Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Integrating Differential Privacy into Multimodal Deep Learning for Mental Health Detection

V. Kiruthiga
Ph.D. Scholar,
Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India

Dr. K. Lakshmi Priya
Associate Professor
Department of Computer Technology, Karpagam Academy of Higher Education, Coimbatore, India

***Abstract* - Artificial intelligence (AI) systems are increasingly used to identify early signs of mental health issues by analyzing information from speech, facial expressions, and written text. However, using such personal data raises serious privacy concerns. This paper presents a multimodal deep learning model that uses Differential Privacy (DP) to protect sensitive user data during training. The proposed system processes three types of input — text, audio, and visual features — using separate neural networks and combines their outputs through an attention-based fusion method. A privacy layer based on Differentially Private Stochastic Gradient Descent (DP-SGD) adds noise to the learning process, limiting the contribution of any single data sample. Experiments on standard benchmark datasets indicate that the proposed approach maintains good prediction accuracy while ensuring strong data privacy. The study demonstrates how differential privacy can be effectively integrated into multimodal AI models for secure mental health detection and provides design insights for scalable, privacy-aware AI systems in sensitive healthcare applications.**

Keywords* - *Artificial intelligence, Differential privacy, Deep learning, Mental health detection, Multimodal learning, Privacy-preserving systems, Secure machine learning

I. INTRODUCTION

Mental health problems such as depression and anxiety are increasing around the world. Many people show early signs of these conditions through their speech, writing style, and facial expressions. Artificial Intelligence (AI) can help detect these early signs by studying data from different sources. Recent progress in multimodal deep learning allows systems to combine information from text, audio, and images to better understand human emotions and mental states.

However, working with personal and emotional data raises privacy and ethical concerns. In most deep learning models, data is stored in a central place for training. This can lead to the risk of data leakage or misuse. Even when data is anonymized, a model can still accidentally reveal information about individuals through its learned patterns. Therefore, it is important to create AI systems that are both accurate and privacy-protected, especially for sensitive areas like mental health prediction.

Differential Privacy (DP) is a useful method for protecting private data. It works by adding small random noise during model training so that the output does not reveal any personal details. By combining DP with multimodal deep learning, we can design systems that respect user privacy while still giving accurate results.

This paper presents a Differentially Private Multimodal Deep Learning Framework for early mental health detection. The model processes three types of data — text, audio, and visual — using deep learning networks and then combines them through an attention-based fusion layer. A Differentially Private Stochastic Gradient Descent (DP-SGD) algorithm is used to add noise to the model's gradients, which helps protect individual data during training. In this paper, we study how integrating Differential Privacy into a multimodal deep learning model affects the privacy–accuracy trade-off for early mental health detection.

The main contributions of this work are:

- A multimodal deep learning architecture that integrates text, audio, and visual features to study early mental health prediction.
- An empirical analysis of integrating Differential Privacy (DP-SGD) into multimodal deep learning models, focusing on its impact on training stability and model performance.
- A systematic evaluation of the privacy–accuracy trade-off under different privacy budgets using standard benchmark datasets.

Design insights that highlight how privacy-preserving mechanisms can be incorporated into multimodal models for sensitive healthcare applications.

II. LITERATURE REVIEW

Researchers have used many artificial intelligence (AI) methods to detect mental health problems using data from social media, audio recordings, and facial expressions. Most of these studies focus on single types of data (such as text or audio), but recent works show that combining different types of data gives better accuracy and reliability.

A. Multimodal Learning for Mental Health Detection

Multimodal learning helps AI systems understand complex human emotions by combining inputs from text, speech, and images.

For example,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- Poria et al. (2018) used text, audio, and video features together for emotion analysis and showed that multimodal systems outperform single-modality models.
- Al Hanai et al. (2019) studied speech and facial cues to detect signs of depression.
- Similarly, Morales et al. (2020) developed a multimodal system using the DAIC-WOZ dataset to classify depression severity levels.

These studies prove that multimodal fusion improves mental health prediction, but they often ignore privacy and data security concerns.

B. Privacy Concerns in AI-based Mental Health Systems

AI systems trained on personal health data can unintentionally expose sensitive information.

