Federated Learning for Enhanced Emotion Detection from Bangla Audio in Decentralized Environments

***Abstract*—The remarkable advantages offered by Machine Learning are astounding; however, achieving optimal performance requires an increased volume of data. Identifying emotions in audio is vital for effective social communication and offers insights into evolving human behaviors. Employing a Federated Learning model diminishes computational time, facilitating emotion detection in speech audio with adaptation to diverse speaker characteristics and improved model robustness through decentralized training. Our research addresses the task of emotion classification in speech audio using federated learning on the SUBESCO dataset, a Bangla Emotional speech dataset comprising 7000 files distributed across 20 speakers and 7 emotion classes. To facilitate federated learning, the dataset is partitioned among 4 clients and treated as individual local datasets. Our research proposes a neural network architecture integrating convolutional and recurrent layers for effective feature extraction from audio data. Local models trained on individual clients exhibit varied accuracies, emphasizing the unique challenges within each dataset. Through federated learning, these local learnings are aggregated to enhance the global model, achieving an accuracy of 90.08% after 150 additional epochs. The model's performance is analyzed, considering individual client discrepancies, global model improvement, and overall robustness. Despite challenges such as data heterogeneity and limited computational resources, our findings showcase the potential of federated learning for emotion classification in speech audio, providing valuable insights for future research in this domain.**

***Index Terms*—Federated Learning, Bangla Emotion Classification, Speech Audio, SUBESCO Dataset, Deep Neural Network, Model Aggregation**

I. Introduction

The ability to identify emotions in Bangla speech has a lot of promise to enhance user experiences with technology and communication. A crucial gap in the literature has to be filled because there hasn't been enough in-depth research done on this subject. Considering the language's complex emotional expression, emotion recognition affects interpersonal relationships, technological development, and social welfare. Our research attempts to create a thorough framework for emotion analysis because it is a distinct linguistic problem to interpret the subtleties of emotions in Bangla speech. Our work aims at building strong models that can identify a broad range of emotions in Bangla speech. This approach promises more nuanced assessments and improved customer sentiment comprehension, opening the door for a variety of applications in fields including mental health and customer service. It's not only a language barrier to identify emotions in Bangla speech; it's also a matter of accepting linguistic and cultural diversity in technology. Closing this gap can lead to more inclusive emotional computing. By enabling technology to recognize and react to emotional cues more effectively, emotion recognition in Bangla speech seeks to enhance communication and promote more comfortable and joyful interactions. This study aims to build effective models and applications for the Bangla language, ensuring they understand its nuances and recognize emotions in Bangla audio speech. This can lead to various possibilities, allowing technology to connect with users in meaningful ways.

II. Related Work

Several works were done in this field. This study [1] utilized the IEMOCAP corpus for assessment, focusing on four primary emotions: happiness, sadness, anger, and neutrality. The evaluation was based on unweighted average recall (UAR), a prevalent metric in speech emotion recognition (SER) literature [13], indicating that the LSTM-based classifier slightly outperformed CNNs, achieving a UAR of 54.8% across 200 communication rounds. The results exhibit promise when contrasted with benchmark performances using variational autoencoders with LSTM (60.91%) and CNN-LSTM (60.23%). Another paper [2] utilized the German Corpus (Berlin Database of Emotional Speech), which comprises around 800 sentences classified into seven emotion classes. Constructed using Keras and Theano frameworks, the model achieved optimal results after 38 epochs in 4.12 minutes. The deep neural network (DNN) demonstrated a validation accuracy of 79.14%, with 2068 accurate predictions and 545 misclassifications, and a testing accuracy of 77.51%, with 2181 correct predictions and 633 misclassifications. Furthermore, the DNN exhibited a notable accuracy of 96.97% in speech emotion recognition for testing files, accompanied by an average prediction confidence of 69.55%. In addition, this paper [3] shows that the federated approach can generate broadly applicable SER models, even when dealing with insufficient labeled data and non-i.i.d. distributions. The improvement in recognition rates is notable, achieving an 8.67% enhancement with just 10% labeled data compared to fully supervised federated alternatives. Moreover,

