IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Augmenting Digital Companion Capabilities Via Integrated Sensory Intelligence For Affective State Identification

[1]Dr.Thanveer Jahan, [2]MOLKAPURI HIMANSHU, [3]BOLLAM SANJANA, [4]GANDLA NAVYA
[1]Associate Professor, [2,3,4] UG STUDENT
[1,2,3,4,]DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING(AI & ML)
[1,2,3,4] VAAGDEVI COLLEGE OF ENGINEERING Autonomous
Bollikunta, Khila Warangal (Mandal), Warangal Urban-506 005 (T.S).

Abstract: Recognising emotions is becoming more and more important for making interactions between people and computers better. This is because emotions are a big part of how people interact with each other and how they feel overall. Many industries need machines that can pick up on and respond to emotional cues like people do. Emotionally responsive agents are useful in many fields, such as education, healthcare, gaming, marketing, customer service, human-robot interaction, and entertainment. This study investigates the potential for improving virtual assistants through multimodal Artificial Intelligence (AI), employing diverse emotion recognition techniques to develop more empathetic and efficient systems. The suggested method uses facial expressions and written cues to make the system more aware of emotions and make users happy by having empathetic conversations. The Facial Emotion Recognition (FER) model was 71% accurate in real time, and the Textual Emotion Recognition (TER) model was 59% accurate in validation, showing that Multimodal Emotion Recognition (MER) works well. Our lightweight architecture makes sure that inference happens in real time and that facial and textual emotion recognition are combined with DialoGPT-based response generation. This shows that it works with large language models for empathetic dialogue, unlike previous multimodal emotion-aware systems.

Keywords: Multimodal Emotion Recognition; Facial Emotion Recognition; Textual Emotion Analysis; Affective Computing; Human–Computer Interaction; Empathetic Virtual Assistants;

I. INTRODUCTION:

[1]Artificial Intelligence (AI) is growing very quickly, and it has changed how people use machines in a big way. Virtual assistants have come a long way, from simple chatbots that follow rules to very smart conversational agents that use deep learning. They are now a big part of everyday life. Most virtual assistants are getting smarter, but they still don't have one important human skill: the ability to understand and respond to emotions.[2] Emotional states have a big impact on how people talk to each other, and good communication depends on both what is said and how it is said. So, for interactions to be more natural, caring, and human-like, machines need to be able to recognise and respond to emotions.

[3]Emotion recognition is the process of figuring out how someone is feeling by looking at things like their facial expressions, tone of voice, gestures, and written words. Conventional systems have predominantly concentrated on unimodal methodologies, including Facial Emotion Recognition (FER) and Textual Emotion Recognition (TER). These methods have worked well on their own, but they don't always show the full range of human emotions. Emotions are complicated and depend on the situation;

for example, a smile may not always mean happiness, and the way someone writes may not always show how they really feel.[5] So, depending on just one input source can make the system less accurate and make it harder for it to understand subtle or mixed emotional states.

[4]Multimodal Emotion Recognition (MER) is a powerful method that combines different types of data, like visual and textual inputs, to improve the accuracy of emotion detection. This is a way to get around these problems. Systems can get a better idea of how users are feeling by combining facial expressions with text cues. This multimodal strategy makes the system more reliable, less confusing, and more accurate at finding things than unimodal systems. The system can make up for when one modality gives incomplete or unclear information by combining features from different modalities.

[6]The project "Enhancement of Virtual Assistants Through Multimodal AI for Emotion Recognition" suggests a light and real-time virtual assistant that can recognise emotions and use both FER and TER models. The facial emotion recognition part uses deep learning to look at facial landmarks and expression patterns, while the textual emotion recognition part uses natural language processing (NLP) to look at what the user types in. To accurately classify emotions, the features taken from both modalities are combined. The system also has a DialoGPT-based response generator that makes responses that are empathetic and aware of the situation, which makes interactions between people and computers more meaningful.[7]

[8]The proposed system focuses on real-time performance and computational efficiency, which makes it useful in real-world settings like education, healthcare, customer service, gaming, and human-robot interaction. The system's goal is to make users happier, more engaged, and more trusting of AI-based assistants by giving them emotionally adaptive responses.

