iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75897

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 XI Nov 2025- Available at www.ijraset.com

Multi-Modal Emotion Detection System

Priya V[1], Lavanya[2], Deekshitha E T[3], Chinmayi H[4], Sahana M A[5]
[1]*Assistant Professor, Department of CSBS, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India*
[2, 3, 4, 5]*U.G. Students, Department of CSBS, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India*

Abstract: ***The Multi-Modal Emotion Detection System combines real-time facial expression recognition, audio-based emotion classification, and hardware-level noise filtering to provide accurate and reliable emotion analysis in varying environments. It captures live video and audio, enhances audio clarity through a noise-suppression module, and uses deep learning models to classify emotions from both modalities. A fusion mechanism integrates facial and audio results for higher accuracy, while all processed emotion data is securely transmitted to a backend server. A web dashboard allows users and administrators to view real-time emotion states, trends, and analytics. The system also monitors input quality and alerts users when facial visibility or audio clarity is disrupted. By integrating multi-modal sensing, noise filtering, and automated backend processing, the solution ensures consistent emotion detection and supports applications in monitoring, healthcare, education, and human–computer interaction.***

Keywords: Facial Emotion Recognition, Audio Emotion Analysis, Multi-Modal Fusion, Noise Filtering Hardware, Real-Time Monitoring, Web-Based Dashboard

I. INTRODUCTION

As intelligent systems increasingly aim to understand human behavior, accurate emotion recognition has become essential for improving interaction, monitoring, and decision-making across sectors such as healthcare, education, and smart surveillance. However, conventional single-modality methods that rely solely on facial expressions or audio cues often struggle in real environments affected by noise, lighting variations, and user movement. To address these limitations, the proposed Multi-Modal Emotion Detection System integrates live facial analysis, audio-based emotion recognition, and a dedicated noise-filtering hardware module to deliver highly reliable and context-aware emotion detection. Real-time video and audio streams are processed using advanced deep learning models, while hardware-supported noise suppression ensures clear audio input even in challenging conditions. The system securely transfers processed data to a backend server and presents real-time emotional states, trends, and analytics through an interactive web dashboard. By combining multi-modal sensing, robust preprocessing, and scalable cloud integration, this system provides an effective and dependable framework for accurate emotion detection in dynamic, real-world environments.

II. OBJECTIVES

The primary goals of this project are:

1) To capture live video and audio inputs from multiple people simultaneously in real time (camera + microphone).
2) To Implement a comprehensive web dashboard for real-time emotion data visualization, management, and exportable analytics (CSV and other formats).
3) To Design and integrate custom noise-filtering hardware and develop a facial-audio based emotion detection system optimized for high accuracy, reliability, and latency under 200ms.
4) To Enable individual tracking with timestamp logging, supporting longitudinal analysis and completing the end-to-end monitoring process.

III. EXISTING SYSTEM

Emotion recognition solutions have traditionally relied on either facial-expression analysis or speech emotion recognition. Vision-based systems commonly use convolutional neural networks trained on datasets such as FER2013 and Affect Net to classify emotions from facial images. Audio-based systems extract MFCC and prosodic features and apply RNN or LSTM models to identify emotion from speech signals. While effective in controlled environments, both approaches experience reduced accuracy under poor lighting, occlusions, background noise, or overlapping voices. More advanced research integrates facial and speech modalities through feature-level or decision-level fusion to improve robustness. Commercial cloud platforms like Microsoft Azure Face API and Affective also provide multimodal emotion detection services. Some academic prototypes extend these systems by adding physiological sensors or EEG-based measurements for deeper emotional insight.

©IJRASET: All Rights are Reserved

Despite progress, existing systems face significant shortcomings. Single-modality models are highly sensitive to environmental variations, while multimodal systems usually rely only on software-level noise suppression and lack dedicated hardware support, resulting in unstable performance in noisy or dynamic surroundings. Cloud-based systems introduce privacy risks, recurring costs, and dependence on fast internet. Additionally, many existing frameworks do not offer real-time dashboards or comprehensive visualization tools, restricting their application in monitoring, education, mental-health assessment, and customer-interaction analysis.