in this study [4] utilized two datasets, IEMOCAP and MSP-Improv, for the development of Speech Emotion Recognition (SER) models. The study introduces SemiFedSER, showcasing its capability to achieve the desired SER performance even with a low local label rate of 20% on IEMOCAP and MSP-Improv datasets. Specifically, the Semi-FedSER framework attains a UAR score of 63.11%, comparable to the fully supervised model's score of 66.70% on the IEMOCAP dataset using the APC speech feature. N. Liu et al. in their study [5] delves into addressing speech emotion recognition's challenge when applied to new domains, proposing "TRaSL" - Transfer Subspace Learning. It deals with unsupervised cross-corpus SER, aligning labeled source speech signals with unlabeled target signals via a projection matrix. By transforming both into a common label space, it ensures similar feature distributions, enabling effective emotion prediction in the target domain. TRaSL outperforms recent cross-corpus SER methods across IEMOCAP, EmoDB, eNTERFACE, and AFEW 4.0 datasets, showcasing improved performance in diverse emotional speech recognition. Prior studies introduce a dynamic fusion framework merging spectrogram-based statistical features with auditory-based empirical features, employing a Kernel Extreme Learning Machine (KELM) as the classifier. Experimentation on Emo-DB and IEMOCAP databases validates this approach, showcasing substantial performance enhancement over current state-of-the-art methods through the integration of auditory-based and spectrogram-based features [6]. This study explores federated learning, both with and without secure aggregators, for supervised and unsupervised speaker recognition. To introduce diverse data without data transmission, a generative adversarial network creates imposter data at the edge. Experimental findings on the Voxceleb-1 dataset reveal the advantages of federated learning in speaker recognition [7]. In another paper, three transformation techniques were explored, with the recurrence plot achieving the highest validation accuracy (86.69%) using the MobileNetV2 model. This approach offers a practical, high-accuracy solution without requiring feature engineering while safeguarding data privacy [16] for widespread deployment [8]. C. Tan, Y. et al. in their paper [9] introduce a federated-based approach for establishing a decentralized local SER model. It computes in a data-efficient FL to train the SER model using labeled samples. The evaluation resulted in an accuracy of 65% to 70% which is lower than the initial SER model accuracy. It managed to maintain a privacy level of (1.92, 10-5)-LDP. The paper highlighted the importance of balancing privacy and accuracy. In addition, this study [10] delves into cross-silo federated acoustic modeling, a strategy to uphold data privacy in ASR systems. It involves collaboration between an ASR service provider and multiple clients, each retaining its own private data storage. The paper compares various federated averaging techniques and investigates the impact of communication [15] frequency. To tackle the challenge of non-identically distributed data among clients, the study introduces client adaptive federated training. This method estimates client-specific transformations to standardize the diverse data across clients. Another study [11] discusses a study comparing different privacy methods in Federated Learning for Speech Emotion Recognition (FL-SER) tasks. It highlights the advantage of tailoring protection methods to specific privacy definitions for better FL-SER accuracy and privacy preservation. The study proposes a broader hypothesis advocating for privacy-aligned protection methods across speech-related FL tasks for improved privacy and accuracy, intending to conduct further experiments in diverse speech-related domains to validate this hypothesis. Furthermore, this study [12] explores the challenge of non-IID data in federated learning and proposes a method to adjust this non-IID nature by employing random client data sampling. It suggests that manipulating the degree of non-IID-ness allows for a flexible balance between cost and quality during training [14]. Initially, enhancing quality in federated learning might increase costs, but through techniques like optimizer configuration, hyperparameter tuning, and regularizer usage, federated learning can achieve quality comparable to IID-level data at lower costs.

