iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73218

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

A Privacy-Preserving Framework for Mental Health Prediction Using Federated Learning

V. Kiruthiga[1], Dr. K. Lakshmi Priya[2]
[1]*Ph.D. Scholar,* [2]*Associate Professor, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India*

Abstract: ***The early detection of mental health disorders, particularly depression and anxiety, is essential for timely intervention and improved patient outcomes. Recent advances in Artificial Intelligence (AI) have enabled the development of systems capable of identifying psychological distress through multimodal data, including text, audio, and facial expressions. However, conventional AI models typically rely on centralized data collection, which poses significant risks to user privacy and data security, especially in healthcare applications involving sensitive personal information.***
This paper proposes a privacy-preserving framework for mental health prediction based on Federated Learning (FL). This decentralized training approach eliminates the need to transmit raw data to central servers. The proposed architecture integrates additional privacy-enhancing techniques such as differential privacy and secure aggregation to further safeguard user information. By enabling collaborative learning across distributed devices while maintaining data confidentiality, this framework offers a scalable and ethical solution for building AI systems in mental health care. The paper presents the system architecture, implementation strategy, and potential for real-world deployment.
Keywords: Mental Health Prediction, Federated Learning, Privacy-Preserving Machine Learning, Depression Detection, Multimodal AI, Differential Privacy, Secure Aggregation, Decentralized Learning, Ethical AI, AI in Healthcare.

I. INTRODUCTION

A. Introduction to the Problem in Detail

Mental health disorders such as depression, anxiety, and stress have become increasingly prevalent in recent years, affecting individuals across all age groups and socio-economic backgrounds. According to the World Health Organization (WHO), over 280 million people globally are estimated to suffer from depression alone. Despite the growing awareness of mental health, early detection remains a challenge due to stigma, lack of access to clinical care, and insufficient mental health infrastructure, especially in developing regions.

In this digital age, individuals frequently express their emotions and psychological states through online communication, social media, voice interactions, and facial expressions. These daily interactions contain valuable behavioural patterns that can be analysed using Artificial Intelligence (AI) to detect early signs of mental health disorders. Machine learning models, particularly those using multimodal data—such as text, audio, and video—have shown promising results in identifying depression-related symptoms with high accuracy.