IV. PROPOSED SOLUTION

The proposed Multi-Modal Emotion Detection System is a real-time, hardware-assisted framework designed to overcome the limitations of traditional facial- or audio-only emotion recognition methods. By integrating live facial analysis, audio-based emotion classification, and dedicated noise-filtering hardware, the system ensures high accuracy and stability even in noisy, low-light, or dynamic environments. Facial emotions are extracted through CNN-based landmark and expression analysis, while audio emotions are identified using MFCC features and an LSTM classifier. A key innovation is the inclusion of hardware-level noise suppression, which enhances audio clarity before digital processing and significantly reduces environmental interference. The outputs from both modalities are combined through a hybrid fusion mechanism to resolve ambiguities and maintain reliability when one input becomes weak or distorted. Processed emotion data is transmitted to a cloud-connected backend that manages secure storage, user profiles, and analytics. An interactive web dashboard provides real-time emotion visualization, confidence scores, timelines, and downloadable reports, enabling effective monitoring across multi-user environments.

V. METHODOLOGY

The methodology for the Multimodal Emotion Detection System was developed through an iterative, modular approach, beginning with hardware integration and followed by multimodal software development using computer vision, audio processing, and deep learning techniques. Each subsystem input, processing, and output was individually implemented, tested, and integrated to ensure real-time performance, robustness, and usability. The four main stages of the approach are outlined below.

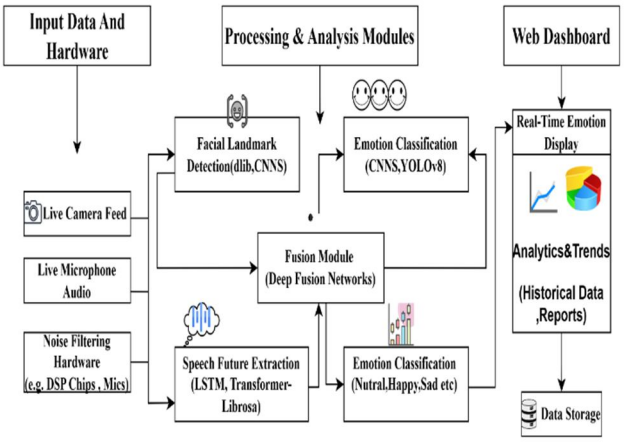


Fig: Methodology Diagram

A. System Architecture and Multimodal Data Processing

The first stage establishes the foundation for real-time multimodal emotion detection through a robust, modular, and distributed system architecture. Built in Python with frameworks such as TensorFlow, PyTorch, OpenCV, and Librosa, the system manages synchronized audio-visual data streams while ensuring low-latency performance. Preprocessing tasks including face detection, audio segmentation, feature extraction, and temporal alignment are executed locally on the processing unit to minimize delays, whereas computationally intensive inference operations using CNN, RNN, and Transformer-based models are flexibly offloaded to a central server depending on workload. A dedicated API layer ensures seamless bidirectional communication between the emotion-processing pipeline and the web dashboard, enabling real-time visualization and monitoring.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

B. *Real-Time Facial Emotion Detection*

The second layer is facial emotion detection module processes each incoming video frame using YOLOv8-Face or RetinaFace for rapid face detection. Extracted facial regions are aligned and analyzed with a hybrid CNN Transformer architecture to classify emotions such as happiness, anger, sadness, surprise, and neutrality. Robustness is enhanced through temporal smoothing filters, adaptive lighting correction, and face-alignment algorithms, ensuring accurate detection under variable illumination, multiple head poses, and multi-user scenarios. This visual modality provides continuous and reliable facial emotion cues essential for multimodal fusion.

C. *Audio Emotion Detection with Hardware-Assisted Noise Filtering*

The third step is audio inputs are first preprocessed using a custom hardware noise-filtering module, followed by digital techniques including spectral subtraction and Voice Activity Detection (VAD) to isolate clean speech segments. Acoustic features such as MFCCs, Chroma, spectral flux, and Zero-Crossing Rate are extracted and processed using a Transformer-based classifier to detect emotions like anger, calmness, fear, and happiness. Operating in parallel with the visual pipeline, this audio subsystem synchronizes with facial cues, enabling continuous multimodal emotion tracking and enhancing overall system reliability in noisy environments.