III. Model Approaches

For our emotion classification task on the SUBESCO dataset [17], we have employed a neural network architecture with a combination of convolutional and recurrent layers. The model architecture, referred to as model1, has been as follows:

A. Convolutional Layers

- Three convolutional layers with varying filter sizes (2048, 1024, 512) have been used for feature extraction from the audio spectrogram.
- Each convolutional layer has been followed by max-pooling, batch normalization, and ReLU activation.

B. Long Short-Term Memory (LSTM) Layers

- Two LSTM layers (256 units followed by 128 units) have been incorporated to capture temporal dependencies in the sequential audio data.

C. Dense Layers with Dropout

- Dense layers with decreasing sizes (128, 64, 32, 16, 8) have been added for higher-level abstraction.
- Dropout layers (0.5 for dense layers, 0.2 for the last two) have been introduced to prevent overfitting.

D. Output Layer

- The final dense layer with softmax activation has facilitated multi-class classification into 7 emotion classes.

This intricate architecture has been designed to extract complex features from audio data, making it well-suited for the nuanced task of emotion classification in speech audio. The combination of convolutional and recurrent layers has enabled the model to discern subtle patterns indicative of different emotional states, contributing to its overall effectiveness.

IV. Methodology

A. Dataset Description

In this study, we utilized the SUBESCO dataset, a comprehensive Bangla Emotional speech dataset. SUBESCO comprises 7,000 audio files, featuring 20 speakers, with a remarkable 350 recordings per speaker, totaling 7 hours of recorded content, contributed by 20 speakers (10 male and 10 female). The audio files are in WAV format, sampled at a rate of 48 KHz, ensuring high-quality recordings for precise analysis. These files span across 7 distinct emotion classes, forming a rich and diverse foundation for our investigation.

- Emotional Classes: The dataset includes seven emotional classes: Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise. These emotions provide a diverse range for emotional speech analysis.
- Corpus Type: The speeches in SUBESCO are acted, allowing for controlled emotional expressions that facilitate research and analysis in emotional speech recognition.
- Speech Content: Each audio file consists of 10 sentences, containing 50 words. The speeches cover a variety of linguistic elements, including 7 vowels, 23 consonants, 6 diphthongs, and 1 nasalization, providing a rich phonetic diversity.
- Clip Duration: Each audio clip has a fixed duration of 4 seconds, ensuring consistency in the dataset and allowing for efficient processing in research applications.
- Total Duration: The cumulative duration of all audio files in the dataset is 7 hours, 40 minutes, and 40 seconds.
- Validator Information: During the dataset creation process, 50 validators were involved in each phase to ensure diverse perspectives in emotional evaluation.
- Subject Evaluation Recognition Rate: The recognition rate in subject evaluation falls within the range of 71-80%, indicating a substantial level of agreement among validators in identifying the intended emotions.

The SUBESCO dataset serves as a valuable resource for researchers and practitioners in the fields of speech processing, emotional speech analysis, and machine learning. Its carefully designed attributes and diverse emotional expressions make it an ideal tool for developing and testing algorithms related to emotion recognition in the context of Bangla speech.

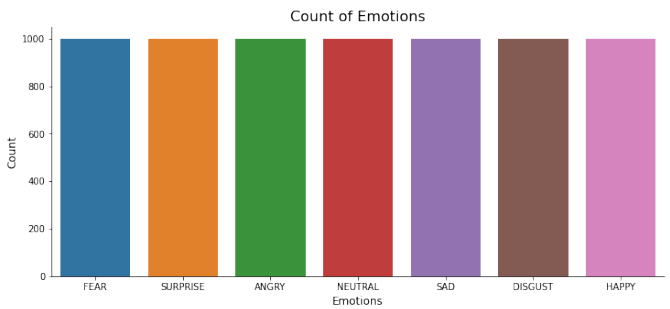

Fig. 1. Distribution of emotion classes.

B. Federated Learning Setup

To align with the principles of federated learning, we have strategically divided the SUBESCO dataset into 4 clients. Each client was treated as an independent entity, managing its unique subset of the data. This segmentation allowed for localized model training while preserving the integrity of the global model.