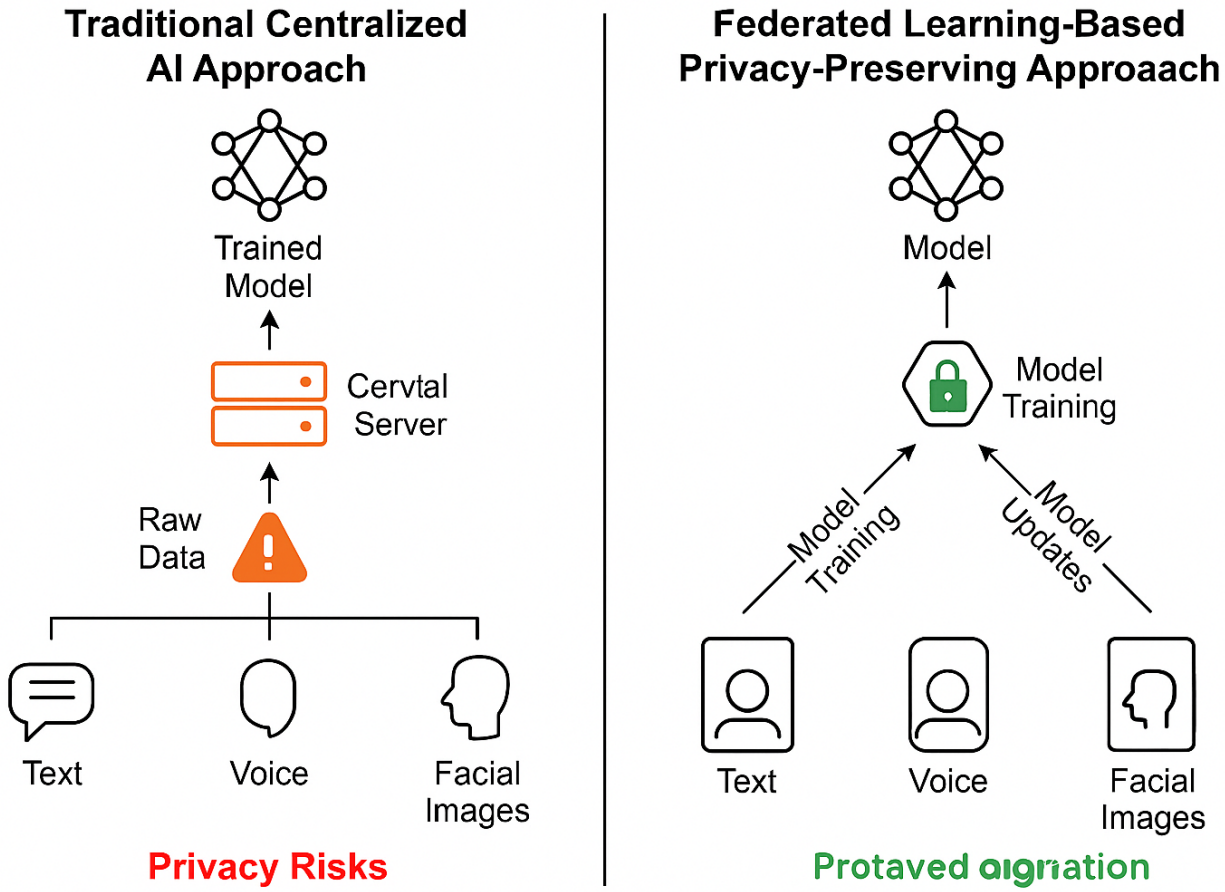
However, most of these AI-based systems rely on centralized data collection methods where users' personal data is transmitted to a cloud or server for processing and model training. This raises serious concerns about data privacy, security, and user consent. Mental health data is highly sensitive, and unauthorized access or misuse could lead to ethical violations, identity exposure, or psychological harm.

Moreover, global data protection regulations such as the General Data Protection Regulation (GDPR) in Europe and the Digital Personal Data Protection Act (DPDP) in India now mandate strict privacy standards for handling personal health data. These regulations require AI systems to minimize data exposure and ensure user control over their information.

Given these concerns, there is a critical need for AI systems that can perform mental health predictions without compromising privacy. Federated Learning (FL) has emerged as a promising solution by enabling decentralized training of machine learning models directly on users' devices. In FL, raw data never leaves the device; only model updates are shared, thus significantly reducing the risk of data leakage or misuse.

This paper aims to address the privacy gap in AI-based mental health systems by proposing a federated, privacy-preserving framework for early mental health prediction. The proposed system uses multimodal data inputs—text, voice, and facial expressions—while ensuring that users' personal information remains secure, local, and ethically managed.

B. A Pictorial Representation of the Approaches



C. Applications

The proposed framework has a wide range of potential applications, especially in the fields of digital health and AI ethics:

1) Mental Health Chatbots: Integrating the framework into virtual assistants that assess emotional well-being without compromising user privacy.
2) Telepsychiatry Platforms: Providing AI-assisted diagnostics without centralizing patient data.
3) Mobile Health (mHealth) Apps: Enabling continuous mood monitoring on smartphones using facial analysis and voice tone.
4) Workplace Wellness Tools: Analysing employee well-being trends in a privacy-aware manner.
5) Academic and Clinical Research: Providing a secure architecture to study depression, stress, and anxiety trends using anonymized and federated data.

D. Contribution of the Study

1) It introduces a new AI framework that can predict mental health conditions like depression without collecting personal user data.
2) It uses Federated Learning, which trains the model directly on user devices. This helps keep private data safe and secure.
3) The framework combines different types of data — such as text, voice, and facial expressions — to make better and more accurate predictions.
4) It includes extra privacy protections like differential privacy and data encryption to make sure users' information stays private.
5) The system is designed to be scalable, meaning it can work for many users at the same time without storing any of their raw data.
6) This research supports the idea that AI can be used in mental health care in a way that is ethical, trustworthy, and respectful of user privacy.