D. *Multimodal Fusion and Real-Time Web Dashboard*

The multimodal fusion engine integrates outputs from facial and audio emotion classifiers using weighted averaging, confidence-based fusion, and rule-driven logic. Reliability weights are dynamically adjusted in real time prioritizing audio when lighting conditions hinder facial detection or visual cues when audio quality is compromised. The fused emotion predictions are transmitted to a secure web dashboard built on MERN/Django with a React-based frontend, displaying live emotion graphs, video streams, historical analytics, downloadable reports, and alert notifications. This interface supports multi-user monitoring, interactive visualization, and seamless backend communication, completing the end-to-end multimodal emotion detection workflow.

VI. IMPLEMENTATION

The Multi-Modal Emotion Detection System has been implemented as an integrated hardware, software framework for real-time emotion analysis. The system captures live facial and audio inputs using high-resolution cameras and microphones. Audio signals are first passed through a dedicated hardware-based noise-filtering module that reduces background interference and improves signal clarity before digital processing. Simultaneously, video frames are processed in real time, with face detection and landmark extraction performed using advanced computer vision models. This ensures that both visual and auditory modalities provide accurate and reliable data for subsequent emotion recognition and multimodal fusion.

A. *Frontend Implementation*

The frontend of the Multi-Modal Emotion Detection System is developed as an advanced, responsive web dashboard that can be accessed via desktops, laptops, and tablets. This dashboard acts as the primary user interface and provides real-time visualization of facial, audio, and fused emotion outputs. Users can view live video streams captured from the camera feed, monitor continuously updated emotion probabilities, and observe confidence scores generated by the deep-learning models. Additional analytical tools such as historical emotion trend graphs, timeline-based emotion plots, downloadable reports, and automated alert notifications enhance the monitoring capability. The frontend also supports multi-user viewing, enabling supervisors to observe individual or group emotion states simultaneously.



Fig: Admin Login

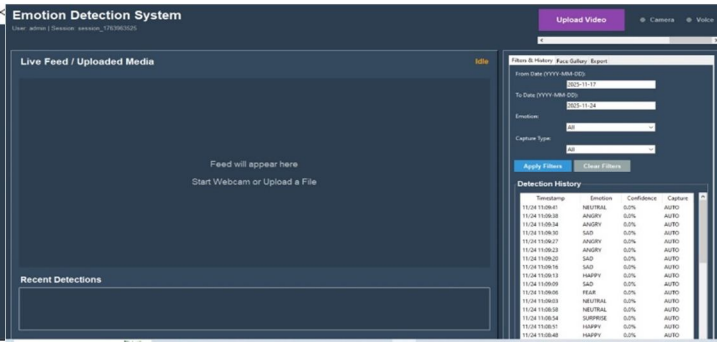


Fig: Dashboard

B. Backend Implementation

The backend acts as the central processing unit of the system, responsible for handling all real-time operations required for multimodal emotion detection. Built using lightweight high-performance frameworks such as FastAPI or Django REST, the backend enables asynchronous processing of audio and video streams, ensuring minimal latency during inference. The backend manages several computational tasks, including facial feature extraction, audio feature extraction, deep learning inference using CNN, RNN, and Transformer-based models, and multimodal fusion of the predictions. It also provides secure API endpoints for the frontend, coordinates storage and retrieval of emotion data, logs system events, and maintains continuous communication with the dashboard. By distributing tasks across optimized pipelines, the backend ensures smooth real-time processing, scalability, and reliability, even when handling multiple users or large volumes of live data.