C. Data Preprocessing Techniques

1) Addition of Noises: To enhance the robustness of our model, we have introduced controlled levels of noise to the audio files. This deliberate introduction of noise aids in simulating real-world scenarios, ensuring the model's resilience in varied acoustic environments.

2) Time Stretching: The time-stretching technique was applied to modify the duration of audio sequences. This not only introduces variability in the temporal aspect of the data but also enables the model to adapt to different speech speeds, contributing to improved generalization.

3) Frequency Shift: We have incorporated frequency shift as a preprocessing step to account for potential variations in pitch across different speakers. This technique has assisted the model in learning invariant representations, promoting better performance across diverse vocal ranges.

4) Pitch Variation: Having introduced pitch variation further diversifies the dataset, exposing the model to a broader range of vocal pitch characteristics. This ensures the model's adaptability to varying emotional expressions and speaking styles.

5) MFCC Feature Extraction: For a compact and informative representation of the audio data, we have extracted Mel-frequency cepstral coefficients (MFCC). These coefficients have encapsulated the spectral features of the audio signal, serving as a foundation for tasks like speech and audio recognition.

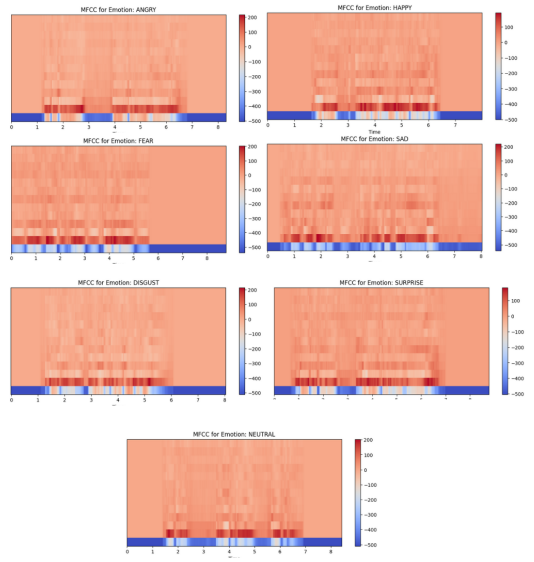

Fig. 2. MFCC plots for 7 emotion classes

In summary, our data preprocessing pipeline has involved a meticulous combination of noise injection, temporal manipulation, pitch adjustments, and the extraction of MFCC features. These steps have collectively contributed to a robust and diverse dataset, laying a solid foundation for the subsequent phases of our federated learning approach.

D. Workflow Diagram

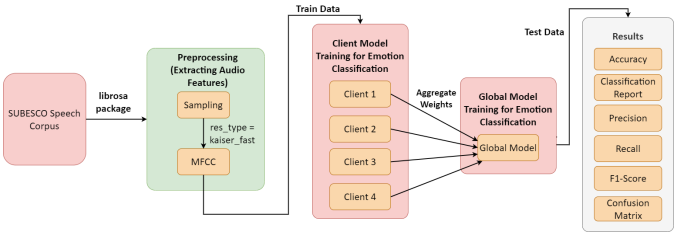

Fig. 3. Workflow

E. Model Architecture

Layer (type)	Output Shape	Param #
conv1d_3	(None, 20, 2048)	12288
max_pooling1d_3	(None, 10, 2048)	0
batch_normalization_3	(None, 10, 2048)	8192
conv1d_4	(None, 10, 1024)	10486784
max_pooling1d_4	(None, 5, 1024)	0
batch_normalization_4	(None, 5, 1024)	4096
conv1d_5	(None, 5, 512)	2621952
max_pooling1d_5	(None, 3, 512)	0
batch_normalization_5	(None, 3, 512)	2048
lstm_2	(None, 3, 256)	787456
lstm_3	(None, 128)	197120
dense_7	(None, 128)	16512
dropout_5	(None, 128)	0
dense_8	(None, 64)	8256
dropout_6	(None, 64)	0
dense_9	(None, 32)	2080
dropout_7	(None, 32)	0
dense_10	(None, 16)	528
dropout_8	(None, 16)	0
dense_11	(None, 8)	136
dropout_9	(None, 8)	0
dense_12	(None, 7)	63

Fig. 4. All layers of information about our model architecture.