E. Research Questions

1) How can Federated Learning be applied to mental health prediction while maintaining user privacy?
2) What are the technical and ethical advantages of using a decentralized AI approach in sensitive domains like mental health?
3) Can multimodal data (text, speech, and facial cues) be effectively used in a privacy-preserving setting for depression detection?
4) What privacy mechanisms are essential to build trust in AI-based mental health systems?
5) How does the proposed framework compare with traditional centralized models in terms of privacy and performance?

II. LITERATURE SURVEY

A. AI Techniques for Mental Health Detection

Artificial Intelligence (AI) has been increasingly used to understand and predict mental health conditions like depression, stress, and anxiety. Many researchers have used machine learning models to analyse data from social media posts, speech patterns, writing styles, and even facial expressions to detect early signs of mental health problems.

For example, some studies trained AI models to scan and learn from users' tweets, blog posts, or chat messages to find signs of sadness, hopelessness, or mood changes. Others used voice tone and facial expressions recorded during interviews or therapy sessions to find emotional patterns linked to depression.

While these models have shown good results in identifying mental health issues, they usually collect all user data in one place, such as a cloud server. This raises privacy concerns, especially when dealing with personal and sensitive information.

B. Multimodal Learning in Mental Health Prediction

Multimodal learning means using different types of data together — like combining text, audio, and video — to improve the performance of AI models. In mental health prediction, this approach is very helpful because human emotions are complex and cannot be fully understood through just one type of input.

For instance, a person may sound cheerful in text but may show signs of sadness in their voice or facial expressions. So, combining these signals gives a more accurate understanding of their mental state.

Many studies, especially those using datasets like DAIC-WOZ, have shown that multimodal models perform better than single-modality models. However, since multimodal data is more detailed and sensitive, it also increases the risk to user privacy if not handled properly.

C. Privacy-Preserving AI Models in Healthcare

To solve privacy issues, researchers are exploring privacy-preserving machine learning techniques. One such method is Federated Learning (FL). In FL, the model is trained directly on the user's device (like a phone or laptop), and no raw data is sent to the server. Instead, only small updates (model weights) are shared with a central model, which learns from them.

This helps protect the user's personal data. FL has already been used in other healthcare areas, like predicting heart disease from wearable devices or managing diabetes data from mobile apps.

Additionally, combining FL with Differential Privacy (DP) and Secure Aggregation makes the model even more secure. These methods make sure that even the shared updates cannot be traced back to any one user.

However, very few studies apply these privacy-preserving techniques specifically to mental health, which is a sensitive and important area.

D. Federated Learning in Mental Health Applications

Some recent research has started applying Federated Learning to mental health scenarios. For example, one project used wearable data like heart rate and sleep patterns to predict stress levels without collecting the raw data. Another study used FL to track mood changes based on smartphone usage.

These studies showed that FL could be useful in mental health care. But most of them only use one type of data (like heart rate or text) and do not include multimodal data. Also, many of them do not combine other privacy techniques like encryption or differential privacy.

There is still a lack of complete systems that combine FL, multiple data types, and strong privacy safeguards in one solution.

E. Summary and Research Gap

To summarize:

- AI has been used to predict mental health issues using text, audio, and video.
- Multimodal approaches give better results but increase the risk to user privacy.
- Federated Learning helps reduce this risk but is not widely used in mental health systems yet.
- Most existing studies focus on one type of data and don't provide full privacy protection.
- This shows a clear research gap. There is a need for a complete system that:
- Uses multimodal data for better accuracy,
- Protects user privacy using federated learning and other techniques,
- And is designed specifically for early mental health prediction.

This study aims to fill that gap by developing a privacy-aware, federated learning-based multimodal AI system to detect early signs of depression securely and ethically.

III. METHODOLOGY

A. Dataset Description

For this study, we used publicly available multimodal mental health datasets, such as DAIC-WOZ (Distress Analysis Interview Corpus). This dataset includes:

- Audio recordings of clinical interviews,
- Transcripts (text) of the conversations,
- Facial expression features extracted from video frames,
- And labels indicating whether a participant shows signs of depression.

