Privacy-First Mental Health Solutions: Federated Learning for Depression Detection in Marathi Speech and Text

Original Scientific Paper

Priti Parag Gaikwad*
Dr. D. Y. Patil Institute of Technology, Electronics and Telecommunication Engineering
COEP Technological University, Pune, Maharashtra
pritigaikwad071@gmail.com

Mithra Venkatesan
Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, Maharashtra
mithra.v@dypvp.edu.in

*Corresponding author

Abstract *– Federated Learning (FL) is a cutting-edge approach that allows machines to learn from data without compromising privacy, making it especially valuable in sensitive areas like mental health. This research focuses on using FL to detect depression through speech and text data from a Marathi-speaking population. Depression, a widespread mental health issue, often leaves subtle clues in the way people speak and write, making speech and text analysis a powerful tool for early identification. However, mental health data is highly personal, and protecting it is crucial. FL addresses this by enabling training across multiple devices without ever sharing the raw data. In this study, we introduce a federated learning framework designed specifically to detect depression in Marathi speakers. The framework combines Natural Language Processing (NLP) for analyzing text and audio processing techniques for studying speech patterns. Using a Marathi dataset that includes both speech and text samples from individuals with and without depression, we train local models on individual devices. These models are then combined into a global model, which is continuously improved through a process called federated averaging.*

Our findings show that this FL-based approach performs well in detecting depression while keeping the data private and secure. This highlights the potential of FL in mental health applications, especially for languages like Marathi, where gathering and processing data centrally can be difficult. By prioritizing privacy, this work opens the door for future research into using federated learning for other regional languages and mental health challenges. The Federated Learning model outperforms the non-FL model, achieving around 97.9% across accuracy, precision, recall, and F1-score, compared to 97.4% without FL. with speech dataset the model demonstrates high parameter values of above 96.0%., This demonstrates FL's effectiveness in improving performance on the text and speech based depression detection task.

Keywords: *Federated Learning, Depression Detection, Mental Health, Speech Analysis, Text Analysis, Marathi Dataset, Privacy-Preserving Machine Learning*

Received: March 14, 2025; Received in revised form: May 27, 2025; Accepted: July 13, 2025

1. INTRODUCTION

Depression is a mental illness from which approximately 264 million individuals suffer worldwide. The subjective self-evaluations and clinical interviews are essential parts of conventional diagnostic techniques, though they are not always reliable and accessible. Nowadays, the use of artificial intelligence in mental health screening, especially when using speech and text data, has opened up new path for early detection and intervention. The utilization of personal and behavioral data, however, raises significant privacy concerns, especially when it is collected and managed in centralized systems [1].

Marathi, a language that is widely spoken in western India, has over 83 million native speakers and is still underrepresented in speech-based mental health research, despite improvements in the diagnosis of depression for high-resource languages. Rich emotional and acoustic clues, including pitch change, pauses, and tone shifts, are present in Marathi speech data and can reveal psychological suffering. However, the centralized collection of such data is frequently impeded by sociocultural and privacy limitations. Privacy-sensitive speech analysis [2, 3] is crucial for assessing mental health issues, but using only privacy-preserving features can lead to information loss and data-sharing risks. Decentralized training meth-

ods like Federated Learning (FL) have been explored to reduce privacy concerns, but their accuracy and overhead can be concerns, especially when deployed on mobile devices. No work has been conducted to establish the accuracy and performance overhead of FL for privacy-preserving speech analysis. In this work, we provide a federated learning system that uses convolutional architectures that have been locally trained on client devices for Marathi speech-based depression diagnosis. This method improves both ethical norms and model robustness by maintaining language diversity and auditory nuance while guaranteeing that the user's data never leaves their device. The data privacy gap is closed through the use of Federated Learning. We trained a model using Federated Learning (FL) that maintains user privacy without passing user-specific raw data to the central server.

Similarly, textual data also provides helpful linguistic indicators of depression, including negative sentiment, self-relatable phrases, and a reduction in syntactic complexity [4]. In this study, Marathi, a low-resource language, has been selected. We trained localized text-based models on dispersed clients using the federated learning paradigm, which enables each instance to contribute to a global, privacy-preserving model while adapting to region-specific language patterns. Considering the multilingual and dialectally diverse user base of Indians, this strategy effectively achieves a balance between generalization and personalization.

This study discusses the importance of Federated Learning (FL) in addressing privacy concerns and improving model robustness by training on distributed data without compromising individual privacy. The proposed framework represents a promising advancement in depression detection, offering a robust and privacy-preserving solution that could significantly impact clinical practice and research in the field of mental health [5]. Ensuring decentralized model training without compromising client data privacy. This is the novel combination of deep learning, attention mechanisms, and privacy-preserving. FL offers a robust and scalable solution for accurate depression detection across low-resource datasets. Here the users' data privacy concern is addressed because their private data is never posted and cannot be accessed by the server. Comparing the effectiveness of a model built using centralized and distributed machine learning techniques is another goal of the research [6].

1.1. OBJECTIVES

- To protect user privacy by implementing Federated Learning (FL), ensuring that sensitive personal data remains on users' devices and is not shared with centralized databases.
- To strengthen security against cyber threats by decentralizing data storage and reducing the risk of breaches and unauthorized access.
- To train AI models without sharing raw data by enabling decentralized learning, allowing models to learn from multiple sources while preserving data confidentiality.
- To maintain high model accuracy across different data sources by ensuring effective and reliable learning even in a decentralized and diverse data environment.
- To make data more usable without compromising privacy by allowing organizations to collaborate on AI model improvements without accessing private user information.
- To encourage safe and secure participation by designing a system where users can contribute to model training while ensuring their data remains protected

1.2. CONTRIBUTION

- This study incorporated Federated Learning (FL) to both speech and text data in Marathi language, which has remained unexplored in existing research work. Hence, this addresses a significant gap in the current literature. FL has been used in the past for privacy-preserving depression detection, mostly in high-resource languages such as English.
- This study employed a novel hybrid DepreLex-BERT model to identify depressive language patterns in Marathi text. Additionally, to detect depression in Marathi speech, this study combines a hybrid feature selection algorithm with an attention-driven CNN architecture.
- This Federated Learning model yields impressive results above 96% accuracy, precision, recall, and F1-score. This illustrates FL's powerful capacity to significantly enhance performance on tasks requiring the identification of depression in both text and speech.

The remainder of the article is structured as follows: Section 2 reviews the related work based on existing research. Section 3 outlines the proposed methodology. Section 4 represents the system implementation. Section 5 presents the results and provides a comparative analysis, and Section 6 concludes the article.

