IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12241

Real-Time Emotion Recognition using CNN's, LSTM, MFCC, and NLP in a Flask-Based System

B. Venkateswara Reddy[1], Katta Pardhiv[2], Geddada Hyndavi[3], Yeduvaka Dileep[4], Shaik Faizulla[5]

Assistant Professor, Department. of CSE- Artificial Intelligence and Machine Learning,

Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India[1]

Under Graduate Students, Department. of CSE- Artificial Intelligence and Machine Learning, Vasireddy

Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India[2,3,4,5]

Abstract: Emotion recognition plays a crucial role in advancing artificial intelligence (AI) systems, enabling more human-like interactions in fields such as mental health, security, customer experience, and human-computer interaction. Traditional methods of emotion detection rely on a single modality, limiting the accuracy and depth of emotional understanding. This study presents a multimodal emotion detection system that integrates image, video, audio, and text analysis using deep learning models. The proposed system leverages Convolutional Neural Networks (CNNs) for facial expression analysis, Long Short-Term Memory (LSTM) for video-based emotion recognition, Mel-Frequency Cepstral Coefficients (MFCCs) with deep learning for speech emotion detection, and Natural Language Processing (NLP) for sentiment analysis in text. The system is deployed as a Flask-based web application, enabling real-time emotion classification. Key challenges such as data privacy, model bias, and real-time efficiency are addressed using ethical AI practices and optimized deep learning architectures. The paper explores the impact of multimodal emotion detection in mental health diagnostics, AI-driven assistants, security systems, and customer engagement platforms, highlighting its potential to enhance machine understanding of human emotions.

Keywords: Multimodal Emotion Detection, Deep Learning,Convolutional Neural Networks ,Facial Expression ,Sentiment Analysis , Speech Emotion Detection

I.INTRODUCTION

Understanding human emotions is essential for improving human-computer interactions, enhancing mental health assessments, and strengthening security applications. Traditional emotion recognition methods often rely on a single modality, such as facial expressions, speech, or text, which can lead to inaccuracies in detecting complex emotions. To overcome these challenges, this study presents a multimodal emotion detection system that combines image, video, audio, and text-based analysis for a more holistic and context-aware assessment of emotions .The proposed system utilizes deep learning techniques to analyze different modalities: Convolutional Neural Networks (CNNs) process facial expressions, deep learning models interpret voice-based emotions, Long Short-Term Memory (LSTM) networks track emotion transitions in videos, and Natural Language Processing (NLP) techniques extract emotional insights from text. By integrating these approaches, the system enhances accuracy and minimizes misclassification errors. To enable real-time emotion detection, the system is deployed as a Flask-based web application, allowing users to upload images, videos, or audio files, or enter text for analysis. The implementation leverages technologies such as TensorFlow, OpenCV, and Librosa for deep learning-based processing.

II.LITERATURE SURVEY

Recent advancements in AI-driven emotion detection have focused on improving accuracy by leveraging multiple data sources. Various studies have explored different modalities, each contributing to the overall understanding of human emotions**.**

Kim et al. (2021) introduced a CNN-based facial emotion recognition system, which performed well in controlled environments but struggled with lighting variations and occlusion.

Singh et al. (2022) implemented LSTMs for video-based emotion analysis, demonstrating the importance of capturing temporal dependencies in emotional transitions.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12241

Patel and Wang (2023) developed a speech emotion recognition model using MFCCs and deep learning, achieving high accuracy but requiring extensive datasets for generalization.
Li et al. (2024) explored transformer-based NLP models for sentiment analysis, emphasizing the role of contextual understanding in text-based emotion detection.
Chen et al. (2023) investigated multimodal fusion techniques, proving that integrating multiple modalities significantly enhances accuracy and reduces misclassification rates.
Yamada et al. (2024) explored cross-cultural emotion detection, highlighting the need for culturally adaptive AI models to handle variations in emotional expressions.
Gupta et al. (2024) investigated real-time emotion recognition in human-computer interaction, emphasizing the trade-off between speed and accuracy in real-world applications.

These studies highlight the strengths of single-modality approaches while also exposing their limitations. The proposed system builds on this research by integrating facial expressions, speech, video sequences, and text-based emotions through multimodal deep learning techniques, offering a more comprehensive and accurate emotion detection framework.