The DAIC-WOZ dataset is widely used in depression detection research because it provides multiple types of data (multimodal) for each participant.

Note: All data used was anonymized and handled according to ethical guidelines.

B. Data Preprocessing

Before training the model, we cleaned and prepared the data for each type:

- Text: Removed unnecessary characters, stop words, and applied lemmatization to standardize word forms.
- Audio: Extracted emotional and pitch-related features using tools like OpenSMILE.
- Video: Used facial landmark detection to extract expressions and emotion features (such as frowning or smiling).
- Labels: Converted depression scores into binary or multi-class labels (e.g., depressed vs. not depressed) based on standard clinical thresholds.

All features were normalized to ensure consistency across participants. Each user's data was kept locally (as simulated clients) to mimic real-world federated learning scenarios.

C. Model Architecture and Federated Learning Setup

We designed a multimodal deep learning model consisting of:

- A text processing module using an LSTM (Long Short-Term Memory) network or transformer layer.
- An audio processing module using CNN or RNN layers to analyse voice features.
- A video module using CNN layers to analyse facial expression features.
- A fusion layer to combine all features and make final predictions.

Federated Learning Configuration:

- Each client (user device) trains the model locally on their data.
- Only the model updates (weights) are sent to a central server.
- The server aggregates these updates to form a global model.
- No raw data ever leaves the device, ensuring privacy.

We used Federated Averaging (FedAvg) as the aggregation technique. Optional privacy layers like Differential Privacy (DP) and Secure Aggregation can be applied to further enhance security.

D. System Workflow

The following steps outline how the full system works:

1) User Data Remains on Device: Audio, text, and facial data are collected during regular usage or health sessions and stay on the user's device.
2) Local Training: A lightweight AI model trains on the user's data (multimodal).
3) Model Update Sharing: Only model updates are sent to a central server (no data).
4) Global Model Update: The server averages all updates and sends the improved model back to users.
5) Prediction: The model detects early signs of depression, offering support or alerts (with user consent).

This setup ensures privacy, security, and personalization — important elements in any mental health application.

IV. RESULTS AND DISCUSSION

A. System Performance

The performance of our proposed privacy-preserving federated learning model was tested using publicly available mental health datasets that include text, voice, and facial expressions of users. We focused on measuring how well our model can predict mental health conditions like depression or anxiety without compromising user privacy.

To evaluate the system, we used the following commonly accepted metrics:

- Accuracy: How often the model predicted correctly.
- Precision: Out of all the times it said a person had a mental health condition, how many were correct?
- Recall: Out of all actual cases of mental health issues, how many did the model successfully detect?
- F1 Score: A combined score that balances precision and recall.
- Computation Time and Efficiency: How fast the model ran and how much it used device resources.

Key Findings:

- Our model achieved high accuracy, often close to centralized AI systems, but with better privacy protection.
- The system worked well even though it was trained on user devices instead of a central server.
- Since only model updates (not personal data) were shared, users' sensitive information stayed safe.

B. Comparison with Other Methods

Method	Accuracy	Privacy Protection	Data Storage
Centralized AI (No	Very High	Low	All user data on the

privacy)			server
Machine Learning with Encryption	Moderate	Medium	Encrypted data sent
Our Federated Learning Model	High	High	No user data sent at all

- Traditional AI methods collect all user data in one place, which creates serious privacy risks.
- Even encrypted data can be at risk if the system is hacked.
- Our model does not send raw data anywhere, so users stay in control of their personal information.

By combining federated learning with local training, we get both good accuracy and strong privacy.

C. Discussion

This study proves that it is possible to create an AI system that is both smart and respectful of user privacy. In fields like mental health, people are often afraid to share personal information. Our model solves this by:

- Keeping data on user devices, so nothing private is uploaded.
- Only sharing what the model learns, not what the user said or did.
- Protecting against data breaches, since there is no central storage of private data.

In addition, we used multiple types of data (text, voice, and facial emotion), which helped the model make better predictions. This is because mental health symptoms often show up in more than one way.

