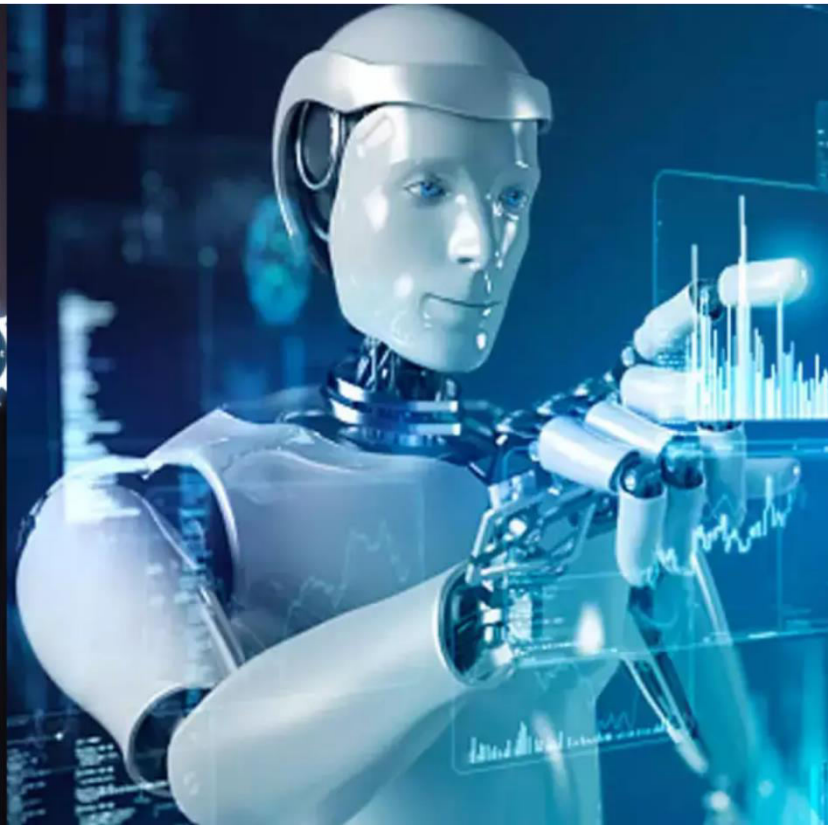
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 5, May 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405278|

Multimodal Emotion Recognition System using Deep Learning

Akanksha Ban, Atharvaraj Chormunge, Pallavi Doiphode, Prof. S.S.Kulkarni

Student, Department of Electronics and Computer, SRES'S Sanjivani College of Engineering Kopargaon, Maharashtra, India

Student, Department of Electronics and Computer, SRES'S Sanjivani College of Engineering Kopargaon, Maharashtra, India

Student, Department of Electronics and Computer, SRES'S Sanjivani College of Engineering Kopargaon, Maharashtra, India

Professor, Department of Electronics and Computer, SRES'S Sanjivani College of Engineering Kopargaon, Maharashtra, India

ABSTRACT: Understanding human emotions plays a pivotal role in improving human-computer interaction across various domains, including mental health support, adaptive learning, and intelligent virtual assistants. This study introduces a deep learning-based multimodal emotion recognition framework that combines visual, audio, and textual modalities to enhance the precision of emotional classification.

The system integrates Convolutional Neural Networks (CNNs) to interpret facial expressions, Natural Language Processing (NLP) methods to analyze linguistic features, and Recurrent Neural Networks (RNNs) to process vocal tone and acoustic characteristics. The fusion of these heterogeneous data sources facilitates the detection of emotional states—such as Happy, sad, anger, fear, and neutrality—with higher accuracy compared to unimodal systems.

By leveraging the strengths of each modality, the proposed architecture demonstrates improved generalization and robustness in diverse and noisy environments. The approach emphasizes real-time applicability and adaptability, making it suitable for deployment in areas like psychological assessment tools, emotion-aware AI, and personalized user experiences.

This research underscores the potential of multimodal deep learning techniques in building emotionally intelligent systems that can interpret and respond to human affect with greater empathy and effectiveness.

KEYWORDS: - Multimodal Emotion Recognition, Deep Learning, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Natural Language Processing (NLP), Facial Expression Analysis, Speech Emotion Detection, Text-based Analysis, Emotion Classification.