V. Result Analysis

A. Local Model Performance

Before delving into the global model's accomplishments, let's scrutinize the individual contributions of each local client model. Following 300 epochs of training on their respective datasets, the local models exhibited varying levels of accuracy:

- Client 1: Achieved an accuracy of 86.37%
- Client 2: Demonstrated a performance of 69.15%
- Client 3: Attained an accuracy score of 81.42%
- Client 4: Showed a commendable accuracy of 76.45%

These divergent accuracies highlight the intricacies within each local dataset and the unique learning challenges faced by individual clients.

B. Aggregated Global Model

Moving to the global scale, the federated learning process seamlessly amalgamated the local models' insights, resulting in a refined global model. After an additional 150 epochs of training, the global model showcased a remarkable accuracy of 90.08%.

C. Interpreting the Results

1) Client Discrepancies: The variance in local model accuracies suggests disparities in the difficulty of emotion classification tasks across different clients. Understanding these differences is crucial for optimizing federated learning strategies tailored to each client's unique data characteristics.

2) Global Model Improvement: The substantial increase in accuracy from the aggregated global model (90.08%) compared to individual local models signifies the efficacy of federated learning. By leveraging the diverse knowledge embedded in local datasets, the global model demonstrates enhanced generalization and overall performance.

3) Robustness Through Aggregation: The federated learning approach proved effective in mitigating the impact of individual client limitations. Aggregating insights from multiple sources not only bolstered the global model's accuracy but also contributed to its robustness in handling diverse emotional speech patterns.