2. RELATED WORKS

The literature highlights gaps and shortcomings in centralized databases, including privacy and security concerns, limited datasets, limited training performance, vulnerability to attacks, advanced encryption, communication issues, and integration of multi-view data with federated learning in existing studies. To address these challenges, Federated Learning (FL) is a distributed machine learning approach designed to handle the decentralization of privacy-sensitive personal data. It is a networked machine learning framework that maintains client confidentiality by training on data from multiple clients without exposing it. The process starts with the creation of an initial global model, which

is then shared with each client, followed by updates using local data. This method minimizes the risk of data breaches by preventing access to raw data.

Y. Cui *et al*. [7] have developed a new method to diagnose depression using speech data while maintaining user privacy by employing federated learning (FL), which ensures that sensitive voice recordings remain on users' devices, with only encrypted model updates being shared with a central server. To address challenges like variations in individual speech patterns and the need for efficient communication, they implement secure aggregation to enhance privacy. This method offering stronger privacy protections. However, the study acknowledges challenges in scaling the system for real-world use and highlights the potential benefits of integrating speech analysis with other data types.

B. S. Reddy *et al*. [8] have proposed a multimodal approach to detecting depression by combining both speech and text data. The system analyzes acoustic features such as pitch, tone, and speech rate alongside linguistic aspects like sentiment, word choice, and syntactic patterns, which may indicate depressive symptoms. By integrating these two modalities, the goal is to improve the accuracy and reliability of depression detection. Some challenges in the proposed work are complexity of data preprocessing, the need for large, high-quality datasets, and ensuring that the system remains effective across diverse populations.

J. Li *et al*. [09] have proposed a privacy-focused method for detecting depression through speech analysis using asynchronous federated optimization. This approach improves traditional federated learning by allowing devices to update the model at different times, reducing. The study shows that this method enhances training efficiency and maintains diagnostic accuracy comparable to existing techniques.

L. Zhang *et al*. [10] have developed a model to detect adolescent depression by analyzing both speech and text from interviews. By combining vocal and linguistic features, the system improves accuracy in identifying subtle depression indicators. Machine learning helps recognize patterns, but challenges like data variability and emotional cue interpretation remain. This study highlights the potential of multimodal methods for reliable mental health assessments.

A. B. Vasconcelos *et al*. [11] have explored using FL to detect depression in social media posts while preserving user privacy. By combining FL with Natural Language Processing (NLP), the approach analyzes linguistic markers of depression—such as negative sentiment and writing style changes—without centralizing sensitive data. FL enables collaborative model training across distributed data, ensuring strong detection performance while maintaining privacy.

A. Kim *et al*. [12] have presented an automatic system for detecting depression using speech signals from smartphones and deep Convolutional Neural Networks (CNNs). By analyzing speech patterns such as rate, tone, and intonation, the model identifies acoustic markers linked to depression. Trained on a large dataset, the deep CNN achieves high accuracy in recognizing depressive states. This research highlights the potential of leveraging everyday smartphone data for mental health monitoring, offering a convenient and low-effort approach to mobile health applications.

L. Liu *et al*. [13] have assessed the accuracy of deep learning models in detecting depression through speech samples. By analyzing studies that use deep neural networks, the authors evaluate model performance, highlighting both strengths and limitations. The findings show that models focusing on acoustic features like pitch, intensity, and speech rate achieve high diagnostic accuracy.

J. Ye *et al*. [14] have presented a multimodal approach to detecting depression by combining emotional audio with evaluative text. By analyzing speech for emotional cues like tone and rhythm alongside text-based sentiment and word usage, the system improves accuracy in identifying depressive indicators. Machine learning techniques help uncover patterns that might be missed when using a single data source. Results show that this combined method outperforms single-modality models, emphasizing the advantage of integrating multiple data types for more reliable depression detection.

L. He and C. Cao [15] have explored the use of Convolutional Neural Networks (CNNs) to detect depression through speech analysis. By examining acoustic features like pitch, energy, and speech rate, the CNN model is trained to identify signs of depression with high accuracy. The findings highlight the potential of advanced speech analysis for real-time mental health monitoring and assessment.

D. Low *et al*. [16], have explored how speech analysis can assist in the automated assessment of psychiatric disorders, with a focus on depression. It examines various studies that use speech processing techniques to identify mental health conditions by analyzing vocal traits like prosody, speech rate, and tone. The review highlights the benefits of speech-based assessments, including their non-invasive nature and ease of use, while also addressing challenges such as variations in speech patterns, background noise, and the need for diverse datasets. The authors discuss potential clinical applications and future research directions for speech-based diagnostic tools.

W. Wu *et al*. [17] have introduced a self-supervised learning approach for detecting depression through speech analysis. The method uses unsupervised learning to extract meaningful patterns from large amounts of unlabeled speech data, which can then be fine-tuned for identifying depression. By leveraging self-supervised techniques, the model overcomes the challenge of limited labeled data, a common issue in mental health research. Findings show that this ap-

proach improves detection performance compared to traditional supervised methods, making it a promising tool for mental health assessment.

X. Xu *et al*. [18] have introduced "FedMood," a federated learning framework designed to detect mood disorders like depression using mobile health data. The system analyzes information from activity logs, speech patterns, and physiological signals while maintaining user privacy by keeping data on local devices. Researchers address challenges such as inconsistent data distribution and propose optimization techniques to enhance model performance. The findings highlight FedMood's potential for scalable, privacy-focused mental health monitoring, though the study notes limitations related to dataset diversity and real-world application.

E. L. Campbell *et al*. [19] have demonstrated how multimodal inputs are complementary; let's look at the diagnosis of Major Depressive Disorder (MDD) utilizing both textual and speech modalities. They illustrate the necessity of sophisticated fusion techniques and show how language and auditory information enhance detection capabilities. However, the cross-lingual environment and lack of privacy issues restrict the applicability of their findings in multilingual, real-world situations.

L. Yang *et al*. [20] have introduced feature-augmenting networks to improve the assessment of depression severity through speech analysis. By enhancing acoustic features like prosody and energy with deep neural network representations. The research highlights the benefits of combining traditional speech features with deep learning for more precise detection.

M. Ahmed *et al*. [21] have explored the use of on-device FL used to detect depression by analyzing text from Reddit posts. To protect user privacy, the approach keeps data on individual smartphones while allowing models to be trained collaboratively, challenges remain, including scalability issues and the need for a substantial amount of labeled training data.

Q. Deng *et al*. [22] have introduced a framework for detecting depression by analyzing both speech signals and textual transcripts using a speech-level transformer model with hierarchical attention mechanisms. By leveraging hierarchical attention, the model highlights crucial cues in both modalities, offering insights into how these features contribute to classification. The research underscores the potential of advanced AI for mental health applications while acknowledging challenges like real-world data variability and the need for effective multimodal alignment. W. Wei *et al*. [23] have introduced a federated contrastive learning method (FedCPC) that can also be applied to identifying depression. The approach uses contrastive pre-training to develop meaningful speech representations from decentralized data while maintaining user privacy. By allowing collaborative model training without sharing raw data, FedCPC enhances both accuracy and security. The study emphasizes the potential of combining federated learning with contrastive techniques to create scalable and privacy-focused healthcare solutions.

