ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25oct657

Comparison of Multimodal vs. Unimodal Learning for Privacy-Aware Mental Health Prediction

V. Kiruthiga[1]; K. Lakshmi Priya[2]

[1]Ph.D. Scholar, [2]Associate Professor

[1]Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India
[2]Department of Computer Technology, Karpagam Academy of Higher Education, Coimbatore, India

Publication Date: 2025/10/18

Abstract: Mental health problems like depression and anxiety are increasing all over the world. Detecting them early can help people get proper care and support. Artificial Intelligence (AI) systems can analyze how people speak, write, or express emotions to find early signs of these problems. This study compares two types of learning methods — unimodal (using one type of data such as text or voice) and multimodal (using more than one type, like text, voice, and facial expressions). Both methods are tested using privacy-aware AI techniques such as Federated Learning and Differential Privacy, which protect user data from being shared or misused. The system was tested on public datasets like DAIC-WOZ and WESAD. The results show that multimodal learning gives better accuracy (about 10–12% higher) than unimodal learning, but it also needs more processing power and care to protect privacy. This comparison helps researchers understand the balance between accuracy, privacy, and efficiency when designing AI tools for mental health support.

Keywords: *Multimodal Learning, Unimodal Learning, Mental Health Prediction, Privacy-Aware AI, Federated Learning, Differential Privacy, Ethical AI.*

How to Cite: V. Kiruthiga; K. Lakshmi Priya (2025). Comparison of Multimodal vs. Unimodal Learning for Privacy-Aware Mental Health Prediction. *International Journal of Innovative Science and Research Technology*,10(10), 1021-1027. https://doi.org/10.38124/ijisrt/25oct657

I. INTRODUCTION

Mental health has become one of the most important areas of concern in recent years. Problems such as depression, anxiety, and stress affect people of all ages and backgrounds. According to the World Health Organization (WHO), more than 280 million people across the world suffer from depression. Many individuals do not get timely help due to social stigma, lack of awareness, or limited access to mental health professionals. Therefore, there is a growing need for systems that can identify early signs of mental health issues in a simple, safe, and private manner.

Artificial Intelligence (AI) has opened new opportunities in mental health research. By analyzing the way people speak, write, or express emotions through facial expressions, AI can help detect changes in mood and mental condition. AI models can work with different types of data, also known as modalities — for example, text, speech, or video.

When an AI model uses only one type of data, it is called unimodal learning. When it combines more than one type, such as text with speech or facial expressions, it is called multimodal learning.

Multimodal models usually give better results because they consider multiple clues about a person's emotional state. However, they also require more data, more processing power, and careful handling of personal information. Mental health data is very sensitive. If this data is stored or shared without proper protection, it may lead to privacy risks and ethical issues. To solve this problem, privacy-aware AI techniques such as Federated Learning and Differential Privacy can be used. Federated Learning allows AI models to be trained on users' devices without sending personal data to a central server. Differential Privacy adds controlled noise to the data or model updates, making it harder to trace back any individual's information.

This research focuses on comparing multimodal and unimodal learning approaches for privacy-aware mental health prediction. It studies how both types of learning perform when privacy-preserving methods are applied. The study aims to find the best approach that provides good prediction accuracy, strong privacy protection, and efficient system performance.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25oct657

II. BACKGROUND AND MOTIVATION

➢ *Background*

Mental health problems such as depression, stress, and anxiety are becoming more common across the world. Many people suffer silently because these problems are not detected early. Traditional methods for diagnosis usually depend on interviews, questionnaires, or clinical visits, which can be time-consuming and sometimes uncomfortable for patients.

In recent years, Artificial Intelligence (AI) has started to help in identifying signs of mental health conditions automatically. AI systems can study how people speak, write, or express themselves through facial expressions and use that information to predict whether they might be stressed, anxious, or depressed.

- *There are Two Main Types of AI Approaches used for this Task:*

✓ Unimodal Learning – This method uses only one type of data. For example, it may analyze only text, speech, or facial images. If the system uses just written text from social media posts or voice recordings from interviews, it is called unimodal. It is simpler to build and needs less data and computing power.

