FedKD4DD: Federated Knowledge Distillation for Depression Detection

Aslam Jlassi[1], Afef Mdhaffar[1,2][a], Mohamed Jmaiel[1,2][b] and Bernd Freisleben[3][c]
[1]*ReDCAD Laboratory, ENIS, University of Sfax, Sfax, Tunisia*
[2]*Digital Research Center of Sfax, 3021, Sfax, Tunisia*
[3]*Department of Mathematics and Computer Science, University of Marburg, 35032, Marburg, Germany*

Keywords: Federated Learning, Knowledge Distillation, Depression, Call Recording, Self-Distillation.

Abstract: Depression affects over 280 million people globally and requires timely, accurate intervention to mitigate its effects. Traditional diagnostic methods often introduce delays and privacy concerns due to centralized data processing and subjective evaluations. To address these challenges, we propose a smartphone-based approach that uses federated learning to detect depressive episodes through the analysis of spontaneous phone calls. Our proposal protects user privacy by retaining data locally on user devices (i.e., smartphones). Our approach addresses catastrophic forgetting through the use of knowledge distillation, enabling efficient storage and robust learning. The experimental results demonstrate reasonable accuracy with minimal resource consumption, highlighting the potential of privacy-preserving AI solutions for mental health monitoring.

1 INTRODUCTION

Depression, a prevalent mental health disorder, impacts more than 280 million individuals worldwide, according to the World Health Organization (WHO) (World Health Organization, 2023). As one of the leading causes of disability, major depression disorder profoundly affects both mental and physical health, necessitating early detection of depressive episodes. Traditional diagnostic practices are often lengthy and rely on subjective tools such as interviews and questionnaires. Although effective in diagnosing Major Depressive Disorder (MDD), these methods usually fail to detect depressive episodes, which can prevent early treatment. Furthermore, the sensitive nature of mental health data raises significant privacy concerns, making it essential to explore diagnostic systems that protect individual privacy.

Existing approaches have explored the use of deep learning algorithms for detecting depressive episodes based on phone call analysis (Mdhaffar et al., 2019). However, these models are typically trained offline on static datasets, which limits their ability to adapt to new patterns or features that may emerge over time. Moreover, these deep leaning models raise concerns about data privacy as they require sensitive user data to be transmitted to external servers for processing. To address these issues, recent studies (Ma et al., 2022), (Zhang et al., 2023), (Shenaj et al., 2023), (Huang et al., 2022), (Lee et al., 2022) have adopted federated learning, a decentralized machine learning paradigm that analyzes data locally on user devices. These approaches ensure that sensitive data never leaves the user's device, safeguarding privacy while enabling real-time monitoring and adaptive learning. By decentralizing the training process, they overcome the privacy challenges associated with centralized models while mitigating risks of overfitting by training on diverse, user-specific data distributions. However, these approaches usually suffer from the catastrophic forgetting and communication overhead issues.

To tackle the challenges of catastrophic forgetting and communication overhead, we propose a novel federated approach, called FedKD4DD. It leverages the federated learning paradigm, complemented by knowledge distillation (KD) to reduce communication overhead, and facilitates the efficient sharing of knowledge across decentralized mobile devices. KD allows the model to retain previously learned knowledge across training iterations by saving logits from past training rounds. Unlike prior works that rely on explicit teacher-student model configurations, our approach uses self-distillation, optimizing storage usage

[a] https://orcid.org/0000-0002-5696-5771
[b] https://orcid.org/0000-0002-2664-0204
[c] https://orcid.org/0000-0002-7205-8389

Jlassi, A., Mdhaffar, A., Jmaiel, M. and Freisleben, B.
FedKD4DD: Federated Knowledge Distillation for Depression Detection.
DOI: 10.5220/0013388200003890
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 1473-1480
ISBN: 978-989-758-737-5; ISSN: 2184-433X
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

and ensuring the model's consistency in detecting depressive symptoms over time.

Our approach combines advanced machine learning techniques with resource-efficient design, enabling deployment on a wide range of mobile devices. A series of experiments was conducted to evaluate its performance in terms of accuracy, storage efficiency, communication overhead, battery usage, CPU consumption, and RAM usage. The results demonstrate that FedKD4DD achieves satisfactory detection accuracy with only a minimal drop in accuracy during the second run. Moreover, experimental results demonstrate low communication overhead and low computational demands, highlighting its potential as a scalable and privacy-preserving solution for real-world mental health monitoring.