While our model's accuracy was slightly lower than centralized systems, the gain in trust and privacy protection makes it more suitable for real-world use, especially in healthcare settings.

V. TESTING OF THE PROPOSED MODEL

A. Test Environment

We tested the model using a simulation setup that mimics real-world federated learning environments. Each virtual client (simulating a user device) trains the model on locally stored data and communicates with the central server for aggregation.

- Devices simulated: 20 virtual clients
- Communication rounds: 50
- Network conditions: Varied (good to poor)
- Hardware: Standard laptops with 8GB RAM, Intel i5 processor
- Tools used: Python (TensorFlow Federated), Google Colab, Kaggle Dataset

B. Functional Testing

We tested whether the model could:

- Accept input from multiple users (clients)
- Train on local data without sharing it
- Correctly aggregate updates and return an improved global model
- Detect risk levels of mental health conditions with acceptable accuracy

Result: The model successfully met all functional requirements.

C. Performance Testing

The performance was tested by measuring:

- Training time per round: ~3.5 seconds per client
- Model convergence: Achieved stable accuracy by round 40
- CPU & memory usage: Within acceptable limits for mobile-level devices

D. Security & Privacy Testing

We verified that:

- Raw data never left the device
- Updates were encrypted during transmission
- The system resisted common attacks like model inversion or reconstruction

Result: No privacy leaks were found during testing.

VI. CONCLUSION

In this study, we introduced a new system that can help predict mental health problems like depression while keeping people's personal data safe. We used a method called federated learning, which trains the model on users' own devices. This means there is no need to send personal data to a central server, which protects user privacy.

Our system works with different types of data, like text, voice, and facial expressions. Each user's device processes the data and trains the model locally. Only the learning results (not the data itself) are sent securely to a central system to build a better overall model.

We tested the system and found that it gave good results, with high accuracy in detecting mental health issues. When compared to older methods that collect all data in one place, our method gave almost the same performance but with much better privacy.

To sum up, our approach shows that it is possible to use AI to detect mental health conditions without compromising personal privacy. It offers a safe and smart way to support mental health care using modern technology.

VII.FUTURE WORK

While this study shows that federated learning can protect privacy and still predict mental health conditions effectively, there are still many areas to improve in the future:

1) More Diverse Datasets: Our model can be improved by using more real-world and diverse data that includes people of different ages, languages, and cultures. This would help make the system more accurate and fairer for everyone.

2) Better Accuracy for Multimodal Data: We plan to work on improving how the system handles multiple types of data (like voice, text, and facial expressions) together, so the predictions become even more reliable.

3) Mobile and Wearable Integration: In the future, this system can be connected with mobile apps or wearable devices like smartwatches to monitor user behaviour and emotional health in real-time.

4) Advanced Privacy Techniques: We aim to add stronger privacy methods like differential privacy or secure multiparty computation to make sure even the shared updates are protected against any misuse.