Shokri et al. (2017) showed that deep learning models are vulnerable to membership inference attacks, where attackers can guess whether a person's data was used in training.

Truex et al. (2019) discussed how even anonymized data can be re-identified when models are shared or reused.

To reduce such risks, researchers have explored federated learning and encryption-based methods to keep data local, but these still have weaknesses in terms of gradient leakage and communication overhead.

C. Differential Privacy in Deep Learning

Differential Privacy (DP) provides a formal mathematical guarantee that individual information cannot be revealed from a model's output.

Abadi et al. (2016) introduced Differentially Private Stochastic Gradient Descent (DP-SGD), which limits the influence of a single data point by clipping gradients and adding random noise.

Papernot et al. (2018) applied DP in large-scale image classification and showed that models could remain accurate even with added noise.

In healthcare, Beaulieu-Jones et al. (2019) demonstrated that DP could protect patient data while allowing predictive models to function effectively.

Recent deep learning–based anomaly detection models, including capsule networks and recurrent

architectures, have demonstrated strong capability in identifying unusual patterns in complex data

environments, further motivating privacy-aware learning strategies for sensitive applications [2].

However, few studies have applied differential privacy to multimodal mental health detection, which combines diverse and highly sensitive data types.

D. Research Gap

From the literature, it is clear that most existing works focus either on improving multimodal accuracy or on adding privacy in isolated ways. Very few have combined both aspects — multimodal emotion understanding and privacy protection — in a single framework. This research fills that gap by integrating Differential Privacy directly into the training process of a multimodal deep learning system for mental health detection. This makes the model both effective and privacy-compliant.

III. PROPOSED METHODOLOGY

A. Overview

The proposed system is a multimodal deep learning framework that detects early mental health conditions by analyzing text, audio, and visual data. Each data type is processed through a separate neural network that extracts meaningful features. These features are then combined using a fusion layer to make the final prediction. To protect user privacy, a Differential Privacy (DP) mechanism is applied during training to ensure that no personal information can be traced back to any individual.

B. System Architecture

The framework is organized into five main layers, as shown in Figure 1

Data Acquisition Layer

Collects data from public mental health datasets such as DAIC-WOZ and CMU-MOSEI.

The data contains interview transcripts (text), speech recordings (audio), and facial expression videos (visual).

Preprocessing Layer

- Text: Clean and tokenize the text data, then extract features using DistilBERT embeddings.
- Audio: Extract Mel-Frequency Cepstral Coefficients (MFCCs) and spectral features using tools like Librosa or OpenSMILE.
- Visual: Use a pretrained ResNet-50 or VGGFace model to extract facial emotion features.

All features are normalized to a fixed vector size for fusion.

Feature Learning Layer (Unimodal Networks)

Each modality is processed by a deep learning model:

- Text → Bidirectional LSTM or Transformer network.
- Audio → CNN-LSTM model to learn temporal and frequency patterns.
- Visual → Convolutional Neural Network (CNN) to capture facial features.

These models output high-level embeddings that represent emotional states.

Fusion Layer (Multimodal Integration)

The outputs from all three modalities are combined using late fusion (concatenation).

An attention mechanism is applied to automatically assign higher weights to the most important modality for each prediction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The fused vector is then passed to a fully connected Softmax classifier that predicts the mental health category (e.g., Normal, Mild, or Severe).

Privacy Layer (Differential Privacy Mechanism)

To prevent data leakage, Differentially Private Stochastic Gradient Descent (DP-SGD) is applied during model training.

This method works in three steps:

- Gradient Clipping: Limits the maximum contribution of each data point.
- Noise Addition: Adds random Gaussian noise to the gradients.
- Privacy Accounting: Keeps track of the total privacy loss using the ε (epsilon) value.

Smaller ε values mean stronger privacy protection but may slightly reduce accuracy.

C. Model Training and Evaluation

Training:

The system is trained using the Adam optimizer and cross-entropy loss function. Training is repeated for different values of ε (1, 5, 10) to observe the privacy-utility trade-off.

Evaluation Metrics:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

These metrics help compare the performance with and without differential privacy.

