WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Federate Learning for Privacy-Preserving Women's Mental Health Emotion Monitoring

[1] S.Selvasundari, [2] A.Shree Lakshmi, [3] Dr Jai Ruby

[1] MCA Student, [2] MCA Student, [3] Associate Professor
Department of Computer Applications and Research Centre, Sarah Tucker College
[1] sselvasundari558@gmail.com, [2] shreelakshmisl20@gmail.com, [3]rubysarah2025@gmail.com

ABSTRACT

This paper presents an AI-powered speech emotion recognition system designed to enhance mental health support for women. The proposed framework analyzes vocal cues such as tone, pitch, and rhythm to detect emotional states including stress, anxiety, and depression. A hybrid deep learning approach is employed, integrating Convolutional Neural Networks (CNNs) for spatial feature extraction from Mel-Frequency Cepstral Coefficients (MFCCs) with Long Short-Term Memory (LSTM) networks for capturing temporal speech dynamics. The model is trained using the RAVDESS dataset, which provides diverse emotional speech samples across multiple speakers. The system architecture comprises modular stages for voice input, preprocessing, feature extraction, classification, and real-time monitoring. Once deployed, the system enables real-time emotion detection and delivers AI-assisted recommendations through a user-friendly interface. By offering early insights into emotional well-being, this work aims to empower healthcare professionals with timely, personalized mental health interventions. The study contributes to proactive mental healthcare and demonstrates the potential of intelligent speech-based technologies in promoting women's emotional well-being.

Keywords: Speech Emotion Recognition, Women's Mental Health, Deep Learning, CNN-LSTM Hybrid Model, MFCC, Real-Time Monitoring.

1.INTRODUCTION

Mental health is an integral component of overall well-being, with emotional regulation playing a vital role in maintaining psychological balance. Women, in particular, are more vulnerable to mental health issues such as anxiety, depression, and chronic stress due to a combination of biological, hormonal, and sociocultural factors. Despite growing awareness, early detection and continuous monitoring of emotional health remain significant challenges, often constrained by the stigma associated with mental illness and the limitations of traditional assessment methods like self-reporting and clinical evaluations.

In recent years, advancements in artificial intelligence (AI) and speech signal processing have opened new avenues for non-invasive, real-time mental health monitoring. Among these, Speech Emotion Recognition (SER) has emerged as a promising approach for identifying emotional states by analyzing vocal features such as tone, pitch, and rhythm. Since human emotions are naturally reflected in speech, SER offers a passive and unobtrusive method to monitor mental well-being, especially useful in long-term, daily-life scenarios.

This research focuses on developing a speech-based emotion recognition system tailored for women's mental health monitoring. The proposed system utilizes a hybrid deep learning architecture that combines Convolutional Neural Networks (CNN) for spatial feature extraction from Mel-Frequency Cepstral Coefficients (MFCCs), and Long Short-Term Memory (LSTM) networks to capture the temporal dynamics of speech. The system is trained and validated using the RAVDESS dataset, which provides a wide range of emotional speech samples across different speakers.

By enabling the automatic detection of emotions through natural speech, this work aims to support early intervention and contribute to more accessible, personalized mental health care for women.

2.LITERATURE SURVEY

Speech Emotion Recognition (SER) plays a vital role in human-computer interaction, enabling systems to recognize and appropriately respond to users' emotional states. Recent advancements in deep learning have led to the development of more accurate, robust, and real-time SER systems. The following literature discusses various models and techniques employed in SER, ranging from hybrid deep learning architectures to gender-specific modeling and applications in mental health monitoring.

[1] Makhmudov proposed a hybrid deep learning architecture combining **Convolutional Neural Networks (CNN)**, **Long Short-Term Memory (LSTM)** networks, and **Attention Mechanisms** for enhanced SER. The CNN extracts spatial features, LSTM captures temporal dependencies in the speech signal, and attention layers help the model focus on emotionally salient parts of the input. Additionally, the authors incorporated psychological principles like

the "peak-end rule" to enhance emotion prediction. Evaluations on TESS and RAVDESS datasets yielded high accuracies (up to 99.8% on TESS), indicating the effectiveness of their model in recognizing emotions from speech.