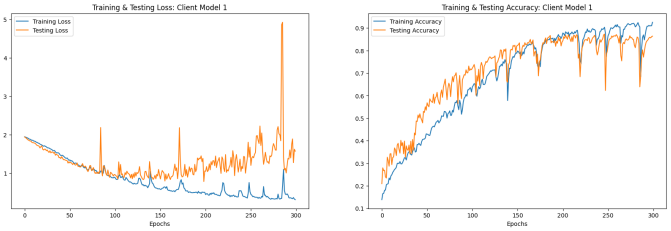

Fig. 5. History plot of client model 1

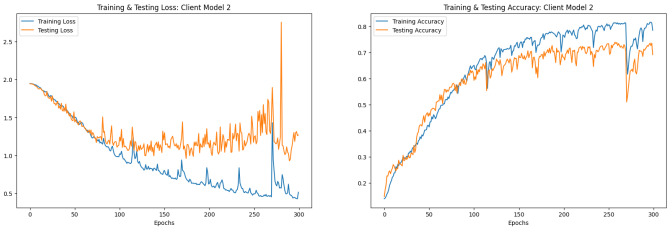

Fig. 6. History plot of client model 2

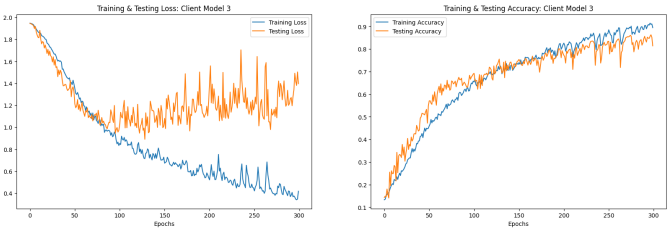

Fig. 7. History plot of client model 3

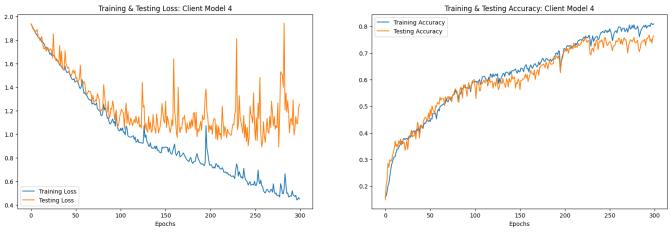

Fig. 8. History plot of client model 4

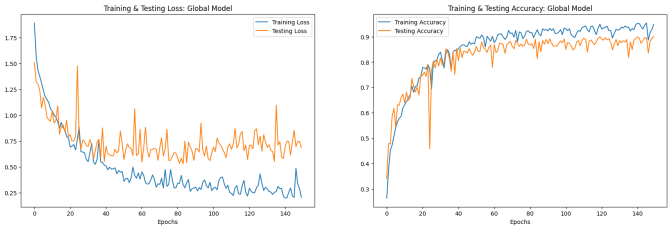

Fig. 9. History plot of global model

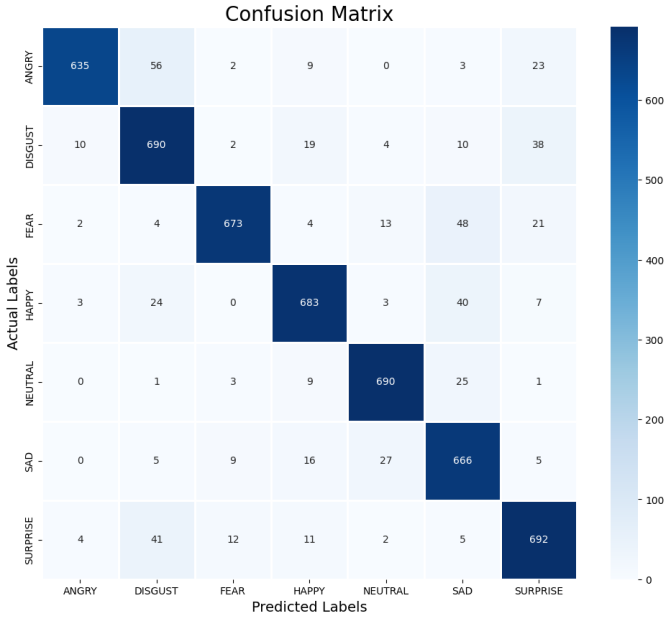

Fig. 10. Confusion matrix of the global model.

We have experimented with 6 alternative models. Upon comparing their performance with our global model, it was observed that the global model exhibited superior accuracy relative to the other models.

TABLE I
REPORT ON EMOTION RECOGNITION CLASSIFICATION

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
KNeighbors Classifier	67.06	67	69	68
Gradient Boost Classifier	64.93	65	65	65
MLP Classifier	63.26	81	71	75
SVM	59.45	59	60	59
Logistic Regression	46.70	45	47	46
Random Forest Classifier	40.78	95	41	56

VI. Challenges and Future Directions

In acknowledging the achievements of our federated learning model, it is imperative to conscientiously recognize the challenges that were encountered during its implementation. Foremost among these challenges were issues related to data heterogeneity, communication overhead, privacy considerations, and constrained computational resources on client devices. The resolution of these challenges forms a central focus for forthcoming research initiatives.

VII. Conclusion

In conclusion, the integration of culturally relevant emotion detection not only enriches human-computer interaction but also opens avenues for cross-cultural research and understanding. This inclusive approach resonates in applications, particularly in the realm of mental health, where emotion detection can significantly enhance assessments and interventions. Future research can focus on practical applications including a speech recognition system for the Bangla-speaking population opens up new horizons for emotion detection. Human-computer interaction largely depends on human emotion it has a huge potential. security, entertainment, healthcare, law enforcement, and customer service are different sectors that carry weight in which the SER model and FL can be used to advance society.

References

[1] S. Latif, S. Khalifa, R. Rana, and R. Jurdak, "Poster Abstract: Federated Learning for Speech Emotion Recognition Applications," 2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), Sydney, NSW, Australia, 2020, pp. 341-342, doi: 10.1109/IPSN48710.2020.00-16.