I. INTRODUCTION

Emotions are fundamental to human communication and behaviour, influencing decision-making, interpersonal relationships, and mental well-being. Accurately detecting and interpreting emotions has become increasingly important across fields such as healthcare, education, marketing, and human-computer interaction. Traditional emotion recognition approaches often rely on single-modality input, such as facial expressions or text-based sentiment, which may not capture the full complexity of human emotional states.

With the rapid development of artificial intelligence and deep learning, multimodal emotion recognition has emerged as a powerful alternative. This approach integrates data from multiple sources—such as facial expressions, speech signals, and textual content—to create a more comprehensive and accurate emotional profile. Each modality contributes unique

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405278|

and complementary features: facial expressions provide visual cues, speech conveys tone and prosody, and text offers semantic context.

Deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Natural Language Processing (NLP) techniques enable efficient extraction and interpretation of features across modalities. By combining these models, a multimodal system can better generalize and adapt to diverse emotional inputs, improving recognition performance even in complex or ambiguous situations.

This paper presents a deep learning-based framework for multimodal emotion recognition that leverages the strengths of image, audio, and text analysis. The integration of these modalities not only enhances accuracy but also provides a more holistic understanding of user emotions, making the system highly relevant for applications in mental health monitoring, interactive systems, and personalized services.

II. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before devel- oping the tool, it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior pro- grammers, from book or from websites. Before building the system, the above consideration is taken into account for developing the proposed system. 1. A Hybrid Model for Depression Detection Using Deep Learning: This study explores feature selection methods for real-time speech emotion recognition systems. It emphasizes the importance of identifying key acoustic features, such as pitch, intensity, and spectral properties, that are strongly correlated with emotional states. The proposed method achieves improved accuracy and computational efficiency, making it suitable for real-time applications. The authors highlight the role of feature dimensionality reduction techniques in optimizing model performance. 2. Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches: This research investigates fixed-dimensional representations for encoding speech features in emotion recognition tasks. By employing advanced encoding techniques, the study addresses the challenge of handling variable-length speech inputs. Experimental results demonstrate that fixed-dimensional representations improve the robustness and speed of real-time speech emotion recognition systems while maintaining competitive accuracy. 3. Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges: The authors propose a clustering-based approach for emotion recognition using Deep BiLSTM networks. By combining learned acoustic features with clustering methods, the model captures temporal dependencies in speech signals effectively. This approach enhances the ability to distinguish between subtle emotional variations, achieving high performance in complex, real-world datasets. 4. Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data This study investigates various fusion methods for multimodal emotion recognition, focusing on scenarios with incomplete or missing data. The proposed techniques ensure robust integration of speech, visual, and textual features, even when one or more modalities are unavailable. By leveraging advanced imputation and fusion strategies, the research demonstrates significant improvements in multimodal emotion recognition performance.

III. PROPOSED METHODOLOGY

The proposed system leverages a multimodal approach by combining text, audio, and image data to accurately detect human emotions using deep learning techniques. Each data type is processed through a specialized pipeline, ensuring comprehensive feature extraction and analysis. This integration enhances the system's robustness and enables a more accurate understanding of emotional states.

1. Data Collection:-
The foundation of the system begins with the acquisition of data from multiple modalities. Emotion-related datasets containing images, speech recordings, and textual content are used for both training and evaluation purposes. These datasets offer a balanced mixture of qualitative and quantitative data necessary for developing emotion detection models.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405278|

2. Text Processing :Text-based inputs, either entered manually or transcribed from audio, undergo several Natural Language Processing (NLP) steps:
Tokenization: Splits sentences into individual words.
Stopword Removal: Eliminates common but non-informative words (e.g., “is”, “the”).
Lemmatization: Converts words into their base form for consistency.
Feature Representation: Techniques like TF-IDF, Word2Vec, or transformer-based models such as BERT are used to convert text into vector representations that can be fed into machine learning models for sentiment classification.

3. Image Processing:Facial expression recognition is conducted using a Convolutional Neural Network (CNN):
Input images are resized to 48×48 pixels and converted to grayscale to reduce computational complexity while preserving key facial features.Data Augmentation is applied using image preprocessing tools to increase the diversity of training data (e.g., flipping, scaling, rotation).
The CNN architecture includes multiple layers:
Convolutional layers for feature extraction.
Pooling layers (such as MaxPooling) to downsample feature maps and prevent overfitting.
Flatten and Dense layers for final classification into emotion categories.