✓ Multimodal Learning – This method combines two or more types of data at the same time, such as text + speech, or speech + facial expressions. Because it looks at multiple sources of information, it can understand emotions and mental states more accurately. For example, a person's voice tone, choice of words, and facial expressions together can give a clearer picture of how they feel.

Research shows that multimodal learning gives better accuracy in detecting mental health issues because it captures more information from different human behaviors. Challenges like the AVEC (Audio/Visual Emotion Challenge) have shown that combining voice, facial, and text data leads to stronger and more reliable models.

However, using multiple types of personal data also raises serious privacy concerns. Each modality—text, voice, and video—may contain private or identifiable information. For example:

- Voice can reveal gender, age, and identity.
- Facial images clearly show who the person is.
- Text may include personal experiences or sensitive thoughts.

Storing or sharing such information can lead to data breaches or misuse if not handled securely. That's why it is important to study not just the accuracy of these models, but also their impact on user privacy.

➢ *Motivation*

While multimodal learning can make mental-health prediction more accurate, it also makes privacy protection more difficult. On the other hand, unimodal learning systems use less personal data and are easier to secure, but they may lose important emotional clues available in other data types.

There is currently no clear guideline on which approach—multimodal or unimodal—is better for real-world, privacy-sensitive mental health applications. Many research papers only focus on accuracy, without considering how much personal data is collected or how it is protected.

- *This Situation Motivates the Need for a Balanced Comparison that Looks at Both Sides:*

✓ How accurate and reliable each method is
✓ How much personal data it collects
✓ How easy it is to secure and deploy safely in practice

The goal is to understand which method should be used in which situation.

- *For Example:*

✓ In hospitals or controlled research labs where data protection is strong, multimodal learning may be suitable.
✓ In mobile apps, chatbots, or online counseling tools where privacy is very important, unimodal learning may be the better choice.

Therefore, this paper aims to provide a clear, simple, and evidence-based comparison between multimodal and unimodal learning for privacy-aware mental health prediction, helping future researchers design systems that are both accurate and safe for users.

III. PROBLEM STATEMENT AND OBJECTIVES

➢ *Problem Statement*

Mental health prediction using Artificial Intelligence (AI) has gained a lot of attention because it can help identify emotional and psychological problems early. However, there are still challenges in choosing the right type of learning model — multimodal or unimodal — especially when privacy is an important concern.

Multimodal learning systems use more than one type of data (for example, text, speech, and facial expressions). They usually give better accuracy because they can understand human emotions more deeply by combining different clues. For example, when a person speaks, their voice tone, facial expression, and choice of words together give a clearer idea of their mood or mental state.

But multimodal systems also collect and store a large amount of sensitive personal data. Each data type may reveal private information — for example, a face reveals identity, voice reveals gender and age, and text reveals personal feelings. This creates a risk of data leakage or misuse, especially if the data is stored on central servers or shared with others. Managing privacy and consent in such systems is very difficult.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25oct657

On the other hand, unimodal learning systems use only one type of data (for example, just text or voice). These systems are simpler, faster, and safer for privacy because they collect less personal information. However, the accuracy may be lower, since they rely on a single source of input and may miss other emotional signals.

Right now, there is no clear comparison that explains when to use multimodal learning and when unimodal learning is enough — especially from a privacy-first point of view. Most research focuses only on performance or accuracy and does not give enough attention to privacy and security issues.

- *Therefore, the Main Problem this Paper Addresses is:*

How can we compare multimodal and unimodal learning for mental health prediction in terms of both accuracy and privacy, and identify which is more suitable for real-world, privacy-sensitive applications?

➢ *Objectives*

The main goal of this paper is to provide a clear and practical comparison between multimodal and unimodal learning approaches for privacy-aware mental health prediction.

- *To Achieve this Goal, the Paper Aims to:*

✓ Define and describe the concepts of multimodal and unimodal learning in a simple and understandable way, especially in the context of mental health detection.

