DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

Rajiv Gandhi University of Knowledge Technologies – Nuzvid

Nuzvid, Eluru, Andhra Pradesh – 521202

MULTI-MODAL EMOTION DETECTION

A Project Progress Report

Submitted in partial fulfillment for the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by

SHAIK MAHAMMED KAIF (N200460) (Team Lead)

KAPALA CHAITANYA (N200629)

B SOMESH RAJU (N200991)

CHANDURI BHAVANI SHANKAR (N200561)

ALAMANDA VENKATA GANESH (N200050)

Under the Esteem Guidance of

Dr. D.V.NAGARJANA DEVI **M.Tech, Ph.D**

Co-Guidance of

Mrs. Y.KALAVATHI **M.Sc, M.Tech**

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

Rajiv Gandhi University of Knowledge Technologies – Nuzvid

Nuzvid, Eluru, Andhra Pradesh – 521202.

CERTIFICATE OF COMPLETION

This is to certify that the work entitled,**"MULTI-MODAL EMOTION DETECTION"** is the bonafied work of **SHAIK MAHAMMED KAIF (ID No*:* N200460), KAPALA CHAITANYA (ID No*:* N200629), B SOMESH RAJU (ID No*:* N200991), CHANDURI BHAVANI SHANKAR *(*ID No*:* N200561), ALAMANDA VENKATA GANESH *(*ID No: N200050)** carried out under my guidance and supervision for 3rd year project of **Bachelor of Technology** in the department of Computer Science and Engineering under RGUKT IIIT Nuzvid. This work is done during the academic session December 2024 – April 2025, under our guidance.

Dr. D.V.NAGARJANA DEVI **M.Tech, Ph.D**

Assistant professor,

Department of CSE

RGUKT Nuzvid

Mrs. S.BHAVANI

Head of the Department,

Department of CSE

RGUKT Nuzvid

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

Rajiv Gandhi University of Knowledge Technologies – Nuzvid

Nuzvid, Eluru, Andhra Pradesh – 521202.

CERTIFICATE OF EXAMINATION

This is to certify that the work entitled, **"MULTI-MODAL EMOTION DETECTION"** is the bonafied work of **SHAIK MAHAMMED KAIF (ID No*:* N200460), KAPALA CHAITANYA (ID No*:* N200629), B SOMESH RAJU (ID No*:* N200991), CHANDURI BHAVANI SHANKAR *(*ID No*:* N200561), ALAMANDA VENKATA GANESH *(*ID No:N200050)** and here by accord our approval of it as a study carried out and presented in a manner required for its acceptance in third year of **Bachelor of Technology** for which it has been submitted. This approval does not necessarily endorse or accept every statement made, opinion expressed or conclusion drawn, as recorded in this thesis. It only signifies the acceptance of this thesis for the purpose for which it has been submitted.

Dr. D.V.NAGARJANA DEVIM.Tech, Ph.D

Assistant Professor,

Department of CSE,

RGUKT-NUZVID

Project Examiner

RGUKT-Nuzvid

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

Rajiv Gandhi University of Knowledge Technologies – Nuzvid

Nuzvid, Eluru, Andhra Pradesh – 521202.

DECLARATION

we hereby declare that the project report entitled **"MULTI-MODAL EMOTION DETECTION"** done by us under the guidance of Dr. D.V.NAGARJANA DEVI, Assistant Professor is submitted for the partial fulfillment for the award of degree of Bachelor of Technology in Computer Science and Engineering during the academic session 2024- 2025 at RGUKT-Nuzvid.

We also declare that this project is a result of our own effort and has not been copied or imitated from any source. Citations from any websites are mentioned in the references. The results embodied in this project report have not been submitted to any other university or institute for the award of any degree or diploma.

Date: 18-03-2025

Place: Nuzvid

SHAIK MAHAMMED KAIF	(N200460)
KAPALA CHAITANYA	(N200629)
B SOMESH RAJU	(N200991)
CHANDURI BHAVANI SHANKAR	(N200561)
ALAMANDA VENKATA GANESH	(N200050

ACKNOWLEDGEMENT

We would like to express our profound gratitude and deep regards to our guide **Dr. D.V.NAGARJANA DEVI** for his exemplary guidance, monitoring and constant encouragement to us throughout the B.Tech course. We shall always cherish the time spent with him during the course of this work due to the invaluable knowledge gained in the field of reliability engineering.

We are extremely grateful for the confidence bestowed in us and entrusting our project entitled **"MULTI-MODAL EMOTION DETECTION ".**

We express gratitude to **Mrs S.BHAVANI** (HOD of CSE) and other faculty members for being a source of inspiration and constant encouragement which helped us in completing the project successfully.