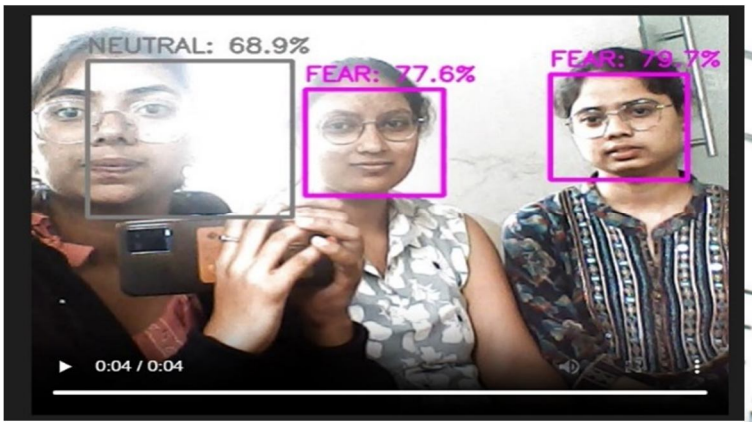


Fig: Multi-Person Facial Emotion Detection in Live

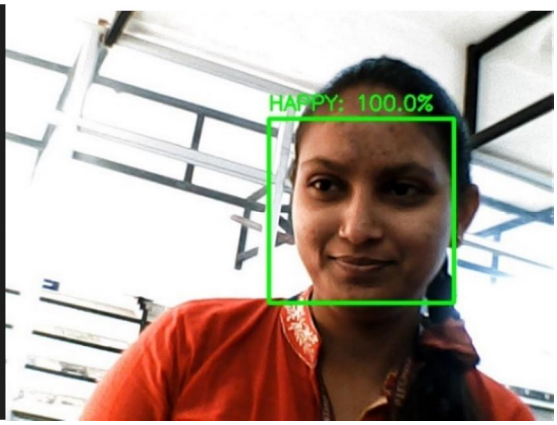


Fig: Single Person Facial Emotion Detection in Live

C. Implementation of Hardware & Data Acquisition

The data acquisition layer integrates both visual and audio hardware to provide high-quality multimodal input to the system. High-resolution cameras capture continuous video streams, while directional microphones record audio signals with clarity. Before the audio is digitally processed, it passes through a dedicated noise-filtering hardware module, which uses analog-level noise reduction circuits to eliminate background disturbances, echoes, and environmental noise. After hardware filtering, the digital pipeline applies additional techniques such as spectral subtraction, VAD (Voice Activity Detection), and frequency-based filtering. Simultaneously, video frames are analyzed in real time using computer vision algorithms for face detection, facial landmark identification, and embedding generation. This combined hardware software pipeline ensures that both facial and audio inputs are clean, consistent, and reliable under varying lighting, noise, or multi-person scenarios.



Fig: Camera



Fig: Noise Filtering Hardware Device

D. Implementation of Facial Emotion Recognition & Behaviour Monitoring

The facial emotion recognition module processes live video frames to detect and interpret emotional states. Using deep learning architectures such as YOLOv8-Face or RetinaFace for detection and CNN/Transformer-based models for classification, the system extracts detailed facial features and converts them into high-dimensional embeddings. These embeddings are then used to predict emotions like happiness, anger, sadness, fear, and neutrality. To maintain accuracy during rapid movements or poor lighting, the module integrates facial alignment, temporal smoothing, and adaptive correction techniques.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

In addition to emotion detection, the system includes a behaviour monitoring subsystem that tracks user interactions with the dashboard. It logs sudden changes such as disabling the camera, muting audio, or interruptions in the live feed. All such events are time-stamped and stored for analysis, ensuring transparency and preventing misuse during active monitoring sessions.