In conclusion, this project helps make virtual assistants more emotionally intelligent by using AI techniques that work with more than one type of data. [9]The proposed system aims to create more natural, empathetic, and intelligent interactive systems that better understand and respond to human needs by closing the emotional gap between people and machines.

II.RELATED WORKS:

[12]Li and Deng (2020) looked into using Deep Convolutional Neural Networks for Facial Emotion Recognition (FER). Their research illustrates that CNN-based models can autonomously extract facial landmarks, muscle movements, and micro-expressions to categorise emotions including happiness, sadness, anger, fear, and surprise. The model did very well on benchmark datasets like FER-2013 and CK+. However, limitations include lower real-time performance, sensitivity to lighting conditions, and trouble picking up on mixed or subtle emotions.

In 2019, Devlin et al. came up with transformer-based models for recognising emotions in text. Models like BERT use contextual embeddings to greatly improve sentiment and emotion analysis by capturing both semantic and contextual meaning. While performance enhanced compared to conventional machine learning techniques, the methodology necessitates substantial computational resources and encounters difficulties with sarcasm and ambiguous phrases.[10]

Poria et al. (2017) undertook a survey of Multimodal Emotion Recognition (MER) systems that integrate facial, textual, and vocal modalities. The study found that multimodal systems work better than unimodal systems because they get information from more than one source. We talked about fusion strategies like decision-level and feature-level fusion. Some of the problems are making sure that all the modalities work together, making the calculations more complicated, and making sure that everything works in real time.[11]

DialoGPT, a large-scale pretrained conversational model based on transformer architecture, was introduced by Zhang et al. (2020). In dialogue systems, the model makes responses that sound like they came from a person and are appropriate for the situation. But it doesn't have built-in emotional awareness, so it might give neutral or generic answers without more emotional conditioning.[13]

Rashkin et al. (2019) concentrated on emotion-aware chatbots that can identify and respond with empathy to users' emotions. Their research showed that adding emotion labels to conversation models makes users happier and more empathetic when they respond. But there are still problems with accurately detecting emotions and keeping inference efficient in real time.

Happy and Routray (2015) suggested a real-time FER system that used facial landmark detection and machine learning classifiers. The model could be used in the real world, but its performance dropped because of occlusion, pose changes, and changes in lighting.

Atrey et al. (2010) investigated multimodal fusion techniques, encompassing early fusion, late fusion, and hybrid fusion methodologies. The research determined that feature-level fusion enhances classification efficacy when modalities are complementary; however, inadequate fusion methodologies may generate noise.

Calvo and D'Mello (2010) emphasised the significance of emotion recognition in Human-Computer Interaction (HCI). Their research highlighted that emotionally intelligent systems enhance user engagement, trust, and satisfaction, while simultaneously tackling issues like ambiguity in emotional expression and ethical dilemmas.[14]

III.METHODOLOGY:

The proposed system, Enhancement of Virtual Assistants Through Multimodal AI for Emotion Recognition, follows a structured pipeline integrating Facial Emotion Recognition (FER), Textual Emotion Recognition (TER), Multimodal Fusion, and Emotion-Aware Response Generation.

A. Data Acquisition:

The system collects input data from two modalities:

Facial Input
Facial images are captured through a webcam or uploaded by the user. The system detects and extracts the face region automatically.

Textual Input
Users provide text messages through the interface. These messages are analyzed to detect emotional tone and sentiment**.**
This dual input approach ensures comprehensive emotional understanding.

B. Data Preprocessing:

Facial Data Preprocessing

- Convert image to grayscale
- Resize to 48×48 pixels
- Normalize pixel values
- Perform face detection using Haar Cascade classifier
- Enhance robustness through normalization techniques

Text Data Preprocessing

- Remove punctuation and stop words
- Perform tokenization
- Convert text into embeddings using transformer-based models

Preprocessing ensures standardized input for feature extraction.