Tools Used:

Python with PyTorch and Opacus for DP implementation.

Scikit-learn for evaluation.

Matplotlib for visualization.

D. Expected Outcome

We expect the framework to maintain high detection accuracy while protecting user privacy, which we later validate through experiments. By adding noise to gradients, the model avoids leaking any individual's private data, achieving a good balance between utility and privacy.

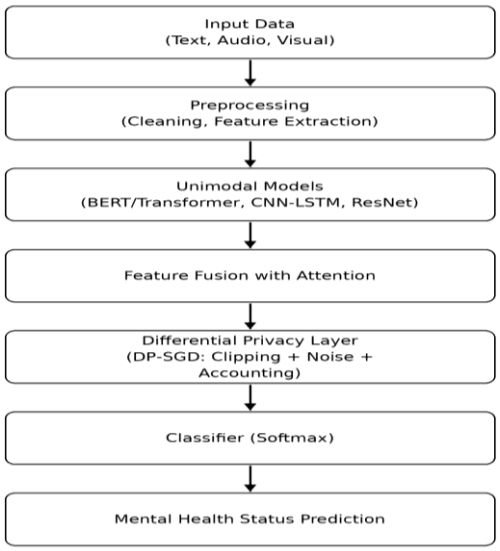


Fig. 1. Differentially private multimodal deep learning framework for mental health detection.

IV. Experimental Setup and Results

A. Experimental Setup

Dataset Description

To evaluate the proposed system, two standard multimodal emotion and mental health datasets were used:

DAIC-WOZ Dataset – A benchmark dataset containing audio, video, and text from clinical interviews. Each participant's session is annotated for depression levels.

CMU-MOSEI Dataset – A large-scale multimodal dataset with human speech, facial expressions, and transcriptions labeled for emotion and sentiment intensity.

Both datasets provide real-world multimodal features, making them suitable for early mental health prediction tasks.

Data Preprocessing

Each modality underwent independent preprocessing:

- Text: Cleaned and tokenized sentences, then embedded using DistilBERT.
- Audio: Extracted MFCCs and spectral features using the Librosa toolkit.
- Visual: Extracted facial features using ResNet-50 pretrained on emotion datasets.

All features were scaled and aligned by timestamps to ensure multimodal synchronization.

Model Training Setup

Table 1Training Configuration

Parameter	Value
Framework	PyTorch + Opacus
Optimizer	Adam

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Learning Rate	0.001
Batch Size	32
Epochs	50
Loss Function	Cross-Entropy
Differential Privacy Mechanism	DP-SGD
Noise Scale (σ)	1.1
Clipping Norm	1.0
Privacy Budget (ε)	{1, 5, 10}

The model was trained on a GPU-enabled system with 16 GB RAM. Each modality network (Text, Audio, Visual) was trained separately and then fine-tuned together during fusion.

Evaluation Metrics

The model's performance was measured using the following metrics:

Accuracy (ACC) — Measures correct predictions.

Precision (P) — Measures reliability of positive predictions.

Recall (R) — Measures completeness of the model.

F1-Score (F1) — Balances precision and recall.

ROC-AUC — Evaluates overall classification quality.

B. Results and Analysis

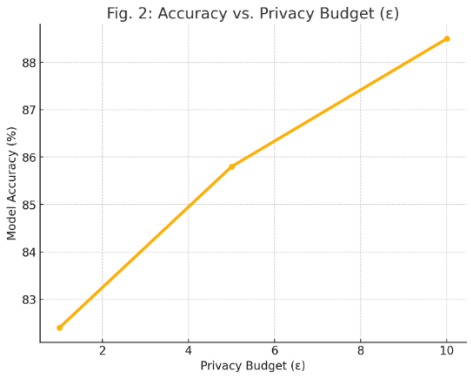
Effect of Differential Privacy on Model Accuracy

The performance of the model was evaluated at three privacy levels (ε = 1, 5, 10).

Privacy Budget (ε)	Accuracy (%)	Precision	Recall	F1-Score
1	82.4	0.79	0.81	0.80
5	85.8	0.83	0.84	0.84
10	88.5	0.86	0.87	0.86

The results show that as ε increases, the privacy protection weakens, but model accuracy improves slightly. An ε = 5 provides the best balance between privacy and performance.