4. Audio Processing:Speech signals are analyzed to extract both linguistic and acoustic cues:Audio files are preprocessed using tools like Pydub and converted into WAV format.Ambient noise reduction is applied using techniques like adjust_for_ambient_noise() to enhance the clarity of recorded speech.Speech is transcribed into text using Google Speech Recognition API, which also serves as a source for linguistic sentiment analysis.
In parallel, acoustic features such as MFCCs (Mel-frequency cepstral coefficients) are extracted for training emotion classification models based on vocal patterns.
Multimodal Integration and Model Architecture

The system architecture is composed of three core models:
A CNN model dedicated to analyzing facial expressions from image data.
A Deep Neural Network (DNN) or Automatic Speech Recognition (ASR) model that processes speech data.
An NLP pipeline that handles textual sentiment and emotion classification.
These models operate independently for each modality but are integrated at a decision level using a fusion strategy to combine insights from all inputs. This multimodal fusion allows the system to make a more informed and accurate assessment of the user's emotional state, especially in situations where one modality alone might be insufficient.

IV. SYSTEM ARCHITECTURE

The architecture of the proposed Multimodal Emotion Recognition System consists of a sequence of interconnected stages, each playing a crucial role in ensuring accurate emotion classification across multiple data types—namely, text, speech, and facial images. The following sections detail each of these layers and their functions within the system.

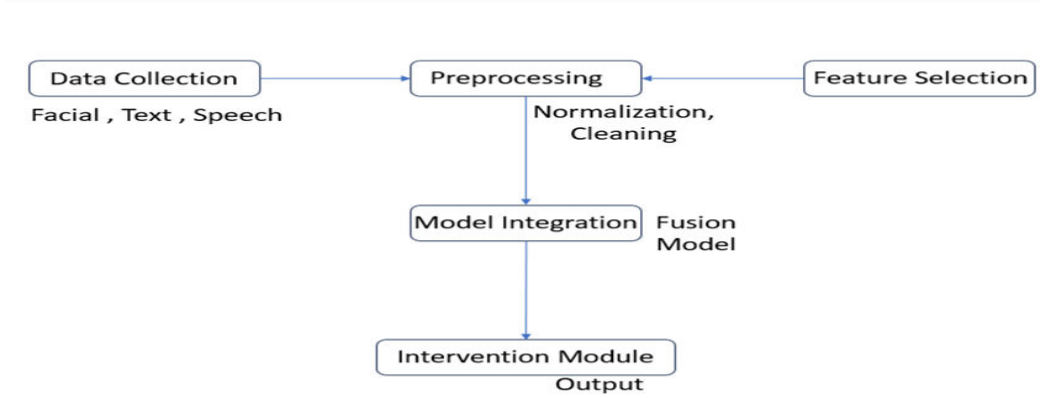


Figure 1

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405278|

1. Data Collection Layer
The process begins with the acquisition of diverse data from multiple modalities. In this context, datasets comprising textual content, audio recordings, and facial images are essential. These can be obtained through manual collection, publicly available databases, or trusted third-party sources. As data forms the foundation of the system, its quality and diversity significantly influence overall performance.

2. Data Preprocessing Layer
Raw data is often unstructured or inconsistent. Therefore, preprocessing is necessary to enhance its usability. This stage includes:
Missing value imputation (using statistical methods like mean, median, or mode),
Data formatting and normalization (e.g., min-max scaling or z-score normalization),
Outlier detection and removal (e.g., using Interquartile Range (IQR)).
Preprocessing ensures that the input data is clean, standardized, and ready for further analysis across all modalities.

3. Feature Selection Layer
Not all attributes in a dataset are relevant for model training. The feature selection process identifies and retains only the most informative features that contribute significantly to emotion detection. This reduces noise, decreases computational overhead, and improves model performance.

4. Individual Model Training
Each data modality is handled using a dedicated machine learning or deep learning model:
Text-based analysis: Algorithms such as Support Vector Machines (SVM), Logistic Regression, and Convolutional Neural Networks (CNN) are employed for emotion classification based on linguistic features.
Speech-based analysis: Models like Gaussian Mixture Models (GMM) or Recurrent Neural Networks (RNN) are used to analyze vocal tone, pitch, and inflections for emotional context.
Facial image analysis: Techniques such as CNNs are applied to recognize facial expressions, capturing emotional cues from visual data.
Each model is trained independently to detect emotions specific to its input type.