2.1. GAPS IDENTIFIED

Following are the gaps identified from the analysis.

- Limited integrated frameworks that assess the ethical effects of employing decentralized learning in mental health.
- Decentralized training methods like Federated Learning (FL) have been explored to reduce privacy concerns, but their accuracy and overhead have been concerns.
- Limited work has been conducted to establish the accuracy and performance overhead of decentralized learning for privacy-preserving speech analysis.
- Another problem is the diversity in the language of the content. People love to talk and express their feelings in their mother tongue. Most of the research in natural language processing (NLP) deals with the English language

3. PROPOSED METHODOLOGY

In mental health studies, large amounts of EHR (electronic health records) may expose owners' sensitive information, such as disease history, medical records, and personal details. In our work, we have proposed a model in which federated learning (FL) will be applied to e-health records of mental health patients. Instead of sending their original data to the server, data owners in the FL train the model locally and provide only the generated gradients. In this study, a practical and privacy-preserving FL architecture is proposed that protects the privacy of all EHR owners and is robust to their training dropout. A novel framework of acoustic, linguistic, and channel attention is proposed, which is shown in Fig. 1.

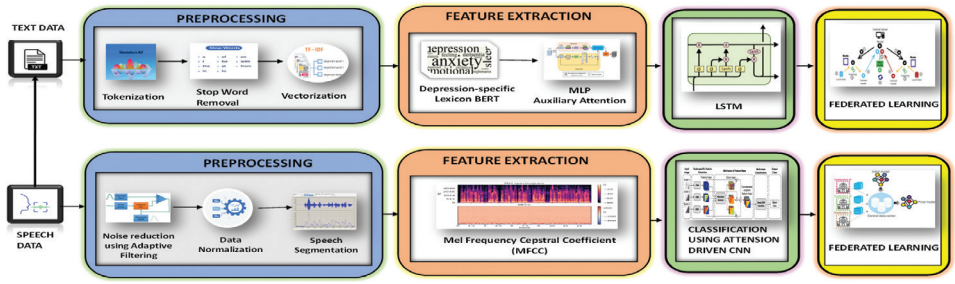


Fig.1. Proposed Methodology

The proposed work focuses on creating a Marathi speech and text dataset. The database contains recordings from various speakers, each differing in gender and age. Ultimately, it is the sentences, rather than individual words, that often convey the emotions present in speech. This speech data is then transformed into a text dataset. Consequently, a key challenge in current research is to enhance emotion recognition accuracy by utilizing Low-Level Descriptors (LLDs) alongside sentence-level features. Traditional methods for speech emotion recognition typically involve three main approaches. Data preprocessing for the speech dataset includes noise reduction, data normalization, and speech segmentation, while the text dataset undergoes tokenization, stop word removal, and vectorization. These processes are essential for analysing the diverse speakers and the emotions they express in their speech. Initially, noise reduction is a crucial step in speech data preprocessing to enhance the clarity of audio signals by minimizing unwanted background noise. In the text procession, the unwanted text is removed, and each word is important for depression-specific words. Text processing comes with several challenges, such as handling noisy data, managing different text formats, resolving ambiguity, and working with large vocabularies to improve language understanding. To tackle these issues, text preprocessing is a vital step in NLP, as it cleans and structures raw text so that machine learning models can analyse it effectively. One common technique used in this process is TF-IDF, which measures word importance by considering both how often a term appears and its relevance across documents. This helps filter out stop words, allowing models to focus on meaningful content. Additionally, FastText-based vectorization represents words as character n-grams, mapping them into high-dimensional vectors. This approach captures sub-word information, making it useful for understanding semantic similarities. Before being processed by an artificial neural network, each word must be converted into a single vector to ensure consistency and enhance model accuracy. Following the initial pre-processing stage, many feature sets were extracted for every audio recording and text dataset together with their corresponding statistical measurements. The spectral characteristics associated with the spectral centroid, MFCCs, which are cepstral characteristics linked to cepstrum analysis, and prosaic components. The initial harmonious signal, the abecedarian frequency, and these features represent the unique features of voice that represent the depression. In the text database, the depression-specific lexicons are extracted from Depression-Specific Lexicons-BERT (DepreLex-BERT), which combines depression-specific lexicons, N-grams, and BERT algorithms. Extensive research has explored word embedding techniques using statistical methods to derive meaningful word representations from text corpora. More recently, deep learning algorithms based on transformers, such as BERT, have been widely adopted for word embedding. This approach significantly enhances the performance of various natural language processing tasks by preserving nuanced meanings and improving contextual understanding. The next block is classification to support the diagnosis of depression; it is therefore necessary to develop depression classification methods. Acoustic feature extraction utilizes a 1D CNN to capture temporal patterns in sound signals, while linguistic feature extraction relies on an LSTM layer to process sequential language data. To improve performance, an attention mechanism is integrated to highlight the most relevant parts of the input, along with a channel attention mechanism that identifies interactions across CNN channels. Finally, the model's performance is evaluated through parameter validation, where metrics such as balanced accuracy, precision, recall, and F1 score are computed using the confusion matrix. The literature identifies several limitations of centralized databases, including privacy and security risks. To address these issues, Federated Learning (FL) has emerged as a distributed machine learning approach designed to handle the decentralization of privacy-sensitive personal data. FL operates within a networked architecture that safeguards client confidentiality by enabling model training across multiple clients without exposing their raw data. A networked machine learning system called federated learning (FL) enables training on data from various clients without jeopardizing client confidentiality. FL's key component is preserving privacy because data collection is unnecessary. The recurrence of the following processes usually results in FL progressing. An initial global model is created and distributed to each client via a centralized server. With local data, each client trains a copy of the model. The model's local modifications are transmitted from each client to the server. The server collects the updates from the clients, updates the global model, and then sends the revised global model back to the clients. These processes are repeated until the model converges or a stopping requirement is satisfied. This reduces the attack surface because training is carried out solely through local model updates without access to raw data.

3.1. FEDERATED LEARNING

We are unable to use patient health data kept in hospitals for centralized learning due to privacy concerns. Federated learning is a new framework that Google has suggested. Each hospital involved in the model training cooperation may keep its own data locally during the federated learning model's training process, eliminating the need for uploading. Every hospital downloads the model from the server for training using its own data. It then uploads the trained model or gradient to the server for aggregation. Finally, the server notifies each hospital of the aggregated model or gradient information. We use the model average approach for training, taking into account the load on communication, connection dependability, and other factors. When the global model parameters are updated in the t round, assuming that K hospitals are involved in federated learning, the K-th participant uses n (1) to cal-

culate the local data average gradient of the current model parameters. The server then aggregates these gradients and uses the updated model parameters to update the global model using equation (2).

$$g_k = \nabla F_k(\omega_t) \quad (1)$$

$$\omega_{t+1} \leftarrow \omega_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} g_k \quad (2)$$

where g_k is the local data's average gradient for the current model parameter, ω_t. The learning rate is denoted by η, where $\sum_{k=1}^{K} \frac{n_k}{n} g_k = \nabla f(\omega_t)$.