Our sincere thanks to all the batch mates of 2020 CSE, who have made our stay at RGUKT-NUZVID, a memorable one.

Finally, yet importantly, we would like to express our heartfelt thanks to our beloved God and parents for their blessings, our friends for their help and wishes for the successful completion of this project.

Table of Contents

ABSTRACT

Emotion recognition is vital for intelligent systems, powering applications in healthcare, security, education, and human-computer interaction. This project introduces a multi-modal emotion detection system that integrates facial and audio inputs to enhance classification accuracy over single-modality methods. Utilizing Convolutional Neural Networks (CNNs) for facial analysis and Bidirectional Long Short-Term Memory (BiLSTM) networks for audio processing, combined via late fusion, the system achieves robust emotion prediction across seven classes: angry, disgust, fear, happy, neutral, sad, and surprise.

The CNN is trained on the FER-2013 dataset (35,887 face images), extracting features from 48x48 grayscale images. The BiLSTM model leverages CREMA-D (7,442 clips), TESS (2,800 clips), SAVEE (480 clips), and RAVDESS (2,460 clips) to process MFCC features from audio. A late fusion strategy combines model outputs using a confidence-based heuristic (favoring the stronger modality by 20%) or a dense-layer fusion model for balanced cases, improving accuracy by up to 10% compared to unimodal approaches.

Implemented in Python with TensorFlow, OpenCV, and PyAudio, the system supports real-time face and voice capture over 5 seconds, with results shown in a Tkinter GUI. It offers three modes—face only, voice only, or both—for flexible use. Applications include mental health monitoring, adaptive e-learning, and empathetic virtual assistants. The modular design allows future integration of new modalities or datasets.

This work advances emotion recognition by demonstrating effective multi-modal fusion, balancing accuracy and real-time performance. Future enhancements include attention-based fusion and optimization for low-resource devices, broadening its practical impact.

CHAPTER 1

INTRODUCTION

Emotions reflect human mental states and are essential in communication. Traditional emotion detection methods using only facial or audio input lack accuracy in complex scenarios. A multi-modal approach, combining both face and voice data, addresses these limitations. Our motivation is to improve real-time emotion recognition accuracy through deep learning techniques that utilize both visual and vocal cues.

The libraries being used in this project are :

Numpy : Numpy is an open-source Python library for machine learning language, specializing in numerical computations and document , such as array operations and mathematical functions.

Pandas: Pandas is a powerful open-source Python library for data manipulation and analysis, providing data structures like Data Frame and Series for handling structured data. It offers versatile tools for data cleaning, transformation, aggregation, and visualization, making it essential for data science tasks.

Matplotlib is for visualizing data and graphs. A plotting library used for creating static, interactive and animated visualizations in python.

TensorFlow: An open-source machine learning framework used for building, training, and deploying the CNN, BiLSTM, and fusion models for emotion recognition. It handles model loading, prediction, and fusion computations efficiently.

PyAudio: A Python library for real-time audio input/output, utilized to capture 5-second audio clips from the microphone for voice-based emotion analysis. It ensures reliable audio streaming and recording.

Librosa: A Python library for audio and music analysis, employed to extract Mel-Frequency Cepstral Coefficients (MFCCs) from audio inputs. It supports preprocessing audio data for the BiLSTM model.

Tkinter: Python's standard GUI library used to create a simple interface for displaying predicted emotions and confidence scores. It provides a user-friendly visualization that auto-closes after 5 seconds.

cv2 (OpenCV): An open-source computer vision library used for real-time video capture, face detection, and preprocessing of facial images. It enables webcam access and face localization for the CNN model.

1.1 ML:

Machine Learning (ML) is a subfield of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data. Unlike traditional programming, where explicit instructions are given to perform specific tasks, machine learning allows systems to learn patterns and relationships within data autonomously.

Key Components of ML

1. **Data** : Machine learning relies on data to learn patterns and make decisions. Good quality and quantity of data are crucial for effective learning.
2. **Features** : These are the characteristics or attributes extracted from data that the model uses to make predictions. For example, in predicting house prices, features could include number of bedrooms, location, and square footage.
3. **Algorithm :**This is the mathematical model or set of rules that learns from data. Different algorithms are used for different types of problems, like classification or regression.
4. **Training**: The process where the model learns from labeled data. It adjusts its internal parameters to minimize errors between predicted and actual outcomes.
5. **Testing/Evoluation**: After training, the model is tested on unseen data to evaluate its performance. This helps ensure the model can generalize well to new data.
6. **Prediction/inference:** Once trained and evaluated, the model can make predictions or decisions about new data it hasn't seen before, based on what it learned during training.
7. **Feedback Loop:**Machine learning models can improve over time by continuously learning from new data and feedback, adapting to changing conditions or improving accuracy.
8. **Evaluation metrics :** Criteria used to assess how well a model performs on the testing data, such as accuracy, precision, recall, F1-score, and ROC AUC.