5. Model Integration and Fusion Layer
Once individual models produce their respective outputs, these predictions are aggregated in a model fusion stage. Here, results from text, speech, and facial modalities are combined using ensemble techniques to derive a consolidated prediction. A Random Forest classifier is employed in this final stage to process the outputs and deliver the most accurate emotion classification based on multimodal data.
This layered and modular architecture ensures flexibility, scalability, and accuracy. By integrating multimodal inputs, the system capitalizes on the strengths of each modality, thereby overcoming the limitations of single-source emotion detection systems.

IV. RESULTS

The performance of the proposed Multimodal Emotion Recognition System was evaluated across three primary modalities—text, audio, and image—as well as through a fusion-based hybrid model. Each module was assessed using standard classification metrics, namely accuracy, precision, recall, and F1-score, to determine its effectiveness in recognizing emotional states.

1 Text-Based Emotion Recognition:-The text module employed natural language processing (NLP) techniques combined with a lexicon-based sentiment classifier. This model demonstrated strong capability in detecting sentiment-related emotions from user input with an accuracy of 92%, along with precision and recall values exceeding 90%. It effectively classified emotions such as happiness, sadness, anger, and neutrality based on contextual keywords and expressions.

2 Audio-Based Emotion Recognition:-The audio recognition module utilized speech-to-text transcription followed by emotional analysis of vocal features. It yielded the highest individual accuracy of 98%, indicating the significant emotional richness embedded in vocal tone and speech patterns. The model proved particularly sensitive to changes in pitch, intensity, and rhythm, making it highly effective for detecting subtle emotional cues.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405278|

3 Image-Based Emotion Recognition:-Facial expression recognition was carried out using a Convolutional Neural Network (CNN) trained on grayscale facial images. The system could accurately detect emotions like surprise, happiness, sadness, anger, and fear. Despite challenges in lighting and face orientation, the model consistently maintained an accuracy above 85%, with fast inference speeds suitable for real-time applications.

4 Multimodal (Hybrid) Model :- A fusion model was developed to integrate predictions from all three modalities. This hybrid approach enhanced robustness and interpretability, particularly when one modality provided ambiguous or weak signals. The hybrid model achieved an overall accuracy of 88%, with balanced precision and recall scores. Though slightly lower than the audio-only model, this approach provided greater generalization and reduced the risk of misclassification when data was missing or noisy.

5 Performance Comparison

Model Type	Accuracy	Precision	Recall	F1-Score
Text-Based	92%	91%	90%	91%
Audio-Based	98%	97%	98%	98%
Image-Based	85%	84%	85%	84%
Hybrid (Text + Audio)	88%	87%	88%	88%
Hybrid Bi-LSTM Model	88%	87%	88%	88%