Every hospital performs one or more gradient descent steps using local data, updates the model parameters locally in accordance with equation (3), and transmits the modified local model parameters to the server. The server then provides the aggregated model parameters to each hospital after computing the weighted average of the model results using equation (4). The enhanced system may select an alternate gradient update optimizer in addition to SGD and can minimise communication requirements by 10-100 times when compared to the simply distributed SGD.

$$\forall k, \omega_{t+1}^{(k)} \leftarrow \overline{\omega_t} - \eta g_k \,, \quad (3)$$

$$\overline{\omega}_{t+1} \leftarrow \sum_{k=1}^{K} \omega_{t+1}^{(k)}, \quad (4)$$

where $\bar{\omega}_t$ is the existing model parameter of the local client.

The literature identifies several gaps and challenges associated with centralized databases, such as concerns around privacy and security, limitations in dataset availability, restricted training performance, vulnerability to attacks, communication inefficiencies, and issues with integrating multi-view data. Additionally, advanced encryption techniques and privacy safeguards often struggle to fully address these challenges. To tackle these limitations, Federated Learning (FL) has emerged as a promising distributed machine learning paradigm. FL enables decentralized processing of privacy-sensitive data by training models directly on clients' devices without transferring raw data to a central server. In this approach, an initial global model is created and distributed to multiple clients. Each client then updates the model using its local data, after which these updates are aggregated to improve the global model. Local and global training are the primary components of the asynchronous federated learning framework. Every device updates its local model during local training using the global model that the server sends. When any device uploads parameters, the server instantly updates the global model. In contrast to the federated average algorithm, noise is introduced before global model updating to prevent parameter leakage. Take federated learning with clients (devices) into consideration. is the local data on the ith device in a horizontal federated learning system. A certain instance is from the ith device. The ultimate goal is to use the distributed local models from every device to train a global model. To achieve this, perceive Eq. (5) as the optimization's ultimate objective.

$$\min F(w) = min\left(\frac{1}{n}\sum_{i\in[n]}^{n} E_{Z^i \sim D^i}(W;\, Z^i)\right) \quad (5)$$

The basic idea behind federated optimization is that there are global epochs, and worker i sends a local model with new information to the server using eqs. (6) and (7).

$$G_t = W_{back}^i - W_{new}^i \quad (6)$$

$$W_{t+1}^k = W_t^k - \partial_k g_t^k \quad (7)$$

Gradients cannot be uploaded straight to the server because the majority of devices are edge devices. Extract the file after uploading the device's most recent model to the server. To enable the server to determine the gradient that the ith device updated, save each model that the ith device uploads to the server in. Is the model on the other device prior to the new one being updated? To prevent the negative effects of the direct upload gradient on the server, use Eq. (6) to indirectly calculate the gradient of the device due to the unreliable connection and poor communication efficiency of edge devices. Following the server's receipt of the local model, equation (7) shows the updated global model. This decentralized process enhances data security by minimizing exposure to potential breaches and attacks, as raw data remains on individual devices throughout the learning process.

Algorithm1:
Federated Learning for Speech and Text Data

Input: Global model parameters W_o, number of global communication rounds T, learning rate η, number of devices n, local dataset D^i on each device i.

Output: Optimized global model W_T

Initialize Global Model:

$Wo \leftarrow$ Random Initialization

For each global round $t \in [1,T]$:

Send current global model W_t to all participating devices {1,2,…,n}.

For each device i in parallel

Download global model W_t.

Perform local training using the speech/text data D^i for E epochs:

$$W_{t+1}^i \leftarrow W_t - \eta \nabla F_i(W_t; D^i)$$

Calculate the local model update:

$$G_t^i = W_t^i - W_{t+1}^i$$

Apply differential privacy (e.g., Gaussian noise) if necessary:

$$\tilde{G}_t^i = G_t^i + noise(\mu, \sigma)$$

Send $\tilde{G}_t^i$ to the server.

Server Aggregation:

Update the global model using Federated Averaging (FedAvg):

$$W_{t+1} = W_t - \eta \sum_{i=1}^{n} \frac{|D^i|}{\sum_{j=1}^{n} |D^i|} \tilde{G}_t^i$$

Repeat Steps 2a to 2c until convergence or maximum rounds T are reached.

Return Optimized Global Model: W_T

4. RESULTS AND DISCUSSION

The following section provides a comparison of baseline models, along with detailed descriptions of the evaluation metrics, outcomes, and datasets. Each subsection delves into these components to offer a comprehensive understanding of the results. The proposed deep learning-based method is evaluated against a benchmark filter using simulations, data preprocessing, and Python 3 libraries, including NumPy, pandas, seaborn, and Scikit-learn. The model is developed with TensorFlow 2.10 and the Keras library.

The dataset's training and testing versions were divided, with 80% going toward training and 20% toward testing. Considering their suitability for classifying the depression for 50 epochs, the technique was trained using the AdamW optimizer with verbose 2, batch size 16, a learning rate of 0.001, and a cross-entropy loss function.

4.1. RESULT OBTAINED FROM THE PROPOSED METHOD

4.1.1. Dataset Description and Transcription

The dataset was collected from 54 non-professional Marathi speakers (both genders) using a microphone at a consistent distance, recorded in a clinic in Pune, Maharashtra, under the supervision of a psychologist. The Beck Depression Inventory (BDI), a widely used self-report tool with 21 items assessing depression symptoms and severity, was used to validate the dataset. The BDI assesses emotional, cognitive, and physical symptoms of depression, scoring each item from 0 to 4 to determine overall depression severity for individuals aged 13 and older. More severe depressed symptoms are indicated by higher overall scores. [24]

The speech Marathi dataset is transcribed into the text Marathi dataset using manual transcription to maintain the 99% accuracy of the dataset. The Marathi speech audio files of 54 individuals are transcribed into the Marathi text dataset and used while implementing the algorithm. All the datasets are labelled according to the BDI tool score and stored in the different folders according to their labelled class for normal, mood disturbance, slightly depressed, moderately depressed, and depressed in different class folders 0, 1, 2, 3, and 4, respectively.

This section presents a comparison of the proposed methodology, which is evaluated both with and without the federated learning approach. The aim is to determine which method performs better based on various parameters. We analyse the effectiveness of the proposed method against current techniques, using metrics such as Accuracy, Precision, Recall, and F1-score. The figure illustrates a comprehensive comparison of the proposed procedures in the context of Centralized and Decentralized Learning.

4.2 PERFORMANCE EVALUATION RESULTS

The section presents the performance evaluation of the model's training and validation results, both with decentralized (Federated Learning) and Centralized (without Federated Learning), using a confusion matrix and ROC curve for assessment.