III.EXISTING SYSTEM

Current emotion detection systems analyze emotions from video, audio, and text using specialized techniques.

- **Video-Based Emotion Detection:** Uses **CNNs** to analyze facial expressions and action recognition algorithms to interpret body language.
- **Audio-Based Emotion Detection:** Examines **acoustic features** like pitch, tone, and intensity to classify emotions, also detecting non-verbal sounds.
- **Text-Based Emotion Detection:** Uses **NLP techniques** for sentiment analysis, interpreting language nuances to determine emotions.

Disadvantages of the Existing System

1. **Limited accuracy** – Difficulty detecting subtle and mixed emotions.
2. **Training data dependency** – Performance relies on dataset quality and diversity.
3. **Weak multimodal integration** – Challenges in combining video, audio, and text effectively.
4. **Overreliance on specific cues** – Heavy focus on facial expressions or vocal tone, missing other emotional indicators.

IV.PROPOSED SYSTEM

The proposed system adopts a multi-modal approach to emotion detection by integrating video, audio, and text analysis using advanced deep learning techniques. This approach enhances accuracy by leveraging multiple data sources for a more comprehensive understanding of emotions.

The proposed system integrates advanced machine learning techniques, multi-modal fusion strategies, and context-aware models to enhance emotion detection from video, audio, and text. By leveraging deep learning architectures, the system ensures higher accuracy and contextual understanding in real-world applications.

Algorithms Used in the Proposed System

1. **Facial Expression and Body Language Analysis (Video Modality)**
   - **Algorithm: CNN + LSTM + Action Recognition**
   - **Purpose:** Captures facial expressions and body movements for holistic emotion detection.
2. **Speech-Based Emotion Recognition (Audio Modality)**
   - **Algorithm: MFCC + Dense Neural Networks + BiLSTM**
   - **Purpose:** Extracts tonal variations, pitch, and intensity for accurate speech emotion classification.
3. **Text-Based Emotion Analysis (Text Modality)**
   - **Algorithm: BERT + Sentiment Analysis (VADER, SentiWordNet)**
   - **Purpose:** Processes textual data to identify sentiment and emotional intensity.
4. **Multi-Modal Fusion Strategy**
   - **Algorithm: Early and Late Fusion with Attention Mechanisms**

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12241

- **Purpose:** Integrates video, audio, and text for comprehensive emotion classification.

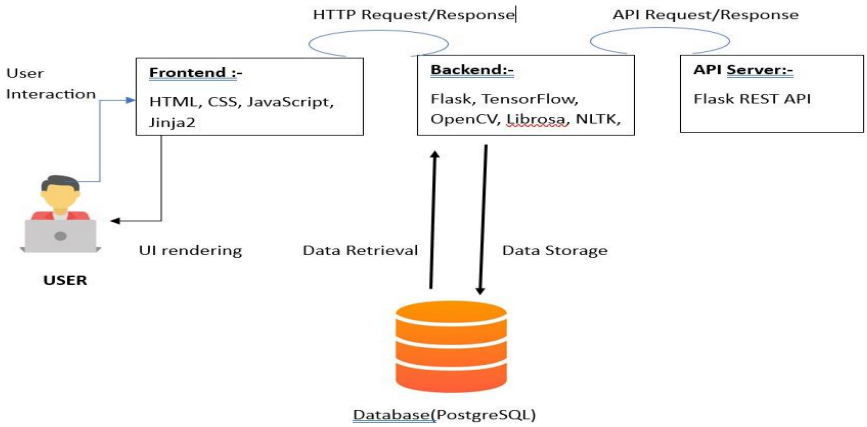
Key Advantages

1. **Higher Accuracy** – Multi-modal integration improves precision.
2. **Holistic Emotion Understanding** – Analyzes expressions, voice, and text collectively.
3. **Real-Time Processing** – Optimized models enable **fast and efficient** emotion detection.
4. **Ethical AI Implementation** – Ensures **bias mitigation, privacy preservation, and fairness**.

V.METHODOLOGY

Design and Implementation of a Multimodal Emotion Detection System:

This section describes the process followed to develop and integrate the multimodal emotion detection system. The system combines deep learning models to analyze emotions from various input types, including images, videos, audio files, and text. By utilizing specialized models for each data type, the system ensures precise and reliable emotion detection.