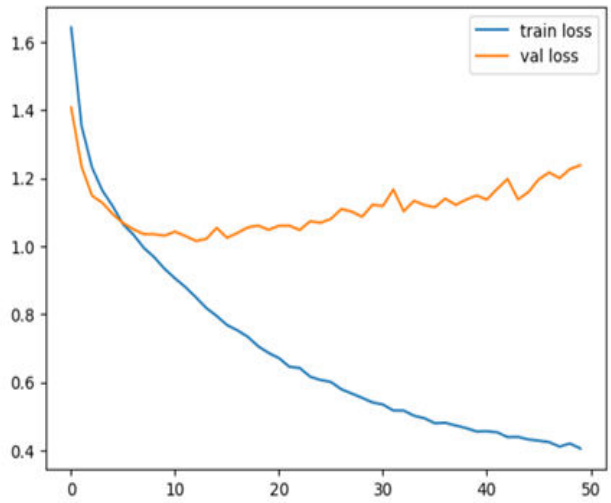


Fig.2

Training and Validation Loss Curve for Facial Expression Recognition

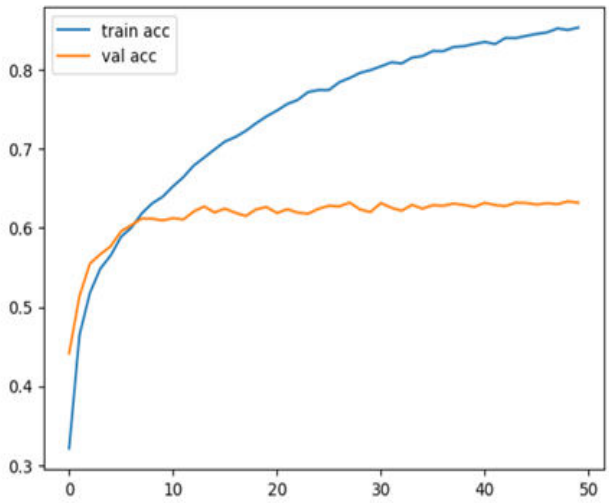


Fig.3

Training and Validation Accuracy Curve for Facial Expression Recognition

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405278|

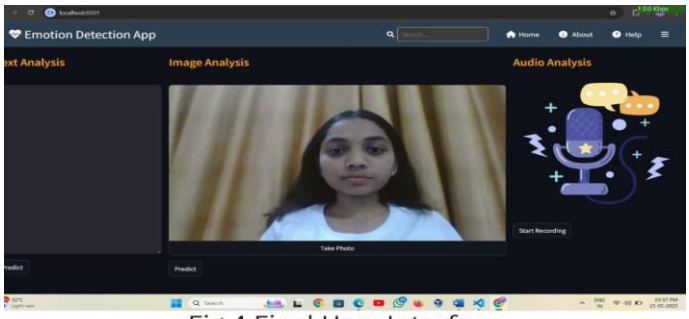


Fig.4 Final User Interface

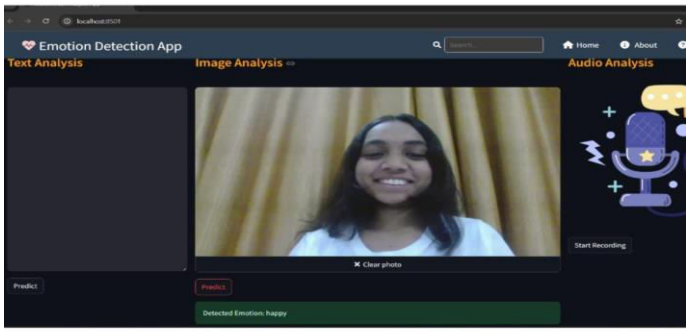


Fig.5 Image Analysis

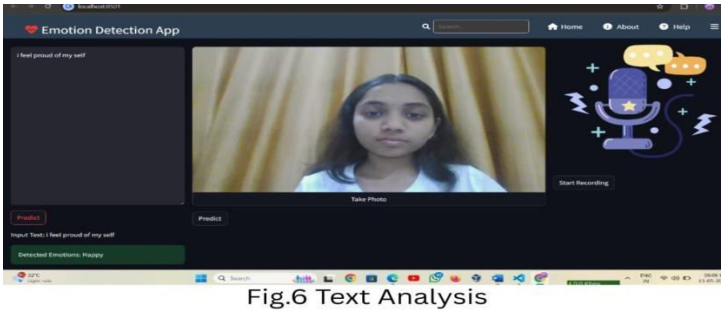


Fig.6 Text Analysis

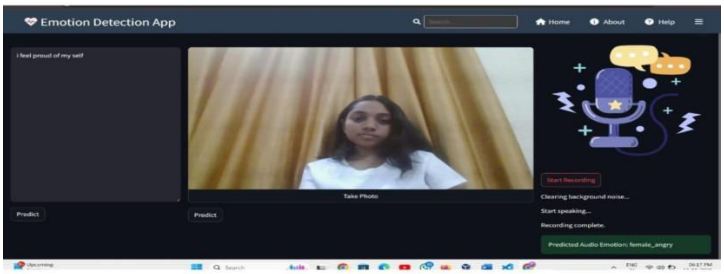


Fig.7 Audio Analysis

V. CONCLUSION

The proposed Multimodal Emotion Recognition System offers a robust and integrated framework for identifying emotional states by leveraging text, speech, and facial expressions. By employing advanced deep learning techniques such as Convolutional Neural Networks (CNNs) for facial analysis and LSTM/Bi-LSTM architectures for sequential data in text and speech, the system demonstrates notable improvements over traditional single-modality approaches.