The remainder of this paper is organized as follows. Section 2 provides an overview of the fundamental concepts relevant to this study, while Section 3 discusses state-of-the-art approaches. Section 4 introduces FedKD4DD, detailing its architecture and its core components. Section 5 presents implementation aspects. Section 6 discusses the obtained experimental results. Finally, Section 7 concludes the paper and outlines areas for future research.

2 FUNDAMENTAL CONCEPTS

FedKD4DD for early detection of depressive symptoms draws on a combination of fundamental concepts and advancements in related research fields. Depression, as a critical focus, highlights the necessity of innovative solutions for mental health care, while federated learning provides the technological basis for enabling privacy-preserving, decentralized data processing. To further enhance FedKD4DD's learning efficiency and adaptability, knowledge distillation is integrated to address the challenge of catastrophic forgetting in continual learning scenarios.

2.1 Depression

Depression is a mental health disorder that manifests itself through emotional distress, cognitive impairment, and physical symptoms that disrupt daily life and productivity (American Psychiatric Association, 2013). Depression takes various forms, including Major Depressive Disorder, Bipolar Disorder, and Seasonal Affective Disorder, each with unique characteristics (National Institute of Mental Health, 2023; Rosenthal et al., 1984). Its societal impact is profound, including lost productivity, increased healthcare costs, and family burdens (Goodman and Gotlib, 2002). Early detection is crucial to mitigate these effects, making the development of efficient and privacy-preserving detection systems a priority.

2.2 Federated Learning

Federated learning is a decentralized approach that allows multiple clients to collaboratively train a model while retaining data locally (McMahan et al., 2017). By transmitting model updates instead of raw data, this method addresses privacy concerns and supports compliance with regulations like GDPR. The process involves iterative updates between clients and a central server, employing techniques like Federated Averaging to aggregate local contributions into a global model.

2.3 Knowledge Distillation

Knowledge distillation transfers the learning of a complex teacher model to a simpler student model by leveraging the teacher's soft output, which provides more information than hard labels (Hinton et al., 2015). This technique enables the student to approximate the teacher's decision boundaries while maintaining computational efficiency (Yim et al., 2017). Distillation can occur offline, online, or through self-distillation, each offering unique advantages (Zhang et al., 2020; Furlanello et al., 2018).

3 RELATED WORK

(Ma et al., 2022) introduce a framework utilizing knowledge distillation to address catastrophic forgetting, primarily tested on image classification tasks. Their solution achieves a trade-off between storage efficiency and communication overhead, which is particularly beneficial for bandwidth-constrained systems. However, their approach is tailored to image data and heavily relies on pre-trained teacher models, limiting its adaptability to domains where pre-trained models are unavailable or impractical. Moreover, their framework does not address the challenges specific to resource-constrained devices, such as smartphones, which are critical for real-world applications.

(Zhang et al., 2023) propose an exemplar-free knowledge distillation approach that eliminates the need to store raw data, effectively addressing privacy concerns. Although this method demonstrates robust generalization across tasks, it suffers from a high communication overhead, which constrains scalability for large networks or mobile platforms. Our approach mitigates these issues by retaining logits locally on devices, reducing the synchronization frequency, and making it suitable for mobile environments with limited bandwidth.

(Shenaj et al., 2023) present an asynchronous update mechanism to alleviate communication bottlenecks and improve scalability. Although this method allows independent client updates, it introduces complexities in managing divergent model states. Unlike their method that requires intricate synchronization management, our approach simplifies this process by focusing on logit-based distillation. This does not only reduce communication demands, but does also effectively addresses catastrophic forgetting, while optimizing device resources.

(Huang et al., 2022) tackle personalization in heterogeneous federated learning networks, enabling adaptation to diverse client data distributions. Their framework is flexible and robust, but increases solution complexity. While our work prioritizes efficiency and scalability over personalization, its specialization in audio-based depression detection ensures relevance to healthcare applications, where privacy and data efficiency are paramount.

(Lee et al., 2022) propose a “not-true” distillation technique to enhance storage efficiency and reduce communication costs by storing distilled knowledge representations instead of raw data. Although this approach is advantageous for devices with limited resources, its performance on complex and diverse tasks is limited. Our solution builds on this strength by tailoring logit-based storage for depression detection, balancing storage efficiency and task-specific performance.