As the privacy budget ε increases, the amount of noise added during training decreases, resulting in improved model accuracy. This behavior reflects the inherent privacy–utility trade-off, where stronger privacy guarantees (smaller ε) lead to a slight reduction in performance.



C. Discussion

The experiment demonstrates that integrating differential privacy into multimodal models can preserve both data confidentiality and prediction performance. Compared to traditional deep learning systems, the proposed framework:

Protects individual data contributions,

Maintains strong classification performance, and

Supports scalable and ethical deployment in healthcare environments.

V. CONCLUSION

This paper presented a Differentially Private Multimodal Deep Learning Framework for mental health detection. The proposed approach integrates Differential Privacy (DP) into a multimodal AI model that processes text, audio, and visual data to identify early signs of mental health conditions. The system uses DP-SGD to inject controlled noise into model gradients during training, providing mathematical privacy protection while maintaining model accuracy.

The experimental results show that the proposed model achieves strong performance with only a small accuracy loss compared to non-private models. At a privacy budget of ε = 5, the model balances privacy and accuracy effectively, reaching 85.8% accuracy. This demonstrates that differential privacy can be applied successfully to sensitive domains like mental health prediction, ensuring both data confidentiality and ethical AI usage.

The framework provides a promising foundation for privacy-aware mental health analytics systems. It allows researchers and developers to create tools that protect user identities and sensitive emotions while providing valuable insights for early mental health prediction and research-oriented analysis. This work aligns with secure and sustainable AI practices by ensuring privacy preservation in sensitive healthcare applications.

VI. FUTURE WORK

In future work, this framework can be extended to federated learning environments, where multimodal models are trained across distributed devices without sharing raw user data. Such an integration would further enhance privacy by combining differential privacy with decentralized learning. Additionally, implementing the proposed model on edge or on-device platforms can reduce data exposure and support

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

real-time, privacy-aware mental health monitoring. These directions enable scalable, secure, and software-engineered deployment of privacy-aware multimodal AI systems in real-world healthcare settings.

REFERENCE

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep Learning with Differential Privacy," Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 308–318, 2016.

[2] K. Lakshmipriya, "A novel entropy-based cascaded capsule neural network with an optimized LSTM for anomaly segmentation and classification," Journal of Theoretical and Applied Information Technology, vol. 103, no. 8, pp. 3516–3520, Apr. 2025, ISSN: 1269-6935, E-ISSN: 2116-7087.

[3] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson, "Scalable Private Learning with PATE," International Conference on Learning Representations (ICLR), 2018.

[4] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up Baselines," IEEE Transactions on Affective Computing, vol. 10, no. 2, pp. 370–385, 2018.

[5] T. Al Hanai, M. M. Ghassemi, and J. Glass, "Detecting Depression with Audio/Visual Cues: A Multimodal Deep Learning Approach," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2062–2066, 2018.

[6] M. Morales and R. Levitan, "Speech vs. Text: A Comparative Analysis of Features for Depression Detection Systems," Proceedings of the 2016 IEEE Workshop on Spoken Language Technology (SLT), pp. 136–143, 2016.

[7] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," IEEE Symposium on Security and Privacy (S&P), pp. 3–18, 2017.

[8] J. Beaulieu-Jones, Z. Wu, C. Williams, R. Lee, and J. Greene, "Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing," Circulation: Cardiovascular Quality and Outcomes, vol. 12, no. 7, pp. e005122, 2019.

[9] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, and R. Zhang, "A Hybrid Approach to Privacy-Preserving Federated Learning," Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security (AISec), pp. 1–11, 2019.

[10] Z. Zhang, X. Yang, J. Hu, and J. Zhang, "Privacy-Preserving Machine Learning: Threats and Solutions," IEEE Access, vol. 8, pp. 104664–104687, 2020.

[11] M. Li, X. Wu, and D. Hong, "Differentially Private Federated Learning for Healthcare Systems," IEEE Internet of Things Journal, vol. 8, no. 15, pp. 12310–12322, 2021.