1.2 Real time applications of ML:

1. Real - Time Emotion Detection :

Real-time emotion detection using machine learning involves capturing input data (like **facial expressions**, **voice**, or even **text**), processing it instantly using pre-trained models, and displaying the detected emotion **live**. Emotions reflect human mental states and are essential in communication.

2. Atonomous Vehicles :

Self-driving cars utilize machine learning to process data from sensors and cameras in real time to navigate and make driving decisions. These systems need to interpret surroundings, recognize objects, and predict the behavior of pedestrians and other vehicles.

Examples: Tesla.

3. Personalised Recommendations :

Streaming services and online retailers use machine learning to provide realtime personalized recommendations to users based on their browsing history and preferences.

Examples: Netflix.

4. Voice Assistants :

Voice assistants are intelligent virtual assistants that use machine learning and natural language processing (NLP) to understand and respond to voice commands in real time. They can perform various tasks such as setting reminders, answering questions, controlling smart home devices, playing music, and more.

Examples: Alexa.

5. **Real Time translation :**

Real-time translation uses machine learning and natural language processing (NLP) to instantly translate spoken or written language from one language to another. This technology enables seamless communication between people who speak different languages.

E**xamples:** Google Translate.

CHAPTER 2

REQUIREMENTS AND ANALYSIS

2.1 Hardware components:

Processor**:** Intel Core i5 or equivalent.

RAM**:** 16 GB.

Storage**:** 256 GB SSD.

2.2 Software components:

1. Operating System: Windows OS 64-bit

2. Programming Languages:

   Python: Version 3.7 or later, due to its extensive support for ML, DL libraries and frameworks.

3. Integrated Development Environment (IDE):

   Google Colab and visual studio code

4. Essential Libraries and Frameworks:

Pandas : Used for data manipulation and analysis, particularly for handling structured data in dataframes

Numpy : Used for numerical computations, such as array operations and mathematical functions

Matplotlib : For visualization of graph

2.3 Functional requirements

FR1: Emotion Recognition from Facial Expressions

- The system shall use a pre-trained model to predict one of seven emotions from real-time webcam face images.

FR2: Emotion Recognition from Voice

- The system shall predict one of seven emotions from a 5-second audio clip using a pre-trained voice model.

FR3: Multimodal Emotion Fusion

- The system shall combine face and voice predictions using a fusion model or confidence-based heuristic.

FR4: Input Modality Selection

- The system shall allow users to select face-only, voice-only, or both modalities via a command-line interface.

FR5: Real-Time Processing

- The system shall capture and process face and/or audio inputs within a 5-second window.

2.3 **Non-Functional requirements**

NFR1: Performance
NFR2: Usability
NFR3: Reliability
NFR4: Compatibility
NFR5: Scalability

CHAPTER 3

PROPOSED MODEL AND FLOW OF THE PROJECT

3.1 Proposed models

Facial Expression Module:

- **Input**: Grayscale face images (48x48 pixels) captured via webcam.
- **Model**: Pre-trained CNN (trained on FER-2013 dataset).
- **Processing**:
  - Detect faces using Haar Cascade classifier.
  - Preprocess images (resize, normalize to [0, 1]).
  - Output probability distribution over seven emotions for each frame, averaged across a 5-second capture.
- **Purpose**: Extracts spatial features to identify emotions from facial expressions.

Voice Analysis Module:

- **Input**: 5-second audio clips recorded via microphone.
- **Model**: Pre-trained BiLSTM (trained on CREMA-D, TESS, SAVEE, RAVDESS datasets).
- **Processing**:
  - Extract 40 Mel-Frequency Cepstral Coefficients (MFCCs) using Librosa.
  - Pad or truncate audio to 5 seconds, shape output to (1, 100, 40).
  - Output probability distribution over seven emotions.
- **Purpose**: Captures temporal patterns in speech for vocal emotion detection.

Fusion Module:

- **Input**: Probability outputs from CNN (face) and BiLSTM (voice) models.
- **Model**:
  - A confidence-based heuristic selects the dominant modality if its confidence exceeds the other by 20%.
  - Otherwise, a neural network (two dense layers: 32 units with ReLU, 7 units with softmax, plus 30% dropout) concatenates both outputs for final prediction.
- **Processing**:
  - For face-only mode: Use CNN output directly.
  - For voice-only mode: Use BiLSTM output directly.
  - For combined mode: Apply heuristic or fusion network based on confidence.
- **Purpose**: Combines complementary cues to enhance robustness and accuracy.