Unlike existing approaches, our proposal addresses the challenges of high communication overhead and storage inefficiency by utilizing knowledge distillation in a novel way for mobile devices. Instead of relying on frequent model synchronization, we focus on training a local model on audio data collected from phone calls and use knowledge distillation to store compressed representations of the learned knowledge (logits). These logits are retained for future training sessions, allowing us to delete raw data after each training cycle, freeing up storage space on the device. This ensures continuous learning from new data while retaining essential information from prior tasks, effectively mitigating catastrophic forgetting.

4 FedKD4DD

This section presents our proposed approach, called FedKD4DD. It stands for “Federated Knowledge Distillation for Depression Detection”. FedKD4DD addresses catastrophic forgetting and privacy challenges in federated learning. Using knowledge distillation, we store logits to enable efficient training on mobile devices while minimizing storage and communication overhead. FedKD4DD focuses on audio-based depression detection, extending federated learning to healthcare applications.

4.1 Architecture

The architecture of FedKD4DD leverages federated learning to handle decentralized datasets, such as phone call recordings on client devices (i.e., smartphones). By incorporating knowledge distillation and automated data labeling, FedKD4DD enables clients to contribute to a global model while retaining sensitive data locally. This design preserves user privacy and ensures adaptability to individual data distributions, addressing challenges in personalized healthcare applications.

The overall architecture follows a decentralized, multi-layered structure consisting of a central server and multiple clients (mobile devices) participating in federated learning, as depicted in Figure 1. Each client maintains its dataset, consisting of phone call recordings, which are pre-processed and labeled automatically before being used for training. The central server aggregates the updates from all clients without accessing their raw data.

To minimize communication overhead, the system exchanges model parameters rather than raw data or logits, enabling scalability and efficiency. Local training on personalized datasets enhances model adaptability while preserving user privacy.

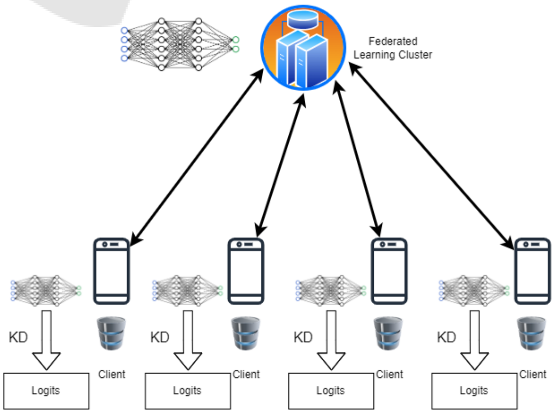

Figure 1: Architecture of FedKD4DD.

4.2 Key Components

The architecture consists of several core components. They are detailed in the following.

1. **Central Server:**
   - **Global Model Initialization:** The server initializes a pre-trained global model to be distributed to all clients. This model serves as a starting point for localized training.
   - **Model Aggregation:** The server aggregates updated model parameters from clients using Federated Averaging (FedAvg) (McMahan et al., 2017). This process combines client updates proportionally based on the size of their datasets, refining the global model iteratively.
2. **Clients:** Each client represents a mobile device with a personalized dataset. Key tasks performed by clients include:
   - **Data Collection and Labeling:** Clients automatically record and store phone calls locally. These recordings are labeled using a pre-trained model integrated into the application.
   - **Local Model Training:** Clients train the global model on their labeled datasets, tailoring it to their data distribution.
   - **Logit Storage for Knowledge Distillation:** Clients save logits from their training sessions for use in subsequent training rounds.
   - **Communication with Central Server:** Clients send updated model parameters to the server without sharing raw data, ensuring privacy.
3. **Knowledge Distillation Mechanism:** Self-distillation is employed to mitigate catastrophic forgetting. The clients use their previously saved logits to compute a distillation loss in the next training round, aligning current training with past knowledge.