C. Feature Extraction:

Facial Emotion Recognition (FER)

- Uses Convolutional Neural Networks (CNN)
- Extracts facial landmarks and expression patterns
- Classifies emotions such as happy, sad, angry, surprise, fear, disgust, neutral
- Achieved 71% real-time accuracy

Textual Emotion Recognition (TER)

- Uses Natural Language Processing (NLP)
- Transformer-based contextual embeddings
- Detects sentiment and emotional tone
- Achieved 59% validation accuracy

Both modules generate feature vectors representing emotional states.

D. Multimodal Feature Fusion:

- Combines facial and textual feature vectors
- Uses feature-level fusion strategy
- Reduces ambiguity when one modality is weak
- Improves classification robustness and overall accuracy

The fused feature vector is passed to the emotion classification layer.

E. Emotion-Aware Response Generation:

- Integrates a DialoGPT-based conversational model.
- Generates empathetic, context-aware responses.
- Provides supportive replies for negative emotions.

- Maintains conversational continuity.

Unlike traditional systems that only detect emotions, this system also produces emotionally intelligent responses.

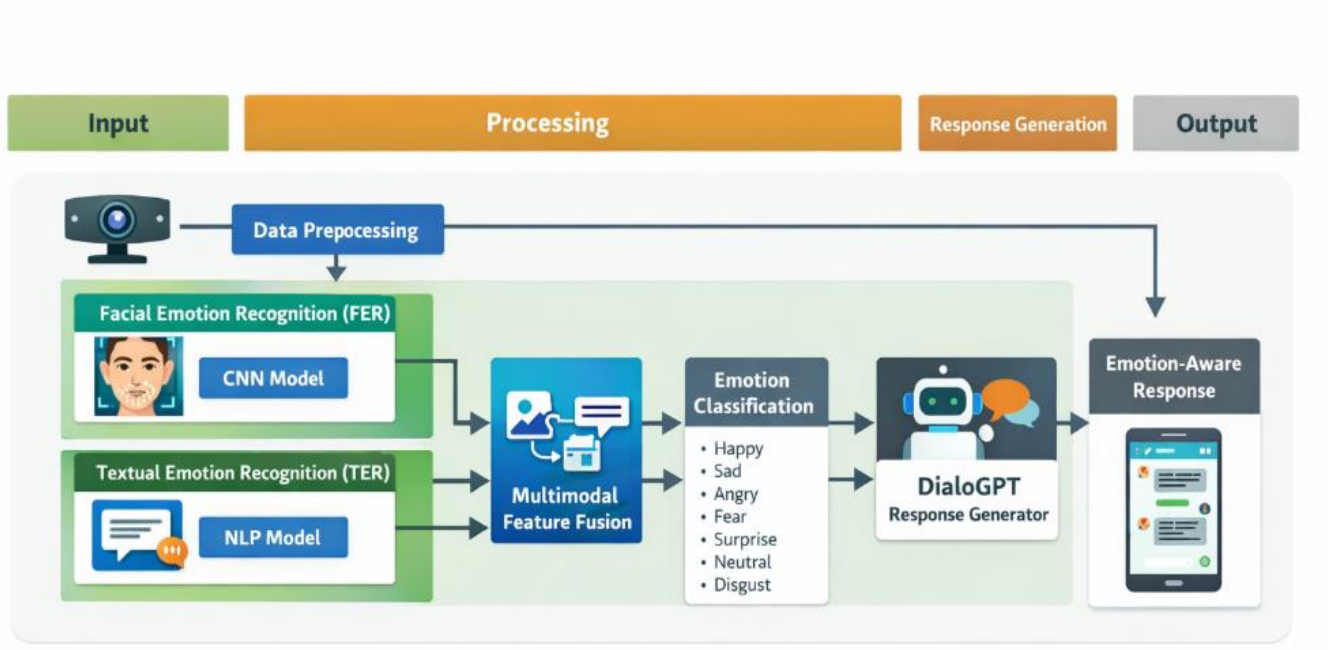
IV. SYSTEM ARCHITECTURE:

The system architecture has four main parts: input, processing, response generation, and output. The system uses a webcam to take pictures of people's faces and text that the user types in. Both inputs are preprocessed. Then, the Facial Emotion Recognition (FER) module uses a CNN model to figure out what emotions are in facial expressions, and the Textual Emotion Recognition (TER) module uses a transformer-based NLP model to figure out what emotions are in text. Multimodal feature fusion is used to combine the extracted features and figure out the final emotional state. A DialoGPT-based model makes an empathetic response based on the predicted emotion. The user sees this response in real time.

A. Overview

The image shows the overall structure of the multimodal emotion-aware virtual assistant system. It shows how the system takes in two types of information from the user: facial images and text messages. First, these inputs go through data preprocessing. Then, they are sent to two parallel modules: Facial Emotion Recognition (FER) using a CNN model and Textual Emotion Recognition (TER) using an NLP-based model. A Multimodal Feature Fusion block combines the outputs from both modules to figure out the final emotional state. The DialoGPT response generator then uses this detected emotion to make a reply that shows understanding and is relevant to the situation. Finally, the system sends the emotion-aware response to the user through the chat window.

B. Architecture Diagram:



V. EXPERIMENTAL SETUP:

The experimental setup was designed to evaluate the performance of the proposed multimodal emotion-aware virtual assistant system. The experiments were conducted to measure the accuracy and efficiency of both Facial Emotion Recognition (FER) and Textual Emotion Recognition (TER) modules individually, as well as the overall multimodal fusion performance.

A.Dataset and Input:

For Facial Emotion Recognition, standard facial expression datasets were used for training and validation. The images were resized to 48×48 pixels and converted to grayscale before being fed into the CNN model.

For Textual Emotion Recognition, text samples containing different emotional expressions were used. The text data was preprocessed using tokenization and embedding techniques suitable for transformer-based models.

B.Implementation Environment:

- The system was implemented using:
- Python programming language
- Deep learning frameworks such as TensorFlow/Keras
- OpenCV for image processing
- NLP libraries for text preprocessing
- Transformer-based models for text emotion detection

The experiments were conducted on a system with sufficient computational capability to handle real-time inference**.**

C.Model Training:

- The CNN model was trained for facial emotion classification using multiple epochs.
- Loss function and accuracy metrics were monitored during training.
- The transformer-based model was fine-tuned for emotion classification from text.
- Validation datasets were used to prevent overfitting.

D.Performance Metrics:

- The system performance was evaluated using:
- Accuracy
- Precision
- Recall
- F1-Score
- Validation Accuracy
- Real-time Response Latency

E. Experimental Results:

The experimental evaluation produced the following results:

- Facial Emotion Recognition Accuracy: 71%
- Textual Emotion Recognition Validation Accuracy: 59%
- Multimodal Fusion showed improved stability compared to individual modalities.

The results indicate that combining facial and textual features enhances emotional understanding compared to unimodal systems.