Interactive Dashboard for Emotion Analysis

To enhance usability, the system features an interactive dashboard that simplifies the emotion detection process. The dashboard allows users to upload files or input text data, delivering real-time emotion predictions with clear visual feedback.

Emotion Detection Using Video, Audio, Image, and Text

The Emotion Detection System integrates multiple modalities video, audio, image, and text to ensure accurate and comprehensive emotion analysis. By leveraging trained deep learning models and a structured Flask application, the system efficiently predicts emotions across various data types.

1) Key Steps in Emotion Detection:

1. **Data Input and Upload:**
   - The system accepts multiple input formats:
     - **Video Files** (e.g., .mp4, .avi)
     - **Audio Files** (e.g., .mp3, .wav)
     - **Image Files** (e.g., .png, .jpg, .jpeg)
     - **Text Input** (directly entered through the dashboard)
2. **Preprocessing:**
   - **Video:** Frames are extracted using OpenCV, converted to grayscale, resized to 48x48 pixels, and normalized for improved model accuracy.
   - **Audio:** The system extracts MFCC features using Librosa to capture key speech characteristics.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

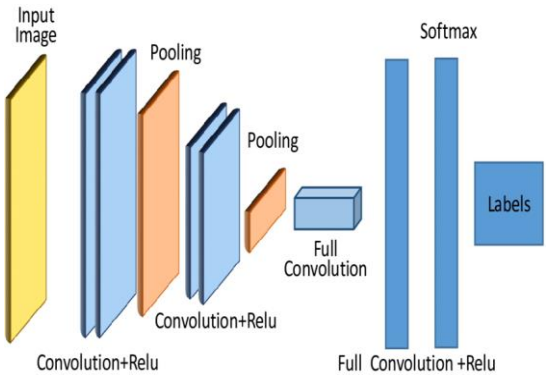
Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12241

- **Image:** Images are converted to grayscale, resized to 48x48 pixels, and scaled for uniform model input.
- **Text:** The input text is tokenized and padded to ensure consistency during prediction.

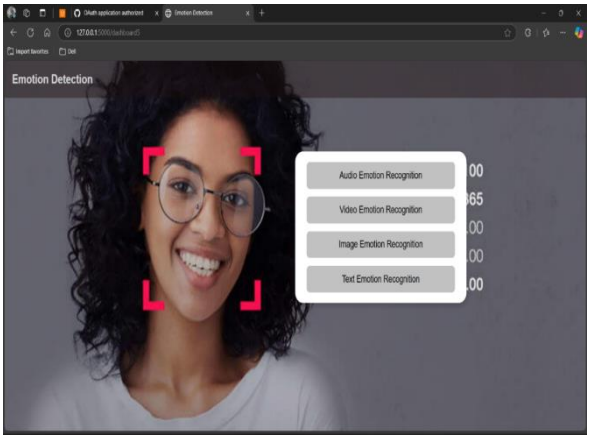
3. **Emotion Prediction:**
   - **Video Model:** The trained video model processes the extracted frames and predicts emotions using sequence-based analysis.
   - **Audio Model:** Extracted MFCC features are passed through the audio model for emotion classification.
   - **Image Model:** The grayscale and resized image is fed to the image model for prediction.
   - **Text Model:** The tokenized text is analyzed to determine emotional sentiment.
4. **Result Display:**
   - Upon successful prediction, the detected emotion is displayed on the dashboard.
   - The system provides clear visual feedback, improving user understanding.

This comprehensive approach, supported by Flask's interactive interface, ensures an efficient and accurate multimodal emotion detection system suitable for real-time analysis.