[2] Ahmed developed an **ensemble model** combining 1D-CNN, LSTM, and GRU layers to improve SER performance. The architecture benefits from the strengths of each model: CNN for local feature extraction, LSTM for long-term dependencies, and GRU for efficient sequential modeling. Furthermore, data augmentation techniques—such as pitch shifting, time-stretching, and noise addition—were employed to enhance generalizability. The ensemble approach achieved state-of-the-art results on multiple datasets, including TESS, RAVDESS, CREMA-D, and SAVEE, demonstrating strong cross-dataset performance.

[3] Al-Hammadi introduced a novel approach using **gender-specific features** to improve SER accuracy. Their method first classifies the speaker's gender and then uses tailored feature extraction pipelines for male and female voices. The architecture integrates CNNs and Bidirectional LSTMs (Bi-LSTM) for sequential feature learning. The results showed a significant improvement in overall recognition accuracy and highlighted that considering gender-specific emotional expression can enhance model performance.

[4] Elsayedpresented a **1D-CNN + GRU** based SER model focused on applications in **mental health monitoring**. Their work aims to embed emotion recognition into intelligent virtual assistants (IVAs) for passive emotional state assessment.

[5] Goodfellow, Bengio, and Courville's textbook offers a comprehensive theoretical foundation for understanding deep learning models, including CNNs, RNNs, LSTMs, and GRUs—all widely used in SER. The text details the mathematical underpinnings, training strategies, regularization methods, and optimization techniques that have been applied in the above research works.

[6] Abbashian provided a foundational review of deep learning methods in speech emotion recognition (SER), comparing traditional machine learning approaches with deep architectures like CNNs, RNNs, and LSTMs, and highlighting their superiority in complex emotional contexts. Building on this, Peng et al. [7] introduced a multi-scale CNN combined with attention mechanisms, enabling the model to focus on emotionally salient parts of speech while capturing multi-resolution features. Muppidi and Radfar [8] proposed a more novel approach using Quaternion CNNs (QCNN), which allowed multidimensional speech features to be processed as a single unit, offering improved accuracy and efficiency over conventional CNNs.

[9] Tzirakis presented a semantic-aware SER system that fused audio features with pre-trained language models, improving classification of emotions that are acoustically similar but contextually different. Daneshfar and Kabudian [10] introduced a compact SER model using deep sparse autoencoders and Extreme Learning Machines (ELMs), which offered high accuracy with reduced computational cost, suitable for real-time systems. A review by *Expert Systems with Applications* [11] further emphasized the importance of multimodal fusion, outlining recent progress in combining audio, visual, and text modalities, while also identifying key challenges like synchronization and modality imbalance.

[12] Li explored deep multimodal SER frameworks using speech, facial, and textual inputs, highlighting various fusion strategies and the integration of transformer-based encoders such as BERT and ViT for richer emotional understanding. Atmaja and Sasou [13] demonstrated the effectiveness of universal speech representations (USRs) like wav2vec 2.0 in eliminating the need for handcrafted features, while still achieving high accuracy with limited training data. Barhoumi and BenAyed [14] designed a real-time SER pipeline using CNN-BiLSTM and data augmentation techniques to ensure robustness in noisy environments, making it ideal for applications like call centers and human-machine interaction. He [15] summarized emerging trends including self-supervised learning, attention-guided fusion, and emotion modeling for under-resourced languages, pointing toward the future direction of SER systems.

3. METHODOLOGY

The research proposes a deep learning-based speech emotion recognition system designed specifically for monitoring women's mental health. The methodology is divided into the following key modules, each playing a crucial role in achieving accurate and real-time emotion detection.