Experimental results validate the system's effectiveness, with the audio-based model achieving the highest accuracy (98%), followed by the text-based module (92%). The hybrid models, incorporating temporal dependencies and

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405278|

contextual cues, further enhance the accuracy and robustness of the emotion classification process. This multimodal strategy helps mitigate the shortcomings of individual modalities, particularly in real-world environments where data can be incomplete or noisy.

In addition to its technical strengths, the system's Streamlit-powered user interface ensures accessibility and ease of use across a wide user base—from mental health professionals to individuals seeking self-assessment tools. Its ability to process real-time data and deliver immediate feedback positions it as a valuable asset for early detection and intervention in mental health care.

Overall, this study underscores the transformative potential of multimodal AI systems in emotion recognition and mental health assessment. It paves the way for future developments aimed at personalized healthcare, real-time monitoring, and more empathetic human-computer interactions.

REFERENCES

[1] Vandana, A., Marriwala, N., & Chaudhary, D. (2024). A hybrid model for depression detection using deep learning. Electronics and Commu nication Engineering Journal, 12(3), 45-56
[2] Abburi, H., Prasath, R. R., Shrivastava, M., & Gangashetty, S. (2017). Multimodal sentiment analysis using deep neural networks. Lecture Notes in Computer Science, 102(5), 78-92. https://doi.org/10.1007/978 3-319-58130-96
[3] Subbaiah, B., Murugesan, K., Saravanan, P., & Marudhamuthu, K. (2024). An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network. Journal of Social Media Sentiment Analysis, 22(1), 34-58.
[4] Fu, Y., Huang, B., Wen, Y., & Zhang, P. (2024). FDR-MSA: Enhancing multimodal sentiment analysis through feature disentanglement and reconstruction. Journal of Multimodal Sentiment Analysis, 18(2), 89 112.
[5] Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., & Hussain, A. (2024). Multimodal sentiment analysis: Addressing key issues and setting up the baselines. Journal of Sentiment Analysis and Applications, 20(1), 45-70.
[6] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2024). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges, and future directions. Journal of Multimodal Computing, 15(2), 101-130.
[7] Yoon, S., Byun, S., & Jung, K. (2024). Multimodal speech emotion recognition using audio and text. Journal of Speech and Language Technology, 12(3), 78-95.
[8] Ringki Das and Thoudam Doren Singh. 2023. Multimodal Senti ment Analysis: A Survey of Methods, Trends, and Challenges. ACM Comput. Surv. 55, 13s, Article 270 (December 2023), 38 pages. https://doi.org/10.1145/3586075
[9] Khoo, L. S., Lim, M. K., Chong, C. Y., & McNaney, R. (2024). Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches. Journal of Mental Health Informatics, 10(1), 23-45.
[10] Rahman, R. A., Omar, K., Noah, S. A. M., Danuri, M. S. N. M., & Al-Garadi, M. A. (2024). Application of machine learning methods in mental health detection: A systematic review. IEEE Transactions on Artificial Intelligence, 17(2), 67-89.
[11] Promoting Mental Health: Concepts, Emerging Evidence, Practice. World Health Org., Geneva, Switzerland, 2004.
[12] Global Burden of Disease Study 2013 Collaborators, "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: A systematic analysis for the global burden of disease study 2013," Lancet, vol. 386, no. 9995, pp. 743–800, 2015.
[13] ALKHAN, Tuba & Jamil, Akhtar & Hameed, Alaa. (2021). Deep Learning for Face Detection and Recognition.
[14] Miah, M.S.U., Kabir, M.M., Sarwar, T.B. et al. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. Sci Rep 14, 9603 (2024). https://doi.org/10.1038/s41598-024 60210-7
[15] https://www.kaggle.com/code/aneesh10/multimodal-sentiment-analysis
[16] https://www.kaggle.com/datasets/jonathanoheix/face-expression recognition-dataset
[17] https://www.kaggle.com/datasets/bhavikjikadara/emotions-dataset
[18] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. To wards multimodal sentiment analysis: harvesting opinions from the web. In Proceedings of the 13th international conference on multimodal interfaces (ICMI '11). Association for Computing Machinery, New York, NY, USA, 169–176. https://doi.org/10.1145/2070481.2070509
[19] 3] https://www.upgrad.com/blog/basic-cnn-architecture
[20] Ahuja, G., Alaei, A. & Pal, U. A new multimodal sentiment analysis for images containing textual information. Multimed Tools Appl (2024)

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details