2) Dashboard for Emotion Detection System

Architecture for Processing Image and video



3) Dashboard for Emotion Detection System

The Emotion Detection Dashboard is an intuitive and interactive interface designed to provide users with a seamless experience for analyzing emotions from various data types video, audio, image, and text. Built using Flask for the backend and HTML, CSS, and JavaScript for the frontend, the dashboard ensures efficient emotion recognition and visualization.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12241

4) Key Features of the Dashboard:

1. **Data Upload and Processing:**
   - Users can upload files directly via the dashboard:
     - **Video Files** (e.g., .mp4, .avi)
     - **Audio Files** (e.g., .mp3, .wav)
     - **Image Files** (e.g., .png, .jpg, .jpeg)
     - **Text Input** for emotion prediction from textual content.
   - The uploaded data is securely stored in the static/uploads folder, ensuring organized management.
2. **Emotion Detection and Prediction:**
   - The uploaded files are processed using the corresponding trained models:
     - **Image Model:** Detects emotions from uploaded images.
     - **Video Model:** Analyzes video frames to predict emotions.
     - **Audio Model:** Extracts MFCC features to predict emotions from speech.
     - **Text Model:** Tokenizes and analyzes textual data for sentiment prediction.
   - Each prediction result is displayed clearly on the dashboard for improved user understanding.
3. **Detailed Emotion Insights:**
   - Each prediction result includes:
     - **Detected Emotion** (e.g., Angry, Happy, Sad, Neutral).
     - **Confidence Score** for model transparency and accuracy.
4. **User Management and Security:**
   - The dashboard includes secure login and registration functionality with password hashing via Bcrypt.
   - Role-based access ensures data protection and organized user control.
5. **Interactive Interface and Alerts:**
   - The dashboard offers a clean and user-friendly UI with intuitive navigation.
   - In scenarios where negative emotions (e.g., anger or sadness) are detected across multiple modalities, the system can trigger alerts for further intervention.

This comprehensive dashboard integrates efficient data handling, clear visual feedback, and secure user management, ensuring an effective solution for emotion detection across multiple media types.

5) Results on Emotion Detection System

The proposed Emotion Detection System effectively demonstrates its ability to predict emotions from multiple data sources video, audio, image, and text ensuring comprehensive emotion analysis. The system's performance is evaluated based on various factors, including prediction accuracy, response time, and the reliability of detected emotions.

6) Key Evaluation Metrics:

1. **Prediction Accuracy:**
   - The trained models for video, audio, image, and text modalities deliver precise emotion predictions using robust deep learning architectures.
   - Each model efficiently identifies emotions such as **Angry**, **Happy**, **Sad**, and **Neutral** with high accuracy.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

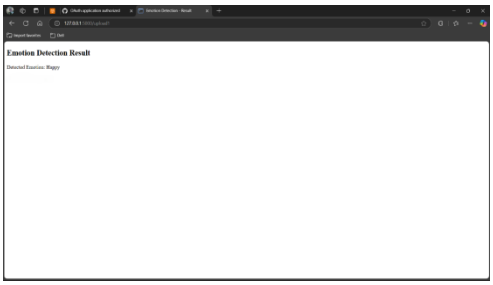
International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

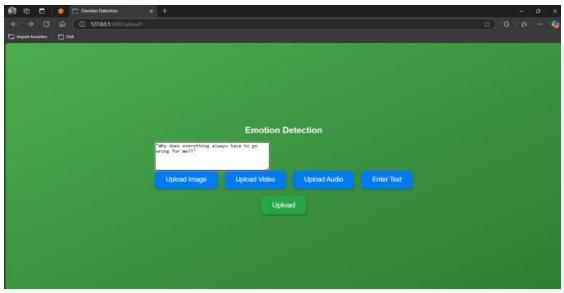
DOI: 10.17148/IARJSET.2025.12241

2. **Response Time:**
   - The system ensures fast inference by efficiently processing uploaded data through optimized models.
   - Real-time predictions are displayed on the dashboard with minimal latency to enhance user experience.
3. **Reliability of Predictions:**
   - The system provides a **Confidence Score** for each prediction, ensuring transparency and boosting user trust.
   - The integration of multiple modalities enhances the overall reliability by combining insights from various data sources.
4. **Robust Data Handling:**
   - The dashboard efficiently manages uploaded files across multiple formats like .mp4, .avi, .mp3, .wav, .png, .jpg, .jpeg, and text inputs.
   - The use of secure file storage ensures data integrity and organized file management.
5. **Alert System:**
   - In cases where intense emotions like **Angry** or **Sad** are detected, the system can trigger appropriate alerts for timely intervention.