3.1 Dataset Collection and Preprocessing

This study employs the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), with emphasis on female speech samples to ensure the dataset's relevance for women's mental health monitoring. To enhance data quality, a series of preprocessing operations were applied, including noise reduction, silence trimming, normalization, and audio segmentation, thereby eliminating unwanted artifacts and improving speech clarity. All audio signals were resampled to a uniform rate of 16 kHz to maintain consistency across recordings. In order to increase dataset diversity and improve model robustness, data augmentation techniques such as pitch shifting and time stretching were employed, producing synthetic variations while preserving emotional integrity.

3.2 Feature Extraction

Feature extraction plays a crucial role in capturing emotion-related acoustic cues. Mel-Frequency Cepstral Coefficients (MFCCs) were selected as the primary features due to their proven ability to represent perceptually relevant aspects of human speech. To enrich the emotional context, additional features such as pitch, energy, and tone were extracted. The features were structured into spectrogram-like matrices, allowing simultaneous preservation of temporal and spectral information for effective learning.

3.3 Hybrid CNN-LSTM Model

A hybrid deep learning architecture combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks was designed for emotion recognition. The CNN component was employed to capture local spatial patterns from MFCC spectrograms, identifying fine-grained acoustic characteristics. These spatially encoded features were subsequently passed into LSTM layers, which modeledtemporal dependencies and captured the dynamic progression of emotional cues in speech. By integrating spatial and sequential representations, the hybrid CNN-LSTM model leveraged the strengths of both architectures, enabling robust classification of diverse emotional states.

3.4 Model Training and Validation

A stratified train-validation-test split was employed to ensure balanced representation of emotional classes. To prevent overfitting and improve generalization, cross-validation techniques and data augmentation were utilized. Model performance was assessed using multiple evaluation metrics, including accuracy, precision, recall, and F1-score, with confusion matrices providing further insight into class-specific misclassifications. These comprehensive evaluations facilitated identification of performance bottlenecks and guided further optimization of the architecture.

3.5 Real-Time Emotion Recognition System

To extend applicability beyond experimental conditions, a real-time system pipeline was developed. The pipeline integrates modules for live audio capture, preprocessing, feature extraction, and classification in a seamless manner, enabling efficient and responsive emotion recognition. A user-friendly graphical interface was designed to display detected emotions while also providing AI-driven mental health suggestions tailored specifically for women. Furthermore, the system was optimized for deployment on resource-constrained platforms, such as smartphones, ensuring accessibility and portability in real-world scenarios.

3.6 Mental Health Monitoring and Alerts

Beyond single-instance recognition, the system was designed to support continuous monitoring of emotional trends over extended periods. By analyzing patterns across time, the system provides users and healthcare professionals with a comprehensive view of emotional well-being. Based on predefined thresholds and patterns of detected emotions, timely alerts and recommendations for intervention can be generated. Recognizing the sensitive nature of mental health data, the system incorporates data privacy and security protocols, ensuring ethical handling of personal information while maintaining user trust.

4.EXPERIMENTAL ANALYSIS

To evaluate the proposed system, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was utilized, with a focus on female speech samples to ensure relevance for women's mental health applications. The dataset includes eight distinct emotional categories: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. All audio samples underwent preprocessing involving noise reduction, silence trimming, normalization, and MFCC feature extraction. The dataset was divided into training (70%), validation (15%), and testing (15%) subsets, ensuring stratified distribution of classes. A hybrid deep learning architecture comprising Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks was employed. CNN layers extracted local spatial features from MFCC representations, while LSTM layers modeled the temporal dependencies of speech signals. The model was trained using the Adam optimizer and categorical cross-entropy loss for 50 epochs with a batch size of 32. Regularization techniques including dropout and early stopping were applied to prevent overfitting. Data augmentation techniques such as pitch shifting and time stretching were incorporated to improve generalization.

Model performance was assessed using standard classification metrics—accuracy, precision, recall, and F1-score—to provide a comprehensive understanding of classification efficacy. Confusion matrices were generated to further analyze class-wise performance and misclassification trends.The CNN-LSTM model achieved an overall classification accuracy of 89.2% on the test set, demonstrating strong performance across most emotional categories. Table 4.1 presents the class-wise results.