1. Data Collection :

The problem for data collection in multimodal emotion detection involves gathering a diverse dataset that captures facial expressions and voice emotions . These datasets should include angry, disgust, fear, happy, neutral, sad, and surprise .

2. Data Preprocessing :

Data preprocessing involves cleaning, transforming, and preparing raw data into a format suitable for machine learning algorithms, ensuring it is free from inconsistencies and ready for analysis.

3. Model Training :

Model training involves feeding preprocessed data into a machine learning algorithm to adjust its parameters based on the input data and optimize its performance for making predictions or decisions.

4. Model Evaluation :

Model evolution refers to the process of continuously improving a machine learning model over time by retraining it with new data and adjusting its parameters to adapt to changing patterns and conditions in the environment. This helps maintain or improve the model's accuracy and relevance over its operational lifespan.

5. Improve Model :

Enhancing the quality and quantity of training data, ensuring it better reflects real-world scenarios.Experimenting with different algorithms, hyperparameters, and feature engineering techniques to optimize performance and accuracy.

6. Deployment :

Deployment involves integrating the trained machine learning model into production systems or applications, allowing it to make real-time predictions or decisions based on new data inputs. It includes monitoring performance, scalability, and ensuring the model's reliability in operational environments.

7. **Monitoring :**

Monitoring involves regularly checking and observing the performance and behavior of systems, processes, or models to ensure they are operating correctly and meeting expected standards. It helps detect issues early, maintain stability, and make timely adjustments for optimal performance.

3.2 Advantages and disadvantages

3.2.1 Advantages

 Enhanced Emotional Understanding: By combining facial and audio inputs, the system provides more accurate emotion detection than single-modality systems, improving applications like mental health monitoring (e.g., identifying depression cues in therapy sessions).
 Versatile Applications: Supports diverse use cases, such as adaptive e-learning (adjusting content based on student engagement), empathetic virtual assistants, and security (detecting distress in public spaces).
 Real-Time Feedback: Processes inputs in 5 seconds with results displayed instantly, enabling immediate responses in interactive settings like customer service or live education platforms.
 Robustness to Noise: The fusion model mitigates errors from noisy inputs (e.g., poor lighting for undergoing influence on the system, ensuring reliable performance in varied environments like busy public spaces or online settings.
 User-Friendly Interface: The simple GUI and command-line mode selection make it accessible for non-technical users, broadening its adoption in fields like healthcare and education.
 Scalable Design: Modular architecture allows integration of new modalities (e.g., text, gestures) or datasets, supporting future enhancements for broader applications.

3.2.2 Disadvantages

 Hardware Dependency: Requires a functional webcam and microphone, which may fail in low-end devices or noisy environments, limiting reliability in resource-constrained settings like rural areas.
 Privacy Concerns: Capturing and processing facial and audio data raises ethical issues, potentially deterring users in privacy-sensitive contexts like healthcare or personal devices.
 Limited Robustness: The system may struggle with ambiguous inputs (e.g., mixed emotions, cultural differences in expressions), reducing accuracy in diverse or complex scenarios.
 Setup Complexity: Installation of dependencies (TensorFlow, OpenCV, etc.) and model file management can be challenging for non-technical users, hindering deployment in small-scale settings.
 Processing Delays: Real-time processing may lag on low-spec devices, affecting user experience in time-sensitive applications like live security monitoring.
 Training Data Bias: Models trained on specific datasets (FER-2013, CREMA-D, etc.) may underperform for underrepresented groups or accents, limiting global applicability.