[2] P. Harár, R. Burget and M. K. Dutta, "Speech emotion recognition with deep learning," 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2017, pp. 137-140, doi: 10.1109/SPIN.2017.8049931.

[3] V. Tsouvalas, T. Ozcelebi and N. Meratnia, "Privacy-preserving Speech Emotion Recognition through Semi-Supervised Federated Learning," 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), Pisa, Italy, 2022, pp. 359-364, doi: 10.1109/PerComWorkshops53856.2022.9767445.

[4] Feng, T., Narayanan, S. (2022) Semi-FedSER: Semi-supervised Learning for Speech Emotion Recognition On Federated Learning using Multiview Pseudo-Labeling. Proc. Interspeech 2022, 5050-5054, doi: 10.21437/Interspeech.2022-141

[5] N. Liu et al., "Transfer Subspace Learning for Unsupervised Cross-Corpus Speech Emotion Recognition," in IEEE Access, vol. 9, pp. 95925-95937, 2021, doi: 10.1109/ACCESS.2021.3094355.

[6] L. Guo, L. Wang, J. Dang, Z. Liu and H. Guan, "Exploration of Complementary Features for Speech Emotion Recognition Based on Kernel Extreme Learning Machine," in IEEE Access, vol. 7, pp. 75798-75809, 2019, doi: 10.1109/ACCESS.2019.2921390.

[7] A. Woubie and T. Bäckström, "Federated Learning for Privacy-Preserving Speaker Recognition," in IEEE Access, vol. 9, pp. 149477-149485, 2021, doi: 10.1109/ACCESS.2021.3124029.

[8] T. Ngo et al., "Federated Deep Learning for the Diagnosis of Cerebellar Ataxia: Privacy Preservation and Auto-Crafted Feature Extractor," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 803-811, 2022, doi: 10.1109/TNSRE.2022.3161272.

[9] C. Tan, Y. Cao, S. Li and M. Yoshikawa, "General or Specific? Investigating Effective Privacy Protection in Federated Learning for Speech Emotion Recognition," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096844.

[10] X. Cui, S. Lu and B. Kingsbury, "Federated Acoustic Modeling for Automatic Speech Recognition," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6748-6752, doi: 10.1109/ICASSP39728.2021.9414305.

[11] S. Mohammadi et al., "Balancing Privacy and Accuracy in Federated Learning for Speech Emotion Recognition," 2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS), Warsaw, Poland, 2023, pp. 191-199, doi: 10.15439/2023F444.

[12] D. Guliani, F. Beaufays and G. Motta, "Training Speech Recognition Models with Federated Learning: A Quality/Cost Framework," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 3080-3084, doi: 10.1109/ICASSP39728.2021.9413397.

[13] D. Waref and M. Salem, "Split Federated Learning for Emotion Detection," 2022 4th Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 2022, pp. 112-115, doi: 10.1109/NILES56402.2022.9942417.

[14] P. Chhikara, P. Singh, R. Tekchandani, N. Kumar and M. Guizani, "Federated Learning Meets Human Emotions: A Decentralized Framework for Human–Computer Interaction for IoT Applications," in IEEE Internet of Things Journal, vol. 8, no. 8, pp. 6949-6962, 15 April15, 2021, doi: 10.1109/JIOT.2020.3037207.

[15] W. Guo, B. Jin, S. C. Sun, Y. Wu, W. Qi and J. Zhang, "Federated Learning of Wireless Network Experience Anomalies Using Consumer Sentiment," 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Bali, Indonesia, 2023, pp. 499-504, doi: 10.1109/ICAIIC57133.2023.10067061.

[16] R. Zeng, B. Mi and D. Huang, "A Federated Learning Framework Based on CSP Homomorphic Encryption," 2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS), Xiangtan, China, 2023, pp. 196-201, doi: 10.1109/DDCLS58216.2023.10167059.

[17] Sultana S, Rahman MS, Selim MR, Iqbal MZ (2021) SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. PLOS ONE 16(4): e0250173. https://doi.org/10.1371/journal.pone.0250173