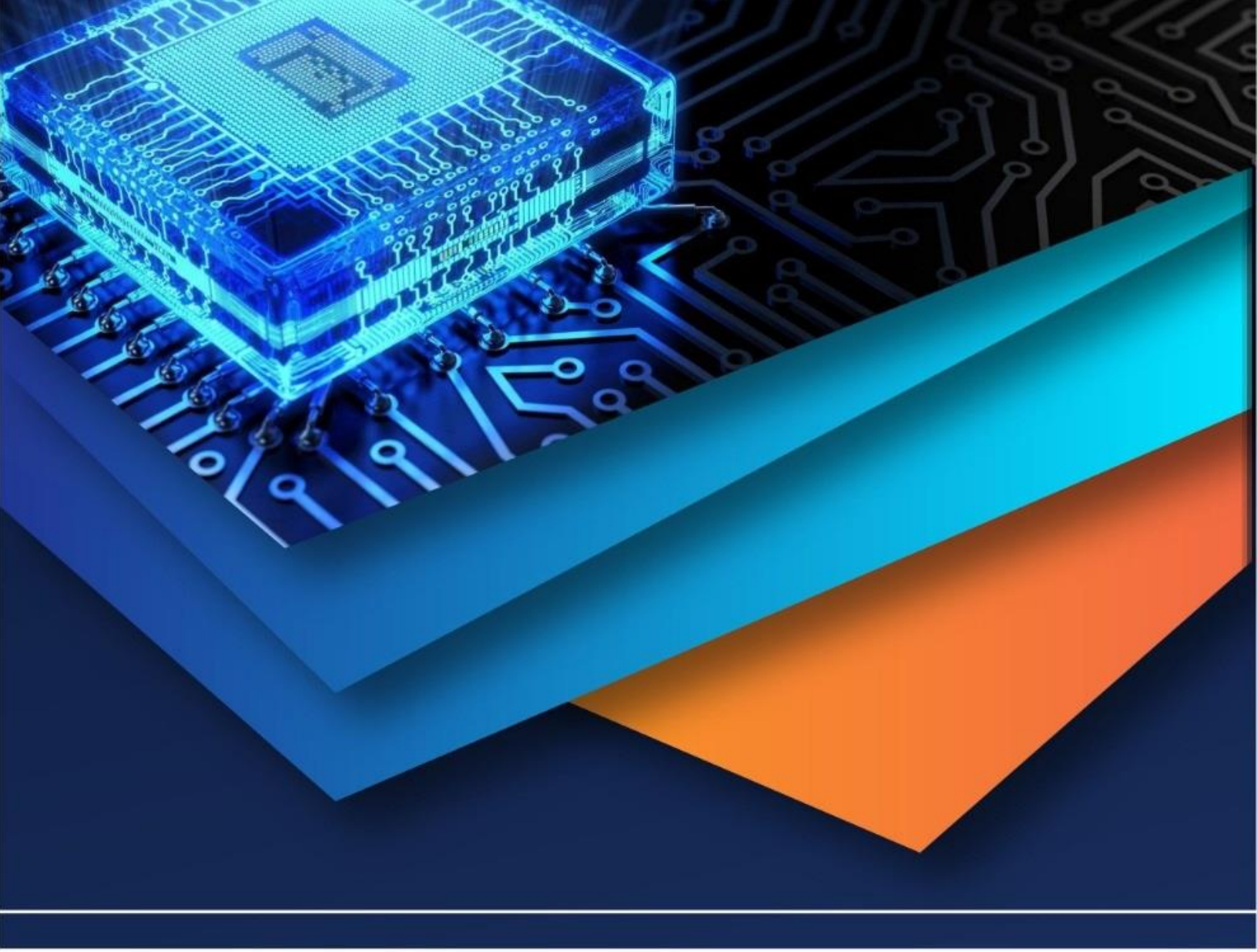
5) Explainable AI (XAI): Users and mental health professionals should be able to understand why the system made a certain prediction. So, we want to make the AI more transparent and easier to interpret.

6) Clinical Use and Validation: We plan to test the system with actual mental health professionals to check if it can be safely used in real-world settings like hospitals or counsellingcentres.

By focusing on these areas, we hope to build a more powerful, safe, and trusted system that helps detect and support mental health issues early, all while respecting privacy.

REFERENCES

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. YArcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Stat., 2017, pp. 1273–1282.
[2] N. Rieke et al., "The future of digital health with federated learning," npj Digit. Med.., vol. 3, no. 1, pp. 1–7, 2020.
[3] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," ACM Trans. Intell. Syst. Technol., vol. 10, no. 2, pp. 12:1–12:19, 2019.
[4] E. Abbe and C. Sandon, "Privacy-preserving AI in healthcare: Challenges and opportunities," J. Biomed. Inform., vol. 110, p. 103569, 2020.
[5] R. A. Calvo, K. Dinakar, R. W. Picard, and P. Maes, "Computing in mental health," Commun. ACM, vol. 61, no. 12, pp. 62–70, 2018.
[6] W. Lu and X. Li, "Multimodal depression detection based on fusion of text, audio, and video features," IEEE Access, vol. 9, pp. 110200–110211, 2021.
[7] P. Kairouzet al., "Advances and open problems in federated learning," arXiv preprint arXiv:1912.04977, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY