for real-time emotion state classification from multimodal streaming. *Methods*, *204*, 340–347.

[8] Gahlan, N., & Sethia, D. (2024). Federated learning in emotion recognition systems based on physiological signals for privacy preservation: A review. *Multimedia Tools and Applications*.

[9] M. Davari, A. Harooni, A. Nasr, K. Savoji, and M. Soleimani, "Improving recognition accuracy for facial expressions using scattering wavelet," *EAI Endorsed Transactions on AI and Robotics*, vol. 3, 2024. DOI: 10.4108/airo.5145.

[10] V. K. Mulukutla, S. S. Pavarala, S. R. Rudraraju, and S. Bonthu, "Evaluating Open-Source Vision Language Models for Facial Emotion Recognition Against Traditional Deep Learning Models," *EAI Endorsed Transactions on AI and Robotics*, vol. 4, 2025. DOI: 10.4108/airo.8870.

[11] Z. Xue, B. Wang, et al., "FDD-YOLO: A Lightweight Multi-scale Prohibited Items Detection Model," *EAI Endorsed Transactions on AI and Robotics*, 2025. DOI: 10.4108/airo.10277.

[12] Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., & Zhang, C. (2021). FedProto: Federated prototype learning across heterogeneous clients. *arXiv preprint*.

[13] Anh, N. Q., Ha, M. H., Nguyen, Q. C., Thi, T. H. N., Vu, Q., Minh-Duc, D. X., & Dinh, T. K. (2024). VNEMOS: Vietnamese speech emotion inference using deep neural networks. In *Proceedings of the 9th International Conference on Integrated Circuits, Design, and Verification (ICDV)* (pp. 97–101). IEEE.

[14] Sahu, A., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., & Smith, V. (2018). Federated Optimization in Heterogeneous Networks. arXiv: Learning.

[15] Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S.J., Stich, S.U., & Suresh, A.T. (2019). SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. International Conference on Machine Learning.

[16] Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., & Zhang, C. (2022). FedProto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 8, pp. 8432–8440).

[17] Tan, Y., Long, G., Ma, J., Liu, L., Zhou, T., & Jiang, J. (2022). Federated learning from pre-trained models: A contrastive learning approach. *Advances in neural information processing systems*, 35, 19332-19344.

[18] Dai, Y., Chen, Z., Li, J., Heinecke, S., Sun, L., & Xu, R. (2023, June). Tackling data heterogeneity in federated learning with class prototypes. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 6, pp. 7314-7322).

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

[20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint*.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[22] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A Database of German Emotional Speech, *Interspeech* (2005).

[23] Nguyen-Duc, Q.-A., Ha, M. H., Dinh, T. K., Pham, M. D., & Van, N. N. (2024). Emotional Vietnamese speech-based depression diagnosis using dynamic attention mechanism. *arXiv preprint*.

[24] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

[25] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022).

[26] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1314–1324).

[27] Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition. *Interspeech* 2019.

[28] Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29, 3451-3460.

[29] Gong, Y., Chung, Y. A., & Glass, J. (2021). AST: Audio Spectrogram Transformer. *Interspeech* 2021.