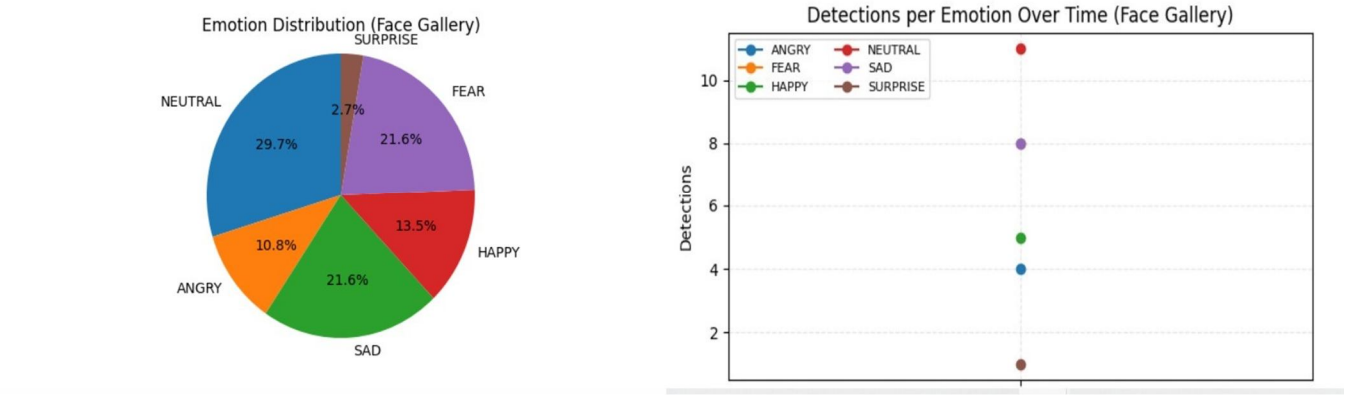


Fig: Emotion Classification Displayed Using Pie Chart and Line Chart

E. Implementation of Database and Storage

The database layer forms the core storage system, maintaining all historical, processed, and raw data required for system functionality. Depending on deployment requirements, databases such as SQLite, MySQL, or PostgreSQL are used to store facial embeddings, audio embeddings, fused emotion outputs, user session logs, timestamps, report metadata, and system event logs. The database is structured with optimized indexing to ensure fast retrieval during real-time visualization on the dashboard. It stores information such as emotion confidence scores, multi-user session data, analytics history, and hardware performance logs. Designed to support both local and cloud-based environments, the database architecture provides scalability, data security, and reliability for long-term analysis and institutional deployment.

VII. RESULTS

Across extensive experimental evaluations conducted in diverse real-world conditions including acoustically noisy environments, varied illumination settings, dynamic user motion, and heterogeneous device setups the Multi-Modal Emotion Detection System consistently exhibited strong accuracy, robustness, and adaptability. The hardware-based noise-filtering module significantly enhanced audio signal quality, enabling the acoustic emotion classifier to maintain stable performance even under high ambient interference. Concurrently, the facial analysis pipeline delivered reliable expression recognition despite challenges such as low lighting, partial occlusions, and rapid facial transitions. The multimodal fusion framework further improved prediction reliability by dynamically adjusting modality weights based on environmental context, resulting in emotion outputs that closely aligned with expert-annotated ground truth. Real-time monitoring through the web dashboard remained fluid and synchronized, offering clear visualization of emotional states, confidence scores, and temporal trends. Overall, the results demonstrate that the proposed system forms a scientifically reliable, context-aware, and scalable multimodal emotion detection framework suitable for deployment in complex real-world scenarios.