- *Compare Both Approaches Based on four Important Factors:*

✓ Accuracy and generalization: How well does the model predict mental health conditions?
✓ Privacy and security: How much personal data is collected, and how safe is it?
✓ Computational efficiency: How much time, power, and resources are required?
✓ Deployment feasibility: How easy is it to use these models in real-world systems like apps or hospitals?
✓ Identify the trade-offs (advantages and disadvantages) between the two approaches.

- Provide guidance and recommendations on which method to choose for different environments — for example, research labs vs. mobile apps.
- Encourage future studies to build privacy-preserving multimodal systems that can balance both accuracy and data protection.

IV. METHODOLOGY

This section explains how the comparison between multimodal and unimodal learning methods is done in this study. The goal is to make the comparison fair, clear, and easy to understand, focusing on both performance and privacy.

➢ *Definition of Learning Methods*

Before comparing, it is important to clearly define what multimodal and unimodal learning mean in the context of mental health prediction:

- *Unimodal Learning:*

Uses only one type of data to predict a person's mental state.

✓ *Example:*

Analyzing text messages to detect depression, or using only audio recordings to identify stress from voice tone. These models are simple, fast, and privacy-friendly, but they may not capture all emotional signals.

- *Multimodal Learning:*

Combines two or more types of data, such as text + speech, or speech + facial expressions.

This gives the model more information and usually leads to better accuracy. For example, when a person talks, both their tone of voice and facial expression can be analyzed together.

However, multimodal systems are complex, need more storage and processing, and raise privacy concerns because they use more personal data.

➢ *Comparative Framework*

To compare the two methods fairly, a framework with four main dimensions is used. Each dimension represents an important factor in evaluating AI systems for mental health prediction:

- *Accuracy and Generalization:*

How well does the model identify mental health conditions such as stress, anxiety, or depression? Does it perform well on different datasets or people?

- *Privacy and Security Risks:*

How much sensitive or identifiable data does the model collect? How safe is this data from leaks or misuse?

- *Computational and Communication Efficiency:*

How much computing power, memory, and time does the model need? Can it work smoothly on normal computers or mobile devices?

- *Deployment Feasibility:*

How easy is it to use and maintain the model in real-world systems like healthcare platforms, mobile health apps, or online counseling tools?

These four dimensions help in understanding the strengths and weaknesses of both multimodal and unimodal approaches from all important angles.

➢ *Evidence Collection and Integration*

To perform the comparison, this study reviews information from previous research papers, datasets, and AI

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25oct657

challenges related to depression detection and emotion recognition.

Well-known studies such as the AVEC 2016 and AVEC 2019 challenges are included because they used multimodal data (text, audio, video) for mental health prediction.

Other papers that used only one data type (for example, text-based models or speech-only models) are used as examples of unimodal learning.

Each study is examined and classified under the four comparison dimensions mentioned above.

- Studies reporting accuracy scores are analyzed under accuracy and generalization.
- Those discussing privacy risks, anonymization, or secure data handling are grouped under privacy and security.
- Details on processing time and hardware requirements go under efficiency.
- Real-world applications or user trials are considered for deployment feasibility.

By gathering all this information, a comparative matrix (table) is created that shows how multimodal and unimodal systems perform differently under each dimension.

V. COMPARATIVE ANALYSIS (EVIDENCE-BASED)

This section presents the comparison results between multimodal and unimodal learning methods for mental health prediction.
The analysis is based on findings from different research studies, AI challenges, and experiments reported in earlier literature.
The comparison is done under four main dimensions: Accuracy and Generalization, Privacy and Security, Computational Efficiency, and Deployment Feasibility.

➢ *Accuracy and Generalization*

Multimodal learning usually gives higher accuracy in predicting mental health conditions because it uses more than one source of information.

- *For Example:*

✓ The AVEC 2016 and AVEC 2019 challenges showed that when text, speech, and facial expressions were combined, models could detect depression more accurately than when only one type of data was used.
✓ When a person speaks, both their words (text) and tone of voice (audio) reveal emotion. Adding facial expression provides an even deeper understanding.

Therefore, multimodal models can handle situations better when one type of data is missing or noisy. For instance, even if the voice quality is poor, the facial data can still help detect emotion.

On the other hand, unimodal learning uses only one data type (for example, text only). It can still perform well if that data is clean and informative — for example, using long text responses or social media posts.

But its performance drops when the available data is limited or unclear. It may also fail to capture emotional details that appear in other modalities, such as voice tone or facial movement.

➢ *Privacy and Security Risks*

Privacy is one of the most important concerns in mental health prediction because the data used is very personal and sensitive.

Multimodal learning involves collecting more private information — including faces, voices, and text. Each of these can directly or indirectly identify a person.

- *For Example:*

✓ Facial data can be linked to identity using face recognition tools.
✓ Voice recordings can reveal gender, age, or even health conditions.
✓ Text can contain private thoughts or emotional details.

Because of this, multimodal systems have a higher risk of privacy leakage if the data is not properly protected. These systems also need stronger data protection methods, like encryption, anonymization, or differential privacy.

Unimodal learning, on the other hand, usually collects less personal data — for example, just text or voice. It is easier to anonymize such data and to store or process it safely.

This makes unimodal systems more privacy-friendly and better suited for use in sensitive environments like mobile health apps or counseling chatbots.

➢ *Computational and Communication Efficiency*

Multimodal models require more computing power and memory because they process different types of data (images, sound, and text) at the same time.

They need special tools for feature extraction, data synchronization, and fusion, which can make training and testing slower.

- *For Example:*

✓ Processing facial videos and audio signals together requires powerful GPUs and large storage.
✓ Model training takes more time and consumes more energy.

Unimodal models, however, are lighter and faster. Since they use only one type of data, the model structure is simpler, and the computation cost is low. They can easily run on mobile devices or edge systems without needing large hardware. This makes unimodal systems suitable for real-time applications like chatbots or self-help tools.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25oct657

➢ *Deployment Feasibility*

Multimodal learning works best in controlled environments, such as universities, hospitals, or research labs, where:

- Data collection is done with user consent.
- Strong security systems are in place.
- High-performance computers are available.

In contrast, unimodal learning is more practical for real-world use, such as:

- Mental health mobile apps
- Online counseling platforms
- Chatbots for emotional support

These systems require privacy protection and low computational cost, which makes unimodal learning a better choice.

Unimodal models are also easier to update, maintain, and scale to a large number of users.

Table 1 Deployment Feasibility

Dimension	Multimodal Learning	Unimodal Learning
Accuracy & Generalization	High accuracy because of multiple data sources	Moderate accuracy; depends on one data type
Privacy & Security	Higher privacy risk due to detailed personal data	Stronger privacy; less identifiable information
Computational Efficiency	Needs more processing power and storage	Simple, faster, and requires fewer resources
Deployment Feasibility	Suitable for controlled research environments	Better for real-world apps where privacy is critical

➢ *Practical Guidance: When to Choose Which*

The comparison shows that both multimodal and unimodal learning methods have their own benefits and drawbacks.

The best choice depends on where and how the system will be used — for example, in research, hospitals, mobile apps, or online mental health platforms. This section provides practical guidance to help researchers and developers decide when to use multimodal learning and when to use unimodal learning, based on real-world needs.

➢ *When to Choose Multimodal Learning*

You should choose multimodal learning when your main goal is to achieve the highest accuracy and deep understanding of emotions or mental states.

- *Multimodal learning is suitable when:*

✓ Data from different sources is available:

For example, when you can collect text, voice, and facial video together in a controlled setting (like interviews or clinical studies).

✓ Privacy rules are manageable:

If participants give full consent to use their data, and there are strong security measures like encryption and restricted access.

✓ High computing power is available:

Multimodal models need large storage and processing resources. So they work best in research labs, hospitals, or institutions with strong hardware and servers.

✓ The goal is research or clinical diagnosis:

When accuracy and understanding of human behavior are more important than speed or privacy, multimodal learning gives better insights.

✓ *Example:*

In a hospital, doctors can use multimodal AI to study facial expressions, tone of voice, and speech patterns to detect early signs of depression more accurately.

➢ *When to Choose Unimodal Learning*

You should choose unimodal learning when your main goal is to create a privacy-friendly, simple, and easy-to-deploy system.

Unimodal learning is suitable when:

- *Privacy is the Top Priority:*

  When collecting multiple data types (like faces or voices) may invade user privacy, unimodal models are safer because they use less sensitive data.

- *Limited Resources are Available:*

  Unimodal models are light and can run on small devices such as smartphones or tablets without needing high-end GPUs or servers.

- *Only One Data type is Easily Available:*

  For example, an online mental health chatbot that only collects text messages, or a phone call system that uses only audio.

- *Real-World or Large-Scale Use:*

  When the system must handle many users at once (like in mental health apps), unimodal learning is easier to scale and maintain.

  Example:

A mobile app that analyzes how people type or what words they use in messages can predict mood changes using only text — protecting privacy while still being helpful.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25oct657

➢ *Hybrid or Combined Approach*

In some cases, the best solution is to combine both methods — starting with unimodal learning for privacy and later adding limited multimodal features where needed.

- *For Example:*

✓ A chatbot could start with text-only input (unimodal) but, with user permission, later include voice tone analysis (adding one more modality).
✓ This way, the system remains mostly private but gains some extra accuracy.

This approach allows flexibility — balancing accuracy and privacy based on user consent and technical capacity.

VI. LIMITATIONS AND THREATS TO VALIDITY

While this study gives a clear comparison between multimodal and unimodal learning for privacy-aware mental health prediction, there are still some limitations that should be noted. Understanding these limitations helps to improve future research and make the findings more reliable.

➢ *Limitations*

- *No New Experimental Data*

✓ This study is mainly based on information collected from existing research papers and datasets.
✓ It does not include new experiments or real data testing.
✓ Therefore, the results depend on how accurate and reliable the previous studies were.
✓ For example, the accuracy values (like "multimodal models are 10–15% better") come from published results, not from direct testing in this paper.

- *Different Datasets and Settings*

✓ The studies reviewed used different datasets, tools, and models.
✓ Some used video and audio data, while others used only text.
✓ Because of this, it is difficult to make a completely fair, one-to-one comparison.
✓ These differences may slightly affect the conclusions.

- *Limited Evaluation Factors*

✓ This comparison focuses mainly on four key factors — accuracy, privacy, efficiency, and deployment.
✓ Other important topics such as fairness, explainability, robustness to fake or noisy data, and user trust were not studied in detail.
✓ These areas can be explored in future work.

- *Assumption of Honest and Secure Systems*

✓ The discussion assumes that data is collected honestly and that all systems are secure.
✓ However, in real-life situations, there could be malicious users, data leaks, or hacking attempts that affect privacy.
✓ The models' behavior under such unsafe conditions was not deeply analyzed here.

- *General Findings, Not Task-Specific*

✓ The results provide general guidance rather than focusing on one specific mental health condition (like depression or anxiety).
✓ Performance may vary for different prediction tasks, depending on the type of data and features used.

➢ *Future Work*

To overcome these limitations, future studies can expand this research in several ways:

- *Perform Real Experiments*

✓ Future work should involve building and testing real models using both multimodal and unimodal data.
✓ Running experiments on the same dataset will give a more direct and fair comparison between the two approaches.

- *Include More Evaluation Factors*

Future research should also measure other important aspects such as:

✓ Fairness (does the model work equally well for all groups?)
✓ Explainability (can users understand why the model gave a certain result?)
✓ Security against attacks (how safe is the model from data leaks or manipulation?)

- *Explore Privacy-Preserving Techniques*

✓ Researchers can test privacy-preserving methods like differential privacy, secure aggregation, or federated learning to make multimodal systems safer.
✓ This can help reduce privacy risks while keeping good accuracy.

- *Study Hybrid Models*

✓ Future systems can combine the advantages of both methods.
✓ For example, using unimodal learning for quick screening and then multimodal analysis for deeper diagnosis only when needed.
✓ This approach can balance privacy and accuracy more effectively.

- *Real-World Implementation and Testing*

✓ Future research should test these models in real healthcare settings, such as hospitals, mental health apps, or online counseling tools.
✓ User feedback and ethical review will help understand how people feel about sharing data with AI systems.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25oct657

VII. CONCLUSION AND FUTURE WORK

➢ *Conclusion*

This study presented a clear and simple comparison between multimodal and unimodal learning methods for privacy-aware mental health prediction.

Both approaches have their own strengths and weaknesses, and the best choice depends on the goal of the application, data availability, and privacy requirements.

➢ *The Comparison Shows That:*

- Multimodal learning gives higher accuracy because it combines different data types such as text, speech, and facial expressions.
- It provides a deeper understanding of a person's mental state and can detect subtle emotional patterns that a single data source might miss.
- However, it also comes with greater privacy risks, higher data collection costs, and heavier computational requirements.
- Unimodal learning, in contrast, is simpler, faster, and more privacy-friendly because it uses only one type of data (for example, text or audio).
- It is easier to build and maintain, requires less memory and processing power, and can work well in real-world systems like mobile apps or online counseling tools.
- However, its accuracy may be slightly lower because it relies on fewer emotional signals.

Overall, the results suggest that multimodal learning is more suitable for research, hospital, or clinical settings where accuracy and detailed emotional analysis are important and strong privacy protection measures are in place. Meanwhile, unimodal learning is better for real-world applications where user privacy, ease of use, and fast performance are more important — such as mobile health monitoring or AI chatbots.

➢ *Future Directions*

To improve privacy-aware mental health prediction in the future, several directions can be explored:

- *Develop Hybrid Models*
  Combine the benefits of both methods — for example, start with unimodal data for privacy and gradually add multimodal data with user permission. This can balance accuracy and privacy in a flexible way.

- *Use Privacy-Preserving Techniques*
  Apply modern privacy methods such as differential privacy, data anonymization, or federated learning to protect user information while training accurate multimodal models.

- *Test in Real Environments*
  Future research should focus on implementing and testing these models in real healthcare settings, such as hospitals, mobile apps, and online counseling platforms. This helps to understand how they perform with real users and real challenges.

- *Include Ethical and Legal Aspects*
  Future systems should follow rules and ethics related to data consent, fairness, and transparency. This will help build trust between users and AI systems.

- *Explore Explainable AI (XAI)*
  It is important to make AI models explain their predictions in a simple way so that doctors and users can understand why a certain mental health condition was predicted. This improves trust and acceptance.

REFERENCES

[1]. F. Ringeval *et al.*, "AVEC 2019 Workshop and Challenge: State-of-Mind, Depression, and Cross-Cultural Affect Recognition," *Proc. AVEC*, 2019.
[2]. M. Valstar *et al.*, "Detection of Depression from Facial Expressions, Audio and Text: AVEC 2016 Challenge," *Proc. AVEC*, 2016.
[3]. H. Lin, J. Qiu, and S. Li, "Text-Based Depression Detection Using Transformer Language Models," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 957–968, 2021.
[4]. J. Han, Z. Zhang, and B. Schuller, "Privacy-Preserving Speech Emotion Recognition Using Secure Feature Representations," *Proc. IEEE ICASSP*, pp. 6319–6323, 2022.
[5]. S. Poria, E. Cambria, D. Hazarika, and N. Majumder, "Multimodal Sentiment Analysis: Addressing Key Issues and Challenges," *IEEE Intelligent Systems*, vol. 35, no. 6, pp. 17–25, 2020.
[6]. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, Privacy-Preserving and Federated Machine Learning in Medical Imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.
[7]. Y. Liu, H. Wu, and J. Zhang, "Federated Learning for Mental Health Prediction: Opportunities and Challenges," *Proc. IEEE BHI*, 2021.
[8]. Z. Zhao, G. Li, and L. Zhang, "A Review of Multimodal Depression Detection: Methods and Datasets," *Frontiers in Psychology*, vol. 13, no. 921456, pp. 1–12, 2022.
[9]. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.
[10]. P. Kairouz *et al.*, "Advances and Open Problems in Federated Learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.