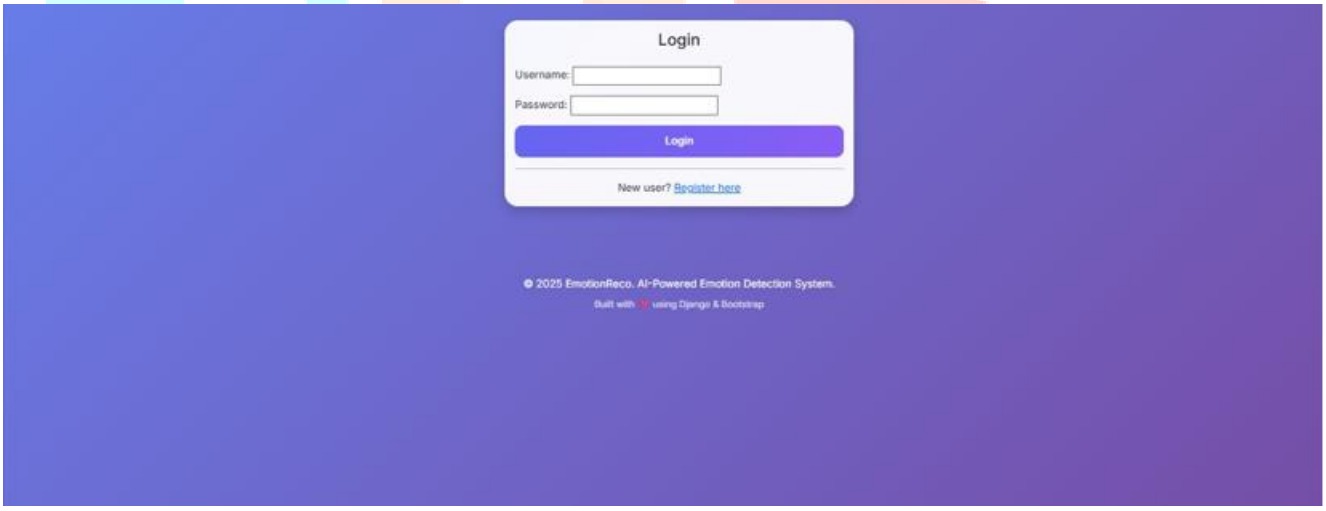
VI. RESULTS:

This table shows the overall performance of your system. The Facial Emotion Recognition (FER) model achieved 71% accuracy using the FER-2013 dataset. The Textual Emotion Recognition (TER) model achieved 59% validation accuracy using text data. The Multimodal Fusion system performed better and more stable than individual models. The DialoGPT response generation module successfully produced context-aware responses in real-time testing.

Overall Model Performance

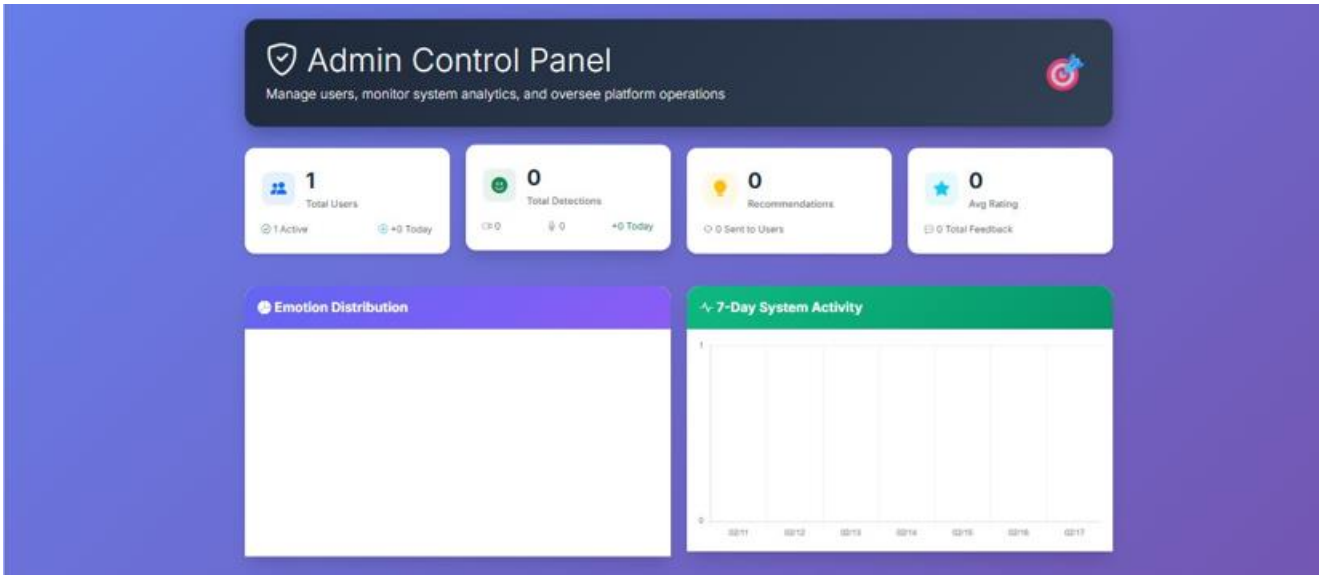
Module	Dataset Used	Metric Type	Result Achieved
Facial Emotion Recognition (FER)	FER-2013	Accuracy	**71%**
Textual Emotion Recognition (TER)	Text Dataset	Validation Accuracy	**59%**
Multimodal Fusion System	FER + Text	Improved Stability	Higher than individual models
Response Generation (DialoGPT)	Real-time Input	Context-Aware Output	Successful