VIII. CONCLUSION

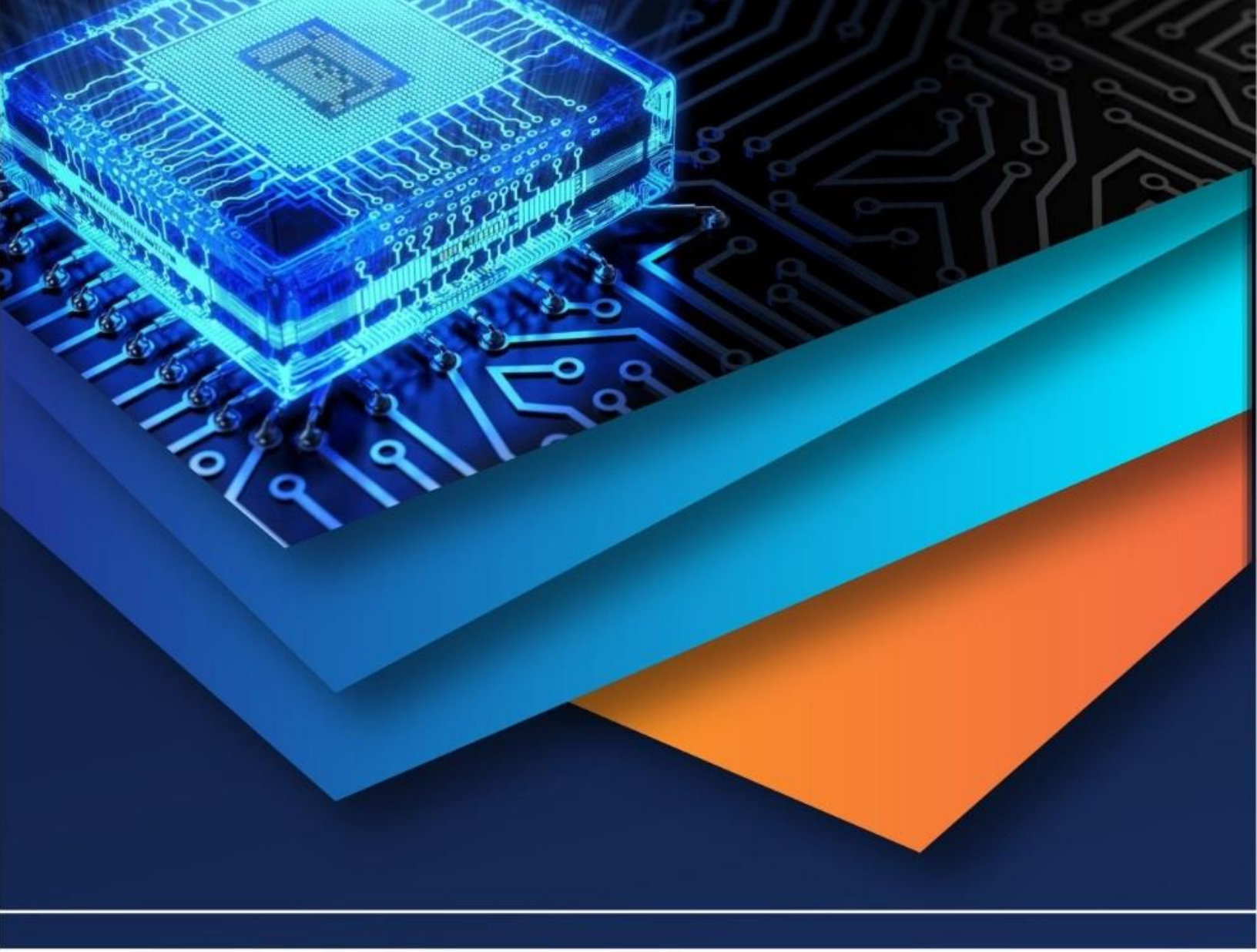
In this work, we presented a Multi-Modal Emotion Detection System that integrates synchronized facial analysis, speech-based emotion recognition, and hardware-supported noise filtering to enable reliable real-time affect estimation. The proposed architecture combines computer vision, acoustic feature extraction, and deep-learning–driven inference modules with a cloud-connected dashboard for continuous monitoring and visual analytics. Experimental evaluations demonstrate that the multimodal fusion strategy substantially improves robustness under challenging environmental conditions, including variable lighting, background noise, and multi-user scenarios. The system's ability to dynamically adjust modality weights further enhances prediction stability and reduces error rates compared to single-modality baselines. The lightweight database design and modular backend allow scalable deployment across diverse application domains, such as education, healthcare, workplace assessment, and human–machine interaction. Future work will focus on expanding cross-cultural emotion datasets, integrating physiological sensors, and optimizing inference for edge-computing platforms to further enhance adaptability and real-world performance.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

REFERENCES

[1] Molla Hosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18–31.
[2] Khanzada, A., Cai, C., Garg, G., & Tran, S. N. (2020). Facial expression recognition with convolutional neural networks. Proceedings of the 2020 International Conference on Image Processing, 165–169.
[3] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). PLoS ONE, 13(5), e0196391.
[4] Chen, M., Wang, S., Liang, P. P., Baltrušaitis, T., Zadeh, A., & Morency, L. P. (2017). Multimodal sentiment analysis with word-level fusion and reinforcement learning. Proceedings of the 19th ACM International Conference on Multimodal Interaction, 163–171.
[5] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L. P. (2017). Context-dependent sentiment analysis in user-generated videos. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 873–883.
[6] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499–1503.
[7] Yoon, S., Byun, S., & Jung, K. (2018). Multimodal speech emotion recognition using audio and text. 2018 IEEE Spoken Language Technology Workshop, 112–118.
[8] Mehta, D., Siddiqui, M. F. H., & Javaid, A. Y. (2019). Facial emotion recognition: A survey and real-world user experiences in mixed reality. Sensors, 18(2), 416.
[9] Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing, 13(3), 1195–1215.
[10] Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, 56–76.
[11] [Russell, S., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson Education.
[12] TensorFlow Team. (2024). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://tensorflow.org/

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)