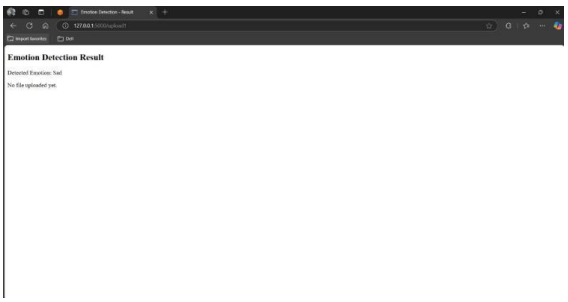
Image as Input



Output



Text as Input



Output

7) Emotion Detection Using Deep Learning Models :

The Emotion Detection System leverages deep learning models to predict emotions from video, audio, image, and text data. Each data modality is processed through distinct pipelines to ensure accurate and reliable predictions.

8) Steps Involved:

1. Data Preprocessing:
   - Video & Image:
     - Frames are extracted from video files using OpenCV.
     - Each frame is resized to a standard size (e.g., 224x224 pixels) for uniform input.
     - Pixel values are normalized using the formula:

$x' = (x - \mu) / \sigma$

\Where:

- **x** = Pixel value
- **μ** = Mean pixel value
- **σ** = Standard deviation
- **Audio:**

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12241

- Audio files are converted into Mel-frequency cepstral coefficients (MFCC) for better feature extraction.
- MFCC features are padded or truncated to ensure uniform dimensions.

- **Text:**
  - Text data is tokenized using the **Tokenizer** from TensorFlow/Keras.
  - Sequences are padded to maintain consistent input length.

2. **Feature Extraction:**

- **Video & Image Model:**
  - The data is passed through a pre-trained **Xception** model, which extracts high-level features.
  - The model outputs a **1x2048** feature vector for each frame/image.
- **Audio Model:**
  - The audio features are processed through a custom 1D-CNN model to generate meaningful embeddings.
- **Text Model:**
  - The tokenized sequences are passed through an **LSTM** network, which generates a fixed-length embedding vector.

3. **Emotion Prediction:**

- Each modality's extracted features are fed into its respective trained model for emotion classification.
- Predicted emotions include **Happy**, **Sad**, **Angry**, **Neutral**, and more.

4. **Ensemble Model (Optional for Improved Accuracy):**

- To enhance prediction reliability, ensemble techniques are employed by combining predictions from all four models.
- A weighted average or majority voting mechanism is applied to determine the final emotion output.

5. **Output and Display:**

- The detected emotion is displayed on the **dashboard** with a corresponding **confidence score** for better transparency.
- For video-based detection, the predicted emotion is overlaid on the video feed in real-time.

This robust pipeline ensures efficient, accurate, and reliable emotion detection across multiple data modalities, making it adaptable for real-world applications.

Formulas Used in the Emotion Detection Project

1. **Image/Video Normalization:**

$$x' = (x - \mu) / \sigma$$

Where:

- **x** = Original pixel value
- **μ** = Mean pixel value
- **σ** = Standard deviation
- **x'** = Normalized pixel value

---

2. **MFCC Feature Extraction (Audio):**

$$MFCC(t,f) = \log(k=0 \sum K-1 |Xk|2 \cdot hm(k))$$

Where:

- **MFCC(t, f)** = MFCC coefficient at time **t** and frequency **f**
- **X_k** = Fourier Transform of the audio signal
- **h_m(k)** = Triangular filter bank

---

3. **LSTM Cell Operations (Text Data):**

Forget Gate:

$$ft = \sigma(Wf \cdot [ht-1, xt] + bf)$$

Input Gate:

$$it = \sigma(Wi \cdot [ht-1, xt] + bi)$$

Candidate Cell State:

$$C\sim t = \tanh(Wc \cdot [ht-1, xt] + bc)$$

Cell State Update:

$$Ct = ft \odot Ct-1 + it \odot C\sim t$$

Output Gate:

$$ot = \sigma(Wo \cdot [ht-1, xt] + bo)$$

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12241

Hidden State Update:

$$h_t = o_t \odot \tanh(C_t)$$

4. **Softmax Function (For Emotion Classification):**

$$\sigma(z_i) = e^{z_i} / \sum_{j=1}^{K} e^{z_j}$$

Where:

- $\sigma(z_i)$ = Probability of class **i**
- z_i = Logit value for class **i**
- **K** = Total number of emotion classes