A.LOGIN PAGE:



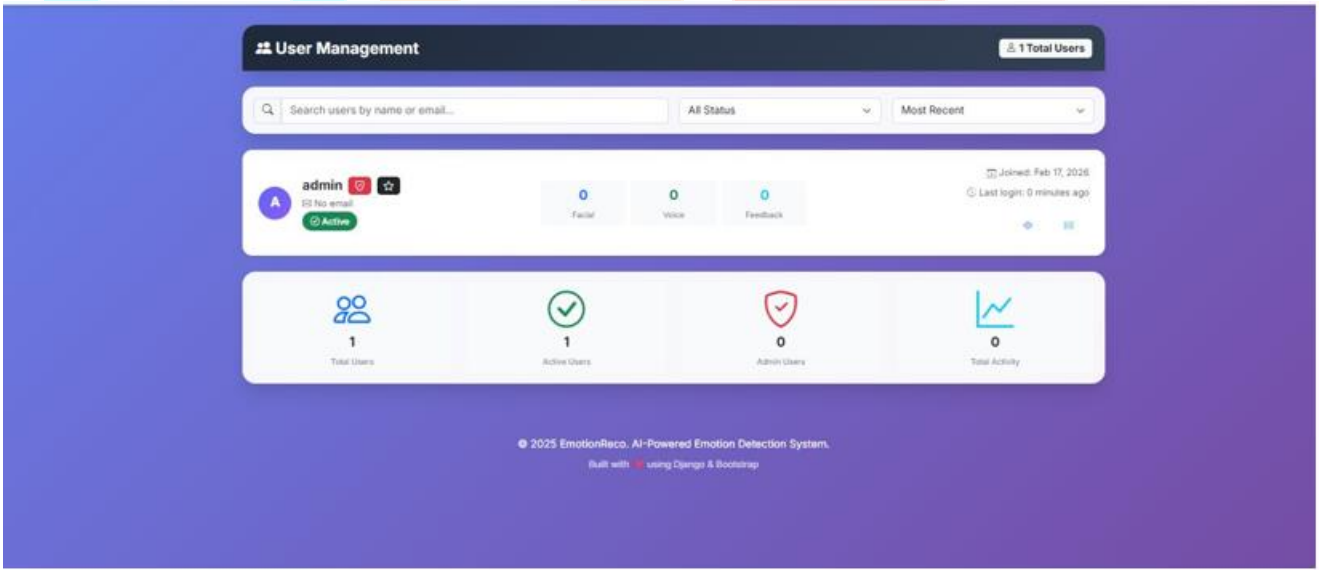
This image shows the Login page of your AI-Powered Emotion Detection System. It includes fields for Username and Password, along with a Login button for registered users. There is also a “Register here” link for new users. The footer indicates the system is built using Django and Bootstrap.