Table 4.1. Class-wise performance of the CNN-LSTM model

Emotion	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Neutral	92.3	90.1	91.2
Happy	89.5	87.7	88.6
Sad	90.2	88.8	89.5
Angry	91.0	89.9	90.4
Fearful	88.4	86.5	87.4

Emotion	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Disgust	85.7	83.9	84.8
Surprised	87.9	85.6	86.7

Higher recognition rates were observed for *neutral*, *sad*, and *angry* classes, while slightly lower scores for *disgust* and *surprised* emotions were attributed to fewer available training samples for those categories. In comparison to traditional machine learning classifiers such as Support Vector Machines (SVM) and Random Forests, which were trained on the same MFCC feature set, the CNN-LSTM model demonstrated a significant improvement of approximately 10–15% in classification accuracy. This validates the efficacy of deep learning in capturing both spectral and temporal characteristics essential for accurate emotion recognition.

5. CONCLUSION

The proposed speech-based emotion recognition system for women's mental health monitoring achieved an overall test accuracy of 89.2% using a CNN-LSTM hybrid model trained on MFCC features, and demonstrated real-time performance with sub-second response times through integration with Google Speech-to-Text API and a rule-based AI Suggestions Module, enabling both acoustic and linguistic emotion analysis to support early and accessible emotional well-being interventions tailored for women.

6. REFRENCES

1. F. Makhmudov, A. Kutlimuratov, and Y.-I. Cho, "Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition," *Applied Sciences*, vol. 14, no. 23, Art. no. 11342, 2024.
2. M. R. Ahmed, S. Islam, A. K. M. M. Islam, and S. Shatabda, "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition," *Expert Systems with Applications*, vol. 223, p. 119987, 2023.
3. A. Al-Hammadi, M. A. Alsulaiman, M. E. Hossain, G. Muhammad, and M. A. Bencherif, "A Deep Learning Method Using Gender-Specific Features for Emotion Recognition," *Sensors*, vol. 23, no. 3, p. 1355, 2023.
4. N. Elsayed, M. Hossny, D. Smith, and S. Nahavandi, "Speech Emotion Recognition using Supervised Deep Recurrent System for Mental Health Monitoring," *arXiv preprint*, arXiv:2208.12812, 2022.
5. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, 1st ed., Cambridge, MA, USA: MIT Press, 2016.
6. B. J. Abbashian, D. Sierra-Sosa, A. Elmaghraby, "Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models," *Sensors*, vol. 21, no. 4, Art. no. 1249, 2021.
7. Z. Peng, Y. Lu, S. Pan, Y. Liu, "Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention," arXiv preprint, 2021.
8. A. Muppidi, M. Radfar, "Speech Emotion Recognition Using Quaternion Convolutional Neural Networks," arXiv preprint, 2021.
9. Panagiotis Tzirakis, Anh Nguyen, Stefanos Zafeiriou, Björn W. Schuller, "Speech Emotion Recognition using Semantic Information," arXiv preprint, 2021.
10. Fatemeh Daneshfar, Seyed Jahanshah Kabudian, "Speech Emotion Recognition Using Deep Sparse Auto-Encoder Extreme Learning Machine …", arXiv preprint, 2021.
11. "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Systems with Applications*, vol. 237, Part C, 2024, 121692.
12. Sunan Li, Yan Zhao, Chuangao Tang, Yuan Zong, "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face," *Entropy*, 2023, 25(10), 1440.
13. B. T. Atmaja and A. Sasou, "Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations," *Sensors*, vol. 22, no. 17, Art. no. 6369, 2022.
14. C. Barhoumi and Y. BenAyed, "Real-time speech emotion recognition using deep learning and data augmentation," *Artificial Intelligence Review*, vol. 58, article 49, 2025.
15. Y. He, "Research Advances in Speech Emotion Recognition based on Deep Learning," *Theoretical and Natural Science*, vol. 86, no. 1, pp. 45–52, 2025.