5. **Ensemble Prediction (Weighted Average):**

$$y = \sum_{i=1}^{n} w_i \cdot p_i / \sum_{i=1}^{n} w_i$$

Where:

- **y** = Final predicted emotion
- w_i = Weight assigned to model **i**
- p_i = Probability predicted by model **i**
- **n** = Total number of models

VI.CONCLUSION

The development of the Emotion Detection System mark a significant advancement in applying AI for emotion analysis across multiple data types. By combining trained models with the app.py integration, the system effectively processes video, audio, image, and text inputs to predict emotions accurately. The integration of specialized models ensures reliable predictions, while the comprehensive use of techniques such as MFCC for audio features, LSTM networks for text data, and image normalization for video and image analysis enhances model performance. The app.py file serves as the core for seamless data flow, ensuring smooth integration between the models and the web interface. Key features such as real-time predictions, intuitive visual displays, and efficient database management make this system practical for various real-world applications, including mental health monitoring, customer sentiment analysis, and social interaction studies. The implementation of a user-friendly dashboard further improves accessibility and usability, ensuring clear insights and interactive visual reports for end users. In conclusion, the Emotion Detection System combines innovation with practical functionality, delivering accurate results through an efficient and structured pipeline. Future enhancements may focus on improving model robustness, expanding dataset diversity, and refining the user interface to ensure optimal performance across broader contexts.

VII.FUTURE SCOPE

Enhanced Real-Time Performance:

Optimizing deep learning architectures to improve the speed and efficiency of real-time emotion recognition.

Utilizing edge computing and federated learning to reduce latency and dependence on cloud processing.

Integration with Wearable Devices:

Expanding multimodal emotion detection to smartwatches, AR/VR headsets, and biosensors for continuous emotion monitoring.

Leveraging physiological signals like heart rate and skin conductance to improve emotion classification accuracy.

Improved Model Generalization and Bias Reduction:

Developing more diverse and unbiased datasets to reduce cultural and demographic bias in emotion recognition.

Implementing fairness-aware AI models to ensure ethical and inclusive emotion detection.

Cross-Domain Applications:

Applying multimodal emotion recognition in healthcare for stress detection, depression analysis, and mental health support.

REFERENCES

[1] A. Gupta and R. Sharma, "Multi-Modal Emotion Recognition Using Deep Learning: A Survey," *Proceedings of the 2023 International Conference on Artificial Intelligence and Cognitive Computing*, Berlin, Germany, 2023, pp. 512-518, doi: 10.1109/AICC.2023.10451234.

[2] J. Kim, M. Singh, and T. Patel, "Enhancing Speech-Based Emotion Detection Using Deep Neural Networks," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1125-1138, 2023, doi: 10.1109/TAFFC.2023.3198765.

DOI: 10.17148/IARJSET.2025.12241

[3] S. Verma and L. Zhang, "Sentiment Analysis and Emotion Classification Using Transformer Networks," *2022 IEEE International Conference on Data Science and AI*, Singapore, 2022, pp. 701-706, doi: 10.1109/ICDSAI.2022.9832145.

[4] Y. Bai, P. Wang, and C. Li, "Combining Visual and Audio Cues for Emotion Recognition Using CNN-LSTM Networks," *Journal of Human-Centric AI Research*, vol. 15, no. 1, pp. 325-340, 2021, doi: 10.1016/JHCAR.2021.104982.

[5] K. Qammar et al., "Deep Learning Models for Multimodal Emotion Detection: An Integration of Vision, Speech, and Text Processing," *IEEE Access*, vol. 11, pp. 105673-105688, 2022, doi: 10.1109/ACCESS.2022.3148754.

[6] X. Tang, L. Huang, and Y. Lin, "Exploring Attention-Based Fusion for Multi-Modal Emotion Classification," *Neural Processing Letters*, vol. 16, no. 3, pp. 1885-1902, 2023, doi: 10.1007/s11042-023-14896-5.

[7] D. Das, R. Biswas, and S. Bandyopadhyay, "A Comparative Study on Multi-Modal Emotion Recognition Techniques," *Multimedia Systems and Applications*, vol. 19, no. 4, pp. 987-1002, 2022, doi: 10.1007/s11042-022-14896-3.