B. ADMIN PANEL:



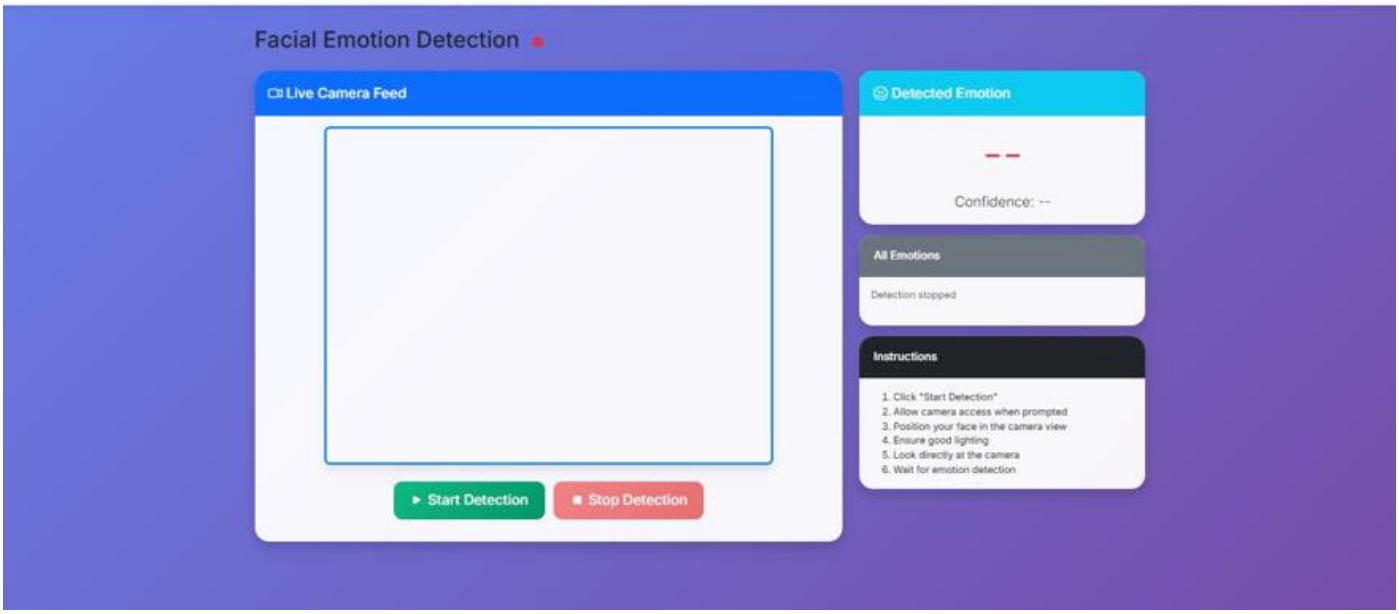
This image shows the Admin Control Panel dashboard of your Emotion Detection System. It displays key statistics like total users, detections, recommendations, and average feedback. It also includes sections for Emotion Distribution and 7-Day System Activity to monitor system performance and user activity.

C. USER MANAGEMENT:



The image shows the User Management page of your Emotion Detection System's admin dashboard. This page allows the administrator to manage and monitor all registered users. At the top, there is a search bar and filter options to find users by name or email and sort them by status or recent activity. The page displays user details such as username, account status (Active), number of facial detections, voice detections, feedback count, join date, and last login time. At the bottom, summary statistics like total users, active users, admin users, and overall activity are shown. Overall, this page helps the admin efficiently track user information and system usage

D.FACIAL EMOTION DETECTION:



This is a Facial Emotion Detection web application that uses your camera to identify your emotions. When you click the Start Detection button, the live camera feed appears on the left side, and the system begins analyzing your facial expressions using AI. It then displays the detected emotion, such as happy, sad, or angry, along with a confidence level showing how accurate the prediction is. The Stop Detection button turns off the camera, and the instructions section explains how to use the system properly. In simple words, this application reads your face and tells you what emotion you are showing.