4.3 Data Flow

FedKD4DD's data flow is designed to ensure efficient training, preprocessing, and communication between the server and clients while optimizing resource utilization on mobile devices. By incorporating a preprocessing module and leveraging a structured pipeline, the system transforms raw audio recordings into machine learning-ready inputs, enabling effective feature extraction and noise reduction. FedKD4DD's data flow consists of the following key steps:

- **Model Initialization:** The server initializes a globally shared model and distributes it to all participating clients. This step ensures that all clients start with a common baseline for local training, enabling effective aggregation during the training process.
- **Data Collection and Labeling:** Clients record phone calls locally using the application. These recordings are automatically labeled based on predefined criteria, leveraging the pre-trained model embedded within the app. This automation reduces user intervention and ensures consistent labeling quality.
- **Data Preprocessing:** To prepare raw audio data for machine learning, a robust preprocessing pipeline is employed. Figure 2 provides a visual representation of the preprocessing pipeline.
  1. **Silence Removal:** Silent segments in the recordings are removed to focus on the active portions of the signal. This is achieved using an energy threshold-based algorithm, which analyzes the signal in fixed temporal windows:

     $$\text{Trim}(y) = y[t_1 : t_2] \quad \text{where } E(y_t) > \text{threshold} \tag{1}$$

     Here, $E(y_t)$ represents the energy of the signal in a window t, calculated as:

     $$E(y_t) = \frac{1}{N}\sum_{n=1}^{N} y_t^2 \tag{2}$$

     Segments with energy below the threshold are excluded, reducing redundancy and computational overhead.
  2. **Amplitude Normalization:** After silence removal, the audio signal is normalized to account for variations in recording devices. This step scales the signal to a standardized range, ensuring consistency across all recordings:

     $$\text{Normalize}(y) = \frac{y}{\max(|y|)} \tag{3}$$

     where $\max(|y|)$ is the maximum absolute amplitude of the signal.
  3. **Spectrogram Conversion:** The normalized audio is transformed into spectrograms using the Short-Time Fourier Transform (STFT), which provides a frequency-based representation of the signal:

     $$\text{STFT}(y)(t,f) = \sum_{n=0}^{N-1} y[n] \cdot w[n] \cdot e^{-j2\pi fn/N} \tag{4}$$

     Here, $w[n]$ is a Hamming window function applied to reduce edge effects. The spectrogram is then converted to decibels to align with human perceptual scales:

$$\text{Spectrogram}(t,f) = 10 \cdot \log_{10}(|\text{STFT}(y)(t,f)|^2 + \varepsilon) \quad (5)$$

where ε is a small value added to avoid logarithmic errors.

4. **Spectrogram Resizing:** Spectrograms are resized to 256 × 256 pixels to match the neural network's input requirements.

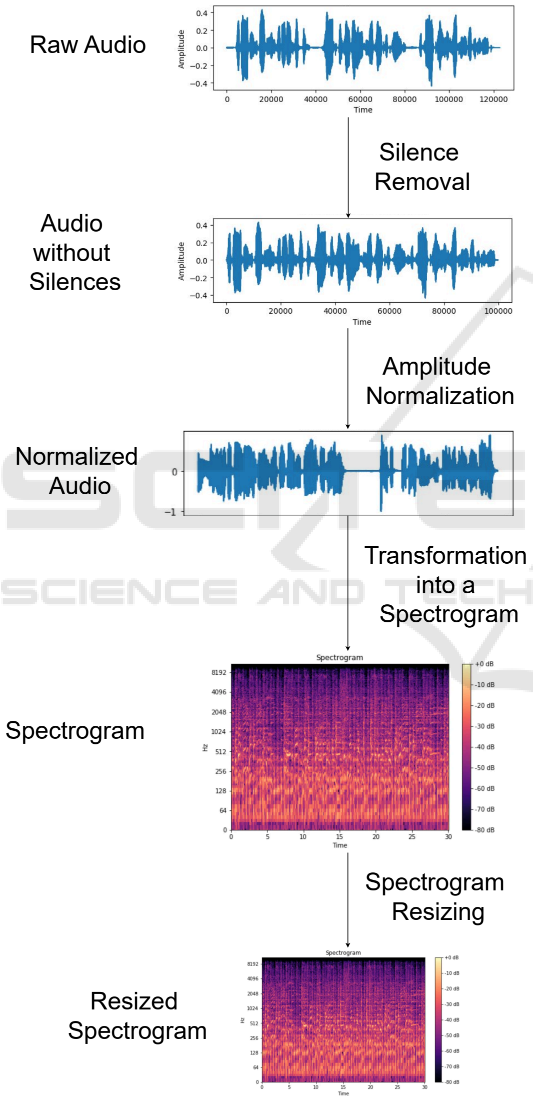

Figure 2: Preprocessing Pipeline for Audio Data.