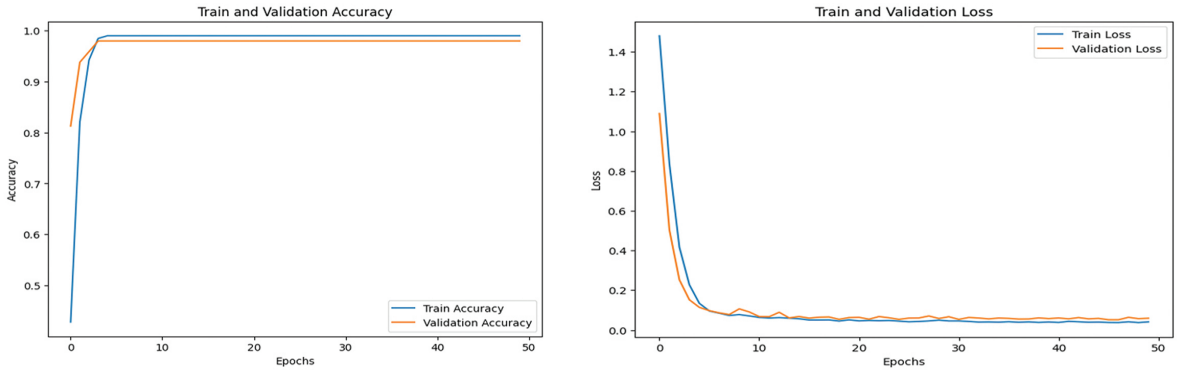


Fig. 2. Accuracy- Loss for training and validation with decentralized Learning for text Moel

Fig. 2 illustrates the training and validation accuracy over 50 epochs for the depression detection text model with federated learning applied. The training accuracy increases quickly, converging at around 99% by the 10th epoch, while the validation accuracy stabilizes close to 98.5%, reflecting strong learning performance with minimal overfitting. Simultaneously, the loss plot shows a sharp decrease in training loss, reaching approximately 0.02 by epoch 10, with the validation loss levelling off around 0.05. Both accuracy and loss exhibit early improvements, indicating rapid convergence and effective optimization using federated learning.

Fig. 3 illustrates the training and validation accuracy over 50 epochs for the depression detection. Speech model with decentralized learning: the training accuracy rises sharply to nearly 98% within the first 5 epochs and reaches 100% after 10 epochs. However, the validation accuracy levels off at approximately 80%, revealing a significant gap between the training and validation performance, suggesting that the model may be overfitting to the training data. The training loss decreases significantly, dropping to around 0.1 by epoch 10, whereas the validation loss exhibits a different behavior, increasing after initially reaching its lowest point of 0.8. This divergence between the training and validation losses further supports the indication of overfitting, where the model performs well on data in the training set but struggles to generalize to unseen data.

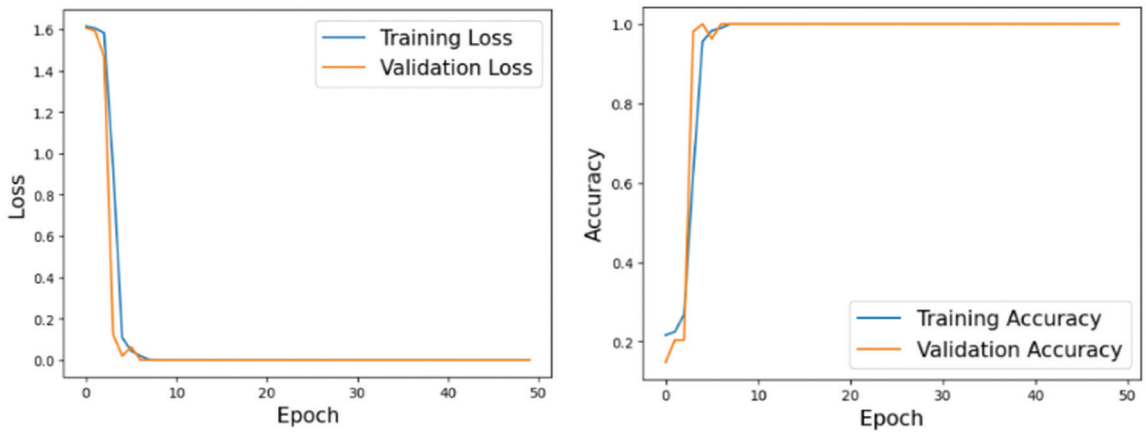
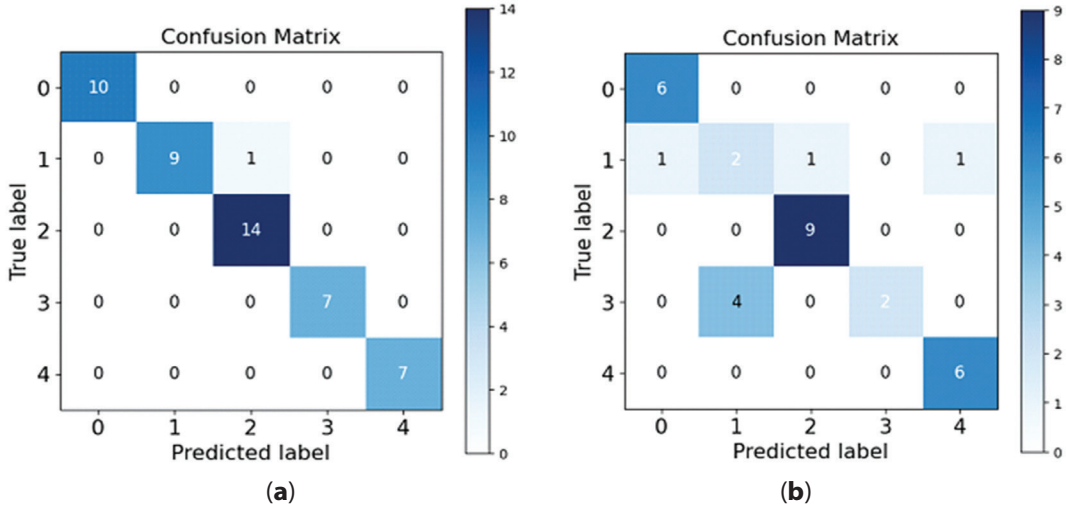


Fig. 3. Accuracy- Loss for training and validation with decentralized learning for Speech Model

Fig. 4 illustrates the confusion matrix for the text model with centralized and decentralized learning and demonstrates strong performance across all classes, with Class 0 correctly predicted 10 times and no misclassifications, Class 1 accurately classified 9 times with only 1 misclassification into Class 2, Class 2 having 14 correct predictions and no errors, and Classes 3 and 4 both showing 7 correct classifications with no misclassifications. This absence of significant errors indicates high accuracy and consistency, suggesting the model generalizes well to unseen data. In contrast, the confusion matrix for the model without federated learning reveals more misclassifications. Class 0 is correctly predicted 6 times but has 1 misclassification into Class 1, while Class 1 has 2 correct predictions with misclassifications into Classes 0, 2, and 4. Class 2 maintains good accuracy with 9 correct predictions, but Class 3 has increased errors, with only 2 correct predictions and 4 misclassifications into Class 4. Similarly, Class 4 is correctly predicted 6 times.