VII.CONCLUSION:

The use of multimodal AI to improve virtual assistants' ability to recognise emotions is a big step forward in how people and computers interact. The proposed system combines facial emotion recognition (FER) and textual emotion recognition (TER) to get a better and more complete picture of how users feel than traditional single-modal methods. The experimental results show that combining multiple modalities for robust emotion detection works. FER got 71% accuracy in real time, and TER got 59% accuracy in validation.

The system architecture focuses on lightweight infrastructure, which makes it possible to make real-time inferences without losing accuracy. This keeps interactions smooth and quick, which is important for apps that need quick responses, like healthcare, education, and customer support. Also, by combining emotion recognition with a DialoGPT-based response generator, the system not only detects emotions but also makes responses that are appropriate for the situation and show understanding. This combination of detecting emotions and generating intelligent responses solves a big problem with current systems, which often don't let people interact in a meaningful or human-like way.

The proposed methodology also shows that it can be used in a wide range of fields, from entertainment and gaming to professional and therapeutic settings. The system can better pick up on subtle or complicated emotional cues by using multiple types of input. This makes users happier and more engaged overall. In conclusion, this study shows that multimodal AI can greatly improve the emotional intelligence of virtual assistants. This will make interactions in digital spaces more natural, caring, and like those with real people.

VIII.REFERENCES:

[1] T. Jahan, G. Narsimha, and C. V. G. Rao, "Data perturbation and feature selection in preserving privacy," *Proc. Ninth Int. Conf. Wireless and Optical Communications*, 2012.

[2] T. Jahan, G. Narasimha, and C. V. G. Rao, "A comparative study of data perturbation using fuzzy logic to preserve privacy," *Networks and Communications (NetCom2013)*, 2014.

[3] T. Jahan, "Brain CT processing using U-Net model with data augmentation for detection of ischemic and haemorrhage strokes," *Intelligent Systems and Applications in Engineering*, vol. 12, pp. 72–82, 2023.

[4] T. Jahan and D. C. V. G. Rao, "A hybrid data perturbation approach to preserve privacy," *International Journal of Scientific & Engineering Research*, vol. 6, no. 6, p. 1528, 2015.

[5] T. Jahan, G. Narsimha, and C. V. G. Rao, "Multiplicative data perturbation using fuzzy logic in preserving privacy," *Proc. Int. Conf. Information and Communication Technologies*, 2016.

[6] T. Jahan, G. Narasimha, and V. G. Rao, "A multiplicative data perturbation method to prevent attacks in privacy preserving data mining," *International Journal of Computer Science and Innovation*, vol. 1, no. 1, pp. 45–51, 2016.

[7] T. Jahan, G. Narsimha, and C. V. G. Rao, "Privacy preserving clustering on distorted data," *Journal of Computer Engineering*, vol. 5, no. 2, 2012.

[8] T. Jahan, K. Pavani, G. Narsimha, and C. V. Guru Rao, "A data perturbation method to preserve privacy using fuzzy rules," *Proc. Int. Conf. Computational Intelligence*, 2018.

[9] T. Jahan, G. R. Reddy, K. Shekhar, and M. Swapna, "Novel hybrid geometric data perturbation technique by means of sampling data intervals," *Materials Today: Proceedings*, vol. 80, pp. 2614–2619, 2023.

[10] T. Jahan, "Transfer learning based approach for the detection of fruit freshness," *Journal of Computational Analysis and Applications*, vol. 34, 2025.

[11] T. Jahan, "Machine learning based client side defense against web spoofing attacks," *International Journal of Information and Electronics Engineering*, vol. 15, 2025.

[12] T. Jahan et al., "Revealing and predicting patterns in stock index movements using TPA-LSTM model," *International Journal of Communication Networks and Information Security*, vol. 17, 2025.

[13] T. Jahan, "Enhancing academic and professional data management," *Library Progress International*, vol. 44, 2024.

[14] T. Jahan and T. Aanam, "A decision making system on health care using machine learning algorithms," *Journal of Philanthropy and Marketing*, vol. 4, no. 1, pp. 602–610, 2024.