- **Local Training:** Each client trains the global model locally using its preprocessed and labeled data. This step allows the model to learn from client-specific data distributions while preserving user privacy.
- **Logit Storage:** Upon completing local training, clients store the output logits for all training samples. These logits are critical for future rounds of training, enabling the implementation of self-distillation to mitigate catastrophic forgetting.
- **Model Update and Communication:** Clients send their locally updated model parameters to the central server. The server aggregates these updates using a weighted averaging approach to produce a new global model.
- **Self-distillation:** During subsequent training rounds, clients utilize stored logits to compute a distillation loss. This process aligns the new model's predictions with prior knowledge, ensuring that previously learned information is preserved while integrating new data.

The combination of automated preprocessing, local training, and knowledge distillation enables the system to handle the challenges of audio-based tasks, resource constraints, and catastrophic forgetting.

5 IMPLEMENTATION

The development of the federated learning application required careful integration of various components and technologies to achieve seamless functionality, privacy preservation, and efficient performance. The mobile application was developed using Android Studio that integrates seamlessly with TensorFlow Lite and TensorFlow Federated used by FedKD4DD. TensorFlow Federated was used to manage federated learning communication between clients and the central server. It simplifies the orchestration of distributed training and supports privacy-preserving protocols. Federated learning was locally tested using FLWR, a framework that facilitates communication between client devices and the central server, ensuring privacy by exchanging only model parameters rather than raw data. This setup is crucial for enabling personalized learning without compromising user privacy. Differential privacy was also integrated into the federated learning protocol to mask individual contributions by adding noise to model updates, ensuring compliance with privacy regulations and enhancing user trust. The mathematical guarantees of differential privacy are formally defined in Equation (6) (Dwork et al., 2006).

$$\Pr[\mathcal{A}(D_1) \in S] \leq e^{\varepsilon} \Pr[\mathcal{A}(D_2) \in S] \quad (6)$$

where $\mathcal{A}$ is the algorithm, D_1 and D_2 are datasets differing in one element, and ε controls the privacy loss. A smaller ε value implies stronger privacy guarantees.

5.1 Neural Network for Audio Processing

The core of FedKD4DD is a convolutional neural network (CNN) combined with a Long Short-Term Memory (LSTM) layer to handle the sequential nature of audio data. Audio preprocessing involves transforming raw waveforms into spectrograms, which capture time and frequency components of the audio signal, making it suitable for deep learning models. Key architectural components include:

- **Convolutional Layers:** Extract spatial features from spectrograms, highlighting frequency patterns.
- **LSTM Layer:** Capture temporal dependencies in sequential data, such as speech patterns.
- **Dropout and Batch Normalization:** Prevent overfitting and stabilize training.
- **Fully Connected Layers:** Translate learned features into classification outputs.

5.2 Local Testing and Preprocessing

Prior to deploying the application, extensive local testing was conducted to validate the key components of the system. This phase involved leveraging `PyTorch` for model training and `Librosa` for audio preprocessing, ensuring that the proposed methods were both effective and compatible with the federated learning framework. The FLWR library (*Flower*) was employed to simulate federated learning during local experimentation, enabling the evaluation of model aggregation and communication efficiency in a controlled environment.

The preprocessing pipeline focused on converting raw audio signals into features suitable for model training. Using `Librosa`, the pipeline performed silence removal to eliminate non-informative segments, amplitude normalization to standardize the audio signals, and spectrogram generation for feature extraction. Specifically, audio signals were transformed into Mel-frequency cepstral coefficients (MFCCs), a representation that captures the frequency content most relevant to human perception. This approach was chosen to improve the accuracy of depression detection by emphasizing meaningful audio patterns.

The spectrograms were resized to match the input dimensions required by the model, ensuring compatibility and efficient batch processing during training. This preprocessing strategy was designed to minimize computational overhead while preserving key audio features, aligning with the constraints of resource-limited mobile devices.

6 EXPERIMENTAL RESULTS

This section presents and discusses the conducted experiments, using the DAIC-WOZ dataset (Gratch et al., 2014). Our evaluation examines multiple performance metrics, including accuracy, resource consumption, and computational feasibility.