The FL model's test evaluation combines predictions from several customers, each of whom contributes a distinct number of samples per class. In contrast, the centralized (non-FL) model is evaluated using a unified, potentially more balanced dataset.

The higher number of errors, especially in Classes 0, 1, and 3, indicates that the model without Federated Learning struggles more with generalization, leading to more prediction inaccuracies compared to the Federated Learning approach.

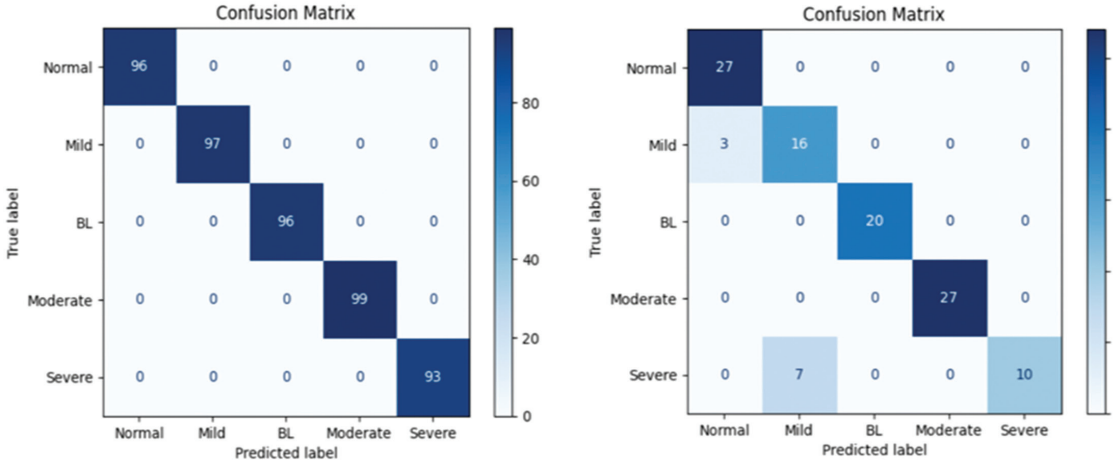


(**a**) (**b**)

Fig. 4. Confusion matrix for the text model with Centralized and Decentralized Learning. (**a**) with FL, (**b**) without FL

The higher number of errors, especially in Classes 0, 1, and 3, indicates that the model without Federated Learning struggles more with generalization, leading to more prediction inaccuracies compared to the Federated Learning approach.

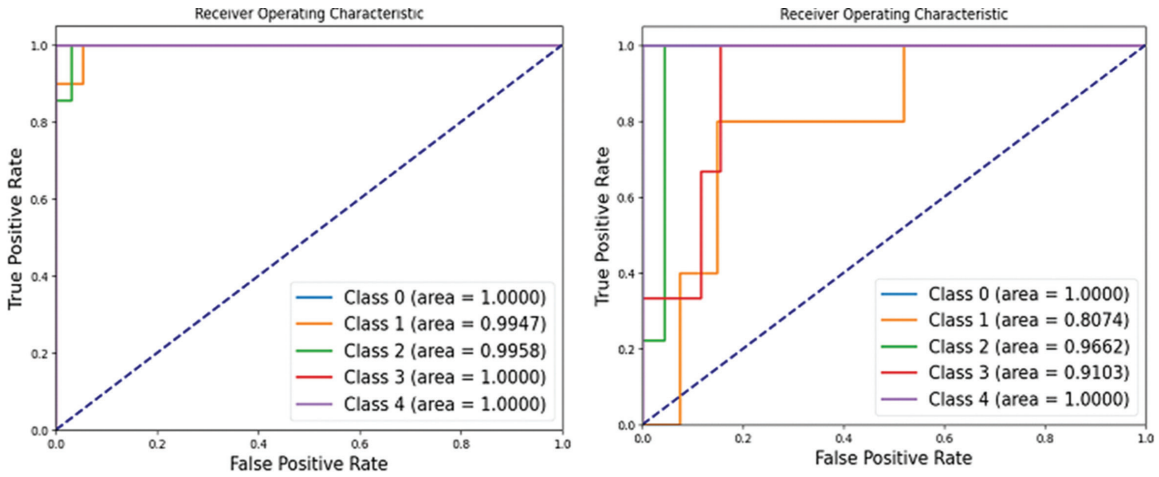
Fig.5 compares the decentralized approach and the centralized approach. shows the confusion The decentralized matrix shows the performance of a classification model for speech across five classes: normal, mild, BL, moderate, and severe. The matrix indicates perfect classification, with all instances correctly predicted for each class (e.g., 96 Normal, 97 Mild, 96 BL, 99 Moderate, and 93 Severe). There are no misclassifications, suggesting 100% accuracy. While centralized, the confusion matrix evaluates a classification model's performance across five classes: Normal, Mild, BL, Moderate, and Severe. The model correctly predicted 27 normal, 16 mild, 20 BL, 27 moderate, and 10 severe instances. However, there are misclassifications: 3 mild instances were incorrectly predicted as normal, and 7 severe instances were misclassified as mild. While the model performs well for normal, BL, and moderate classes, the errors in mild and severe classes indicate room for improvement, particularly in distinguishing between mild and severe cases. Instead of being centrally pooled, data in federated learning is dispersed among several clients, each of which represents a distinct local dataset. The total class distribution in the test phase may therefore be different from that in the centralized (non-FL) assessment, where the dataset is uniformly available, when combining predictions from all clients in the FL configuration. These differences in class-wise sample counts reflect the variability introduced by FL's decentralized nature and demonstrate the model's performance under realistic, heterogeneous conditions.



(**a**) (**b**)

Fig. 5. Confusion matrix for the speech model with Centralized and Decentralized Learning (**a**) with FL, (**b**) without FL

Fig. 6 illustrates the ROC curve for the text model with centralized and decentralized learning.



(**a**) (**b**)

Fig. 6. ROC curve for the model with Centralized and Decentralized Learning. (**a**) with FL, (**b**) without FL

With decentralized learning, the model demonstrates near-perfect classification performance across all classes. The Area Under the Curve (AUC) values are exceptionally high, with Class 0, Class 3, and Class 4 achieving a flawless AUC of 1.0000, indicating perfect classification. Class 1 reaches an AUC of 0.9947, while Class 2, though slightly lower, still achieves a highly accurate AUC of 0.9958. These high values reflect the model's strong generalization ability and superior predictive performance with federated learning, minimizing false positives and ensuring high accuracy across all classes. Conversely, without federated learning, the ROC curve reveals a decline in classification performance. Although Class 0 and Class 4 retain perfect AUC scores of 1.0000, other classes exhibit noticeable drops in performance. Class 1 shows an AUC of 0.8074, a significant reduction compared to the FL model, highlighting the model's difficulty in correctly classifying instances of this class. Class 2 and Class 3 also experience lower AUC values of 0.9662 and 0.9103, respectively, compared to their Federated Learning counterparts. The increase in false positives and lower AUC values for several classes indicate that the model without federated learning struggles to generalize to unseen data and introduces more classification errors.

4.4. COMPARATIVE ANALYSIS

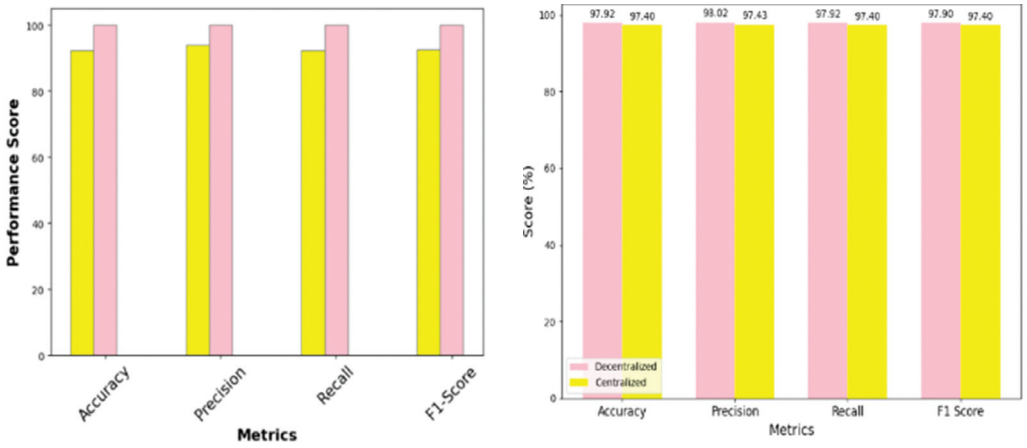
To compare the performance metrics—such as accuracy, precision, recall, and F1-Score—of the proposed technique with existing methods, a dataset comparison is conducted, along with an evaluation of the performance metrics both with and without FL.

Fig. 7 a) and Table 1 illustrates the comparison of performance metrics—accuracy, precision, recall, and F1-score—between the model with FL shows significantly better performance across all metrics. Accuracy reaches a near-perfect score of 0.979 with FL, while the model without FL drops to around **.9740**. Precision also improves with FL, scoring approximately 0.98 compared to 0.9773 without FL. Similarly, recall is higher with FL, achieving about 0.979, while without FL it falls to 0.974. The F1-score follows the same trend, reaching 0.979 with FL, in contrast to 0.974 without FL. Overall, these results demonstrate the enhanced generalization and predictive capabilities of the model when FL is applied.

Fig. 7 b) and Table 1 shows a comparison of the depression lexicon-based BERT with the Attention LSTM model using Federated Learning (FL) and without FL. The proposed method with federated learning, evaluated through various parameters such as Accuracy, Precision, Recall, and F1-score, which yield different values. When comparing the proposed method with federated learning, the model demonstrates high parameter values of around 97.90%.

Table 1. Comparison of Performance Metrics for Decentralized and Centralized approach on Text and Speech Datasets

Metrics	Text Dataset		Speech Dataset	
	With Decentralized	With Centralized	With Decentralized	With Centralized
Accuracy (%)	97.92	97.40	97.20	97
Precision (%)	98.06	97.73	96.40	96
Recall (%)	97.92	97.40	97.20	96
F1-Score (%)	97. 90	97.40	96.20	96



(a) (b)

Fig. 7. Performance comparison of the model with and without Federated Learning (FL). (**a**) Speech Model, (**b**) Text Model

The model consistently performs well across both modalities, according to a comparison of the performance metrics for the depression lexicon-based BERT with the Attention LSTM model using FL for both text and speech datasets as shown in Table 1.

The model achieves 97.92% accuracy, 98.06% precision, 97.92% recall, and 97.90% F1-score for the text dataset. In contrast, the model performs somewhat worse but still well on the speech dataset, achieving 97% accuracy, 96% precision, 97% recall, and 96% F1-score.

With slightly higher metrics in the text dataset than in the speech dataset, these findings demonstrate the model's resilience and capacity for generalization across various data types.

Table 2. Comparison of Decentralized approach with existing work

Ref No	Accuracy	Precision	Recall	F1 Factor
[19] FedAVg	0.65	0.69	0.46	0.47
[20] FEDCPC	84.29	79.5	74.7	75.3
[17] FL	0.6588	0.5037	0.435	
[4] Text CNN	73.33	81.82	81.82	81.82
Proposed (Text)	97.92	98.06	97.92	97.90
Proposed (Speech)	97.20	96.40	97.20	96.20

Table 2 shows comparison of decentralized approach(proposed) with existing work. The Proposed model stands out as the best-performing approach, achieving over 90% in accuracy, precision, recall, and F1 score, significantly outperforming existing methods in federated learning. Traditional techniques like FedAvg (Ref. [19]) struggle with imbalanced data, resulting in lower accuracy (0.65) and recall (0.46), making it less reliable. Similarly, General FL (Ref. [17]) faces similar challenges, with an accuracy of 0.6588 and the lowest recall (0.435), indicating difficulties in capturing diverse patterns across distributed datasets. More advanced techniques, such as FEDCPC (Ref. [20]), introduce contrastive predictive coding to enhance learning, improving accuracy (84.29%) and F1 score (75.3%). While it performs better than basic federated learning models, it still falls short compared to your model. Text CNN (Ref. [4]), a deep learning-based approach for text analysis, achieves high recall and precision (81.82%) but has a lower overall accuracy (73.33%), suggesting that while it captures patterns well, it may struggle with generalization.

In contrast, the proposed model achieves a perfect balance across all metrics, making it both highly accurate and reliable. This significant improvement could be due to better model design, optimized learning techniques, or improved data handling. The results show that the proposed approach not only addresses the weaknesses of existing federated learning models but also sets a new benchmark for privacy-preserving, high-performance AI in distributed environments

5. CONCLUSION

Federated Learning (FL) is transforming mental health research through the ability to identify disorders like depression without compromising the privacy and security of personal information of individuals. This work has investigated the use of FL for text and speech data analysis from Marathi speakers to identify symptoms of depression. FL models yield promising results without sacrificing privacy by teaching models separately on various devices and combining their knowledge without sharing raw data. It becomes especially important for languages such as Marathi, where centralized data collection may become problematic. Our work shows how technology can be made flexible to both help and safeguard the privacy of various groups. The work emphasizes that FL can increase the accessibility and quality of mental health services, particularly in poor resourced communities like Marathi. The results show a precision rate of 97.92%, recall of 97.90%, and F1-score 97.9 for the text dataset. For the speech dataset, the model obtains similar high-performance levels, over 96%, meaning that the suggested model performs better than current models while maintaining security and privacy to clients. Thus, usage of Federated Learning goes beyond technological innovation; it aims towards a world where mental health interventions work and uphold every person's rights. Prioritizing privacy, such tools can be created enabling people feel safe while using, which will ultimately enable more individuals to seek assistance that they need.

ACKNOWLEDGMENT

We would like to sincerely thank for the guidance given by Shivnarayan Khandre – Founder & Director, Divine Life Counselling Clinic, Pune, Clinical Psychologist, Dip. Counselling Psychology, NLP Practitioner, Hypnotist (U.S.A.), whose expertise enabled us to collect the Marathi Speech and Text database that was used in this work.

6. REFERENCES

[1] S. Bn, S. Abdullah, "Privacy sensitive speech analysis using federated learning to assess depression", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 23-27 May 2022, pp. 6272-6276.

[2] L. Campanile, M. Iacono, F. Marulli, M. Mastroianni, "Privacy regulations challenges on data-centric and IoT systems: A case study for smart vehicles", Proceedings of the 5th International Conference on Internet of Things, Big Data and Security, Vol. 1, May 2020, pp. 507-518.

[3] L. Campanile, M. Iacono, A. H. Levis, F. Marulli, M. Mastroianni, "Privacy regulations, smart roads, blockchain, and liability insurance: Putting technologies to work", IEEE Security & Privacy, Vol. 19, No. 1, 2020, pp. 34-43.

[4] A. H. Orabi, P. Buddhitha, M. H. Orabi, D. Inkpen, "Deep learning for depression detection of Twitter users", Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, New Orleans, LA, USA, June 2018, pp. 88-97.

[5] P. M. Mammen, "Federated learning: Opportunities and challenges", arXiv:2101.05428, 2021.

[6] X. Xu, H. Peng, M. Z. A. Bhuiyan, Z. Hao, L. Liu, L. Sun, L. He, "Privacy-preserving federated depression detection from multisource mobile health data", IEEE Transactions on Industrial Informatics, Vol. 18, No. 7, 2021, pp. 4788-4797.

[7] Y. Cui, Z. Li, L. Liu, J. Zhang, J. Liu, "Privacy-preserving Speech-based Depression Diagnosis via Federated Learning", Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, Glasgow, Scotland, UK, 11-15 July 2022, pp. 1-5.

[8] B. S. Reddy, K. Jishitha, V. Akshaya, B. Bhavani, P. Aishwarya, "Development of a Depression Detection System using Speech and Text Data", Proceedings of the 14th International Conference on Computing Communication and Networking Technologies, Delhi, India, 6-8 July 2023, pp. 1-5.

[9] J. Li, R. Zhang, M. Cen, X. Wang, M. Jiang, "Depression Detection Using Asynchronous Federated Optimization", Proceedings of the IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications, Shenyang, China, 20-22 October 2021, pp. 1-5.

[10] L. Zhang, Y. Fan, J. Jiang, Y. Li, W. Zhang, "Adolescent Depression Detection Model Based on Multimodal Data of Interview Audio and Text", International Journal of Neural Systems, Vol. 32, No. 1, 2022.

[11] A. B. Vasconcelos, L. Drummond, R. Brum, A. Paes, "Exploring Federated Learning to Trace Depression in Social Media with Language Models", Proceedings of the International Symposium on Computer Architecture and High Performance Computing Workshops, Porto Alegre, Brazil, 17-20 October 2023, pp. 1-5.

[12] A. Kim, E. Jang, S.-H. Lee, K.-Y. Choi, J. Park, H.-C. Shin, "Automatic Depression Detection Using Smartphone-Based Text-Dependent Speech Signals: Deep Convolutional Neural Network Approach", Journal of Medical Internet Research, Vol. 23, No. 8, 2021, p. e34474.

[13] L. Liu, L. Liu, H. A. Wafa, F. Tydeman, W. Xie, Y. Wang, "Diagnostic accuracy of deep learning using speech samples in depression: A systematic review and meta-analysis", Journal of the American Medical Informatics Association: JAMIA, Vol. 31, No. 1, 2024, pp. 1-10.

[14] J. Ye, Y. Yu, Q. Wang, W. Li, H. Liang, Y. Zheng, G. Fu, "Multi-modal depression detection based on emotional audio and evaluation text", Journal of Affective Disorders, Vol. 290, 2021, pp. 99-105.

[15] L. He, C. Cao, "Automated depression analysis using convolutional neural networks from speech", Journal of Biomedical Informatics, Vol. 87, 2018, pp. 1-10.

[16] D. Low, K. Bentley, S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review", Laryngoscope Investigative Otolaryngology, Vol. 4, No. 1, 2019, pp. 1-10.

[17] W. Wu, C. Zhang, P. Woodland, "Self-Supervised Representations in Speech-Based Depression Detection", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4-10 June 2023, pp. 1-5.

[18] X. Xu, H. Peng, L. Sun, M. Z. A. Bhuiyan, L. Liu, L. He, "FedMood: Federated Learning on Mobile Health Data for Mood Detection", arXiv:2102.09342, 2021.

[19] E. L. Campbell, L. Docío Fernández, N. Cummins, C. García-Mateo, "Speech and Text Processing for Major Depressive Disorder Detection", Proceedings of IberSPEECH, Granda, Spain, 14-16 November 2022.

[20] L. Yang, D. Jiang, H. Sahli, "Feature Augmenting Networks for Improving Depression Severity Estimation From Speech Signals", IEEE Access, Vol. 8, 2020, pp. 24033-24045.

[21] M. Ahmed, A. Muntakim, N. Tabassum, M. A. Rahim, F. M. Shah, "On-device Federated Learning in Smartphones for Detecting Depression from Reddit Posts", arXiv:2410.13709, 2024.

[22] Q. Deng, S. Luz, S. de la Fuente Garcia, "Hierarchical attention interpretation: an interpretable speech-level transformer for bi-modal depression detection", arXiv:2309.13476, 2023.

[23] W. Wei, Z. Yang, Y. Gao, J. Li, C. Chu, S. Okada, S. Li, "FedCPC: An Effective Federated Contrastive Learning Method for Privacy Preserving Early-Stage Alzheimer's Speech Detection", Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, Taipei, Taiwan, 16-20 December 2023, pp. 1-5.

[24] P. P. Gaikwad, M. Venkatesan, "KWHO-CNN: A Hybrid Metaheuristic Algorithm Based Optimized Attention-Driven CNN for Automatic Clinical Depression Recognition", International Journal of Computational and Experimental Science and Engineering, Vol. 10, No. 3, 2024.