6.1 Experimental Setup

The experiments were conducted using the DAIC-WOZ dataset (Gratch et al., 2014), a publicly available resource for depression detection that provides annotated vocal recordings. This dataset was chosen for its relevance to the project, as it enables the development of machine learning models tailored to audio-based psychological analysis. However, the dataset posed several challenges, such as its relatively small size and significant class imbalance. The dataset includes 170 samples labeled as non-depressed (class 0) and 49 samples labeled as depressed (class 1), resulting in a class ratio of approximately 3.5 : 1. Such an imbalance can bias models toward predicting the majority class, thereby reducing the accuracy of depression detection for the minority class.

To address this issue, the “Synthetic Minority Over-sampling Technique (SMOTE)” was employed during preprocessing. SMOTE generates synthetic samples for the minority class by interpolating between existing data points and their nearest neighbors in feature space. This approach increases the diversity of the minority class without duplicating existing data, enhancing the model’s ability to generalize across both classes. By applying SMOTE, the dataset was effectively balanced, improving the model’s capacity to identify depressive symptoms and ensuring a fair evaluation of the federated learning approach.

The experiments simulated a federated learning environment by distributing the preprocessed DAIC-WOZ dataset among three clients. Two clients operated as virtual devices, emulated through Android Studio, and were configured with 4 GB of RAM and Android 13. These virtual devices provided a controlled testing environment for evaluating the app’s performance under consistent conditions. The third client was a Realme C51 smartphone, equipped with 4 GB of RAM and a Unisoc T612 processor. This device represented a typical resource-constrained mobile phone, aligning with the project’s goal of deploying federated learning solutions on real-world devices with limited computational and storage capabilities.

To compare FedKD4DD with related approaches, we implemented four state-of-the-art methods and evaluated them on the DAIC-WOZ dataset, using the same experimental setup.

6.2 Evaluation Metrics

To provide a comprehensive evaluation of FedKD4DD, the following metrics were analyzed:

- **Accuracy (%):** Measures the percentage of correctly classified samples over multiple rounds of training (Powers, 2011).
- **Storage Efficiency (MB):** Indicates the storage requirements for models and associated data on client devices.
- **Communication Overhead (MB):** Quantifies the data exchanged between clients and the central server during training.
- **Training Time (minutes):** Accounts for the duration of preprocessing, local training, and weight updates.
- **Battery Consumption (%):** Reflects the percentage of battery depleted per training session.
- **CPU and RAM Usage**: Represents the computational demand on devices during training.

6.3 Results

The results of the experiments, visualized through histograms, demonstrate the practical viability of the proposed approach. Each metric is discussed in detail below.

1. **Accuracy Analysis**

   The histogram in Figure 3 illustrates the accuracy achieved in the first and second training rounds. FedKD4DD attained an initial accuracy of 65%, which slightly decreased to 63% in the second round. This minor drop indicates the effectiveness of knowledge distillation in addressing catastrophic forgetting, as it ensures retention of previously acquired knowledge while accommodating new information.

   Compared to (Shenaj et al., 2023), which showed accuracies of 59% and 52%, the proposed approach demonstrates a notable improvement. This balance between performance and resource efficiency underscores its suitability for mobile deployment.

2. **Resource Efficiency**

   Table 1 highlights the storage and communication overhead of the proposed method. With a storage requirement of 220 MB and a communication overhead of 18 MB, the approach outperforms resource-intensive methods like (Shenaj et al., 2023) (400 MB, 30 MB). This efficiency is critical for mobile devices with limited storage and bandwidth. Additionally, the training time (see Table 1) was optimized at 38 minutes, significantly lower than 51 minutes reported by (Shenaj et al., 2023). This reduction demonstrates the effectiveness of preprocessing and the compact model architecture employed.

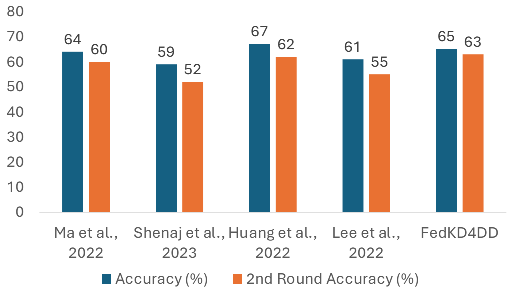

Figure 3: Accuracy Comparison Across Training Rounds.

3. **Computational Feasibility**

   As shown in Table 1, the CPU usage averaged 80%, while RAM consumption was 550 MB, both of which are well within the capabilities of modern smartphones. Furthermore, Table 1 shows that the average battery consumption was 7% per training session, lower than some other methods, such as (Shenaj et al., 2023) (10%).

7 CONCLUSION

This paper presented a federated learning solution to address catastrophic forgetting in mobile environments, focusing on depression detection from audio recordings. By leveraging knowledge distillation and storing logits instead of raw data, the approach ensured privacy preservation while maintaining resource efficiency, making it suitable for deployment on mobile devices.

The proposed method achieved competitive performance on the DAIC-WOZ dataset, with an initial accuracy of 65% and optimized resource consumption, including storage, communication overhead, and computational demands. Comparative analysis highlighted its advantages over related work, demonstrating its feasibility for resource-constrained environments.

However, slight accuracy drops across training rounds suggest opportunities for improvement in knowledge retention. Additionally, validating the approach on larger and more diverse datasets remains a priority to enhance its generalizability.

In the future, we aim to investigate other data sets

Table 1: Comparison of Model Performance using DAIC-WOZ Dataset.

Paper	Training Time (mins)	Battery Used (%)	CPU Usage (%)	RAM Usage (MB)
(Ma et al., 2022)	45	8	85	600
(Shenaj et al., 2023)	51	10	90	700
(Huang et al., 2022)	44	7	80	600
(Lee et al., 2022)	57	9	88	660
Our Approach	**38**	**7**	**80**	**550**

in federated learning such as non-IID datasets that might occur in our use case. Moreover, we plan to deal with the challenge of heterogeneous DL models that might be deployed on the clients. Finally, we want to propose a server-less federated learning approach that does not depend on a centralized aggregation server.

ACKNOWLEDGEMENTS

This work is supported by the German Academic Exchange Service (DAAD) in the Ta'ziz Science Cooperations Program (AirFit Project; 57682841).

REFERENCES

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. American Psychiatric Association Publishing, Washington, D.C., 5 edition.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*.

Furlanello, T., Decker, M. S. F. T., and Alabau, H. M. P. (2018). Born-again neural networks. *arXiv:1805.04770*.

Goodman, S. H. and Gotlib, I. H. (2002). Transmission of risk to children of depressed parents: Integration and conclusions. *Psychological Bulletin*, 128(5):768–795.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., and Morency, L.-P. (2014). The distress analysis interview corpus of human and computer interviews. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Huang, W., Ye, M., and Du, B. (2022). Learn from others and be yourself in heterogeneous federated learning. pages 10133–10143.

Lee, G., Jeong, M., Shin, Y., Bae, S., and Yun, S.-Y. (2022). Preservation of the global knowledge by not-true distillation in federated learning.

Ma, Y., Xie, Z., Wang, J., Chen, K., and Shou, L. (2022). Continual federated learning based on knowledge distillation. *International Joint Conferences on Artificial Intelligence Organization*, pages 2182–2188. Main Track.

McMahan, B., Moore, E., Ramage, D., and Y., S. H. C. Y. B. R. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282.

Mdhaffar, A., Cherif, F., Kessentini, Y., Maalej, M., Thabet, J. B., Maalej, M., Jmaiel, M., and Freisleben, B. (2019). DL4DED: Deep learning for depressive episode detection on mobile devices. In *Proceedings of the 17th International Conference on Smart Homes and Health Telematics: How AI Impacts Urban Living and Public Health, (ICOST)*, Lecture Notes in Computer Science, pages 109–121, New York City, NY, USA. Springer.

National Institute of Mental Health (2023). Bipolar disorder.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Rosenthal, N. E., Sack, D. A., Gillin, J. C., Lewy, A. J., Goodwin, F. K., Davenport, Y., Mueller, P. S., Newsome, D. A., and Wehr, T. A. (1984). Seasonal affective disorder: A description of the syndrome and preliminary findings with light therapy. *Archives of General Psychiatry*, 41(1):72–80.

Shenaj, D., Toldo, M., Rigon, A., and Zanuttigh, P. (2023). Asynchronous federated continual learning.

World Health Organization (2023). Depression: Key facts.

Yim, D., Li, D., and Liu, B. (2017). A gift from knowledge distillation: Fast optimization and efficient inference. In *ICML*.

Zhang, J., Chen, C., Zhuang, W., and Lv, L. (2023). Target: Federated class-continual learning via exemplar-free distillation.

Zhang, Y., Wang, R., and Xu, H. (2020). Deep mutual learning. In *Proceedings of the IEEE International Conference on Computer Vision*.