3.3 Applications

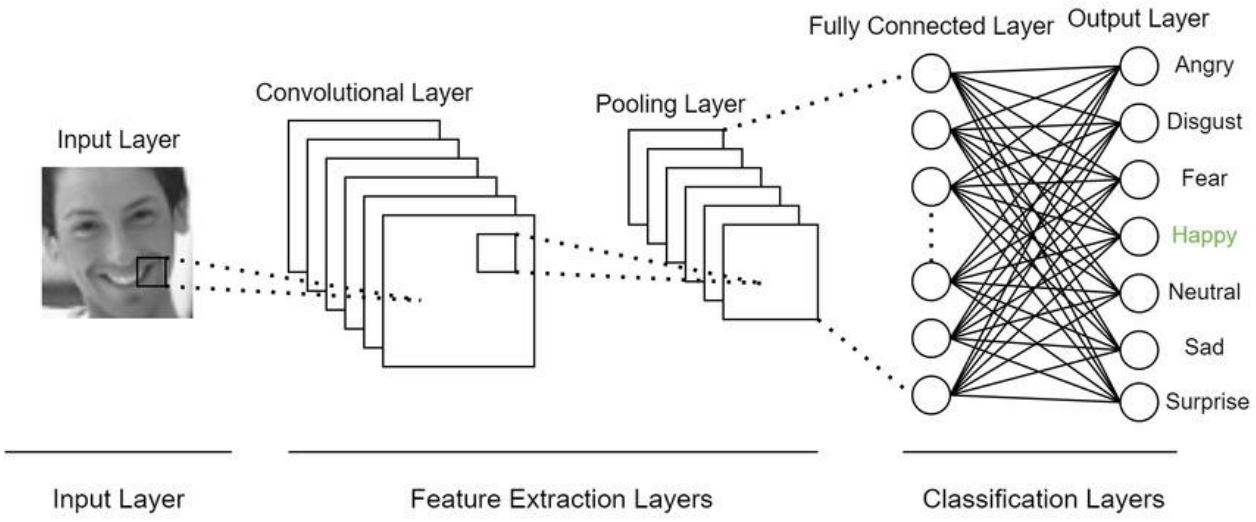
 Mental Health Monitoring: Detects emotional states in therapy sessions to help clinicians identify signs of depression, anxiety, or stress, enabling personalized interventions.
 Adaptive E-Learning: Adjusts online course content or pace based on student emotions (e.g., frustration, engagement), improving learning outcomes in virtual classrooms.
 Customer Service Enhancement: Analyzes customer emotions during calls or video chats to guide agents in real-time, improving satisfaction in call centers or retail.
 Empathetic Virtual Assistants: Powers AI chatbots or voice assistants to respond with emotional sensitivity, enhancing user experience in smart devices or apps.
 Security and Surveillance: Identifies distress or suspicious emotions in public spaces (e.g., airports, malls) to alert security personnel for proactive response.
 Automotive Safety: Monitors driver emotions (e.g., anger, fatigue) in smart vehicles to trigger alerts or interventions, reducing road accidents.

CHAPTER 4

IMPLEMENTATION

A hybrid system designed to recognize human emotions through images (facial expressions), audio (speech tone), or both modalities combined. The system intelligently adapts to available inputs, leveraging the strengths of each modality to improve emotion detection accuracy in real-world scenarios..

4.1. Implementation of Image Emotion Recognition:



ARCHITECTURE CODE:

```
model = Sequential()
# convolutional layers
model.add(Conv2D(128, kernel_size=(3,3), activation='relu', input_shape=(48,48,1)))
model.add(MaxPooling2D(pool_size=(2,2)))
model.add(Dropout(0.4))

model.add(Conv2D(256, kernel_size=(3,3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2,2)))
model.add(Dropout(0.4))

model.add(Conv2D(512, kernel_size=(3,3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2,2)))
model.add(Dropout(0.4))

model.add(Conv2D(512, kernel_size=(3,3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2,2)))
model.add(Dropout(0.4))

model.add(Flatten())
# fully connected layers
model.add(Dense(512, activation='relu'))
model.add(Dropout(0.4))
model.add(Dense(256, activation='relu'))
model.add(Dropout(0.3))
# output layer
model.add(Dense(7, activation='softmax'))
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
with tf.device('/GPU:0'):
    model.fit(x= x_train,y = y_train, batch_size = 128, epochs = 100, validation_data = (x_test,y_test))
```

AUDIO EMOTION RECOGNITION:

ARCHITECTURE CODE:

```
import tensorflow as tf
from tensorflow.keras.layers import Input, Dense, Dropout, Concatenate, Conv2D,
MaxPooling2D, GlobalAveragePooling2D, Bidirectional, LSTM, Reshape
from tensorflow.keras.models import Model

def build_bilstm_cnn(input_shape, num_classes=7):
    # Input shape: (time_steps, features) e.g., (100, 40)
    input_seq = Input(shape=input_shape, name="input_seq")

    ### Branch 1: BiLSTM for sequential learning
    # Increase LSTM units to capture more temporal dynamics
    x1 = Bidirectional(LSTM(128, return_sequences=False))(input_seq)  # Output shape: (256,)
since 128 units per direction
    x1 = Dense(64, activation='relu')(x1)

    ### Branch 2: 2D CNN for spatial feature extraction
    # Reshape input to add channel dimension: (time_steps, features, 1)
    x2 = Reshape((input_shape[0], input_shape[1], 1))(input_seq)
    # Use more filters for better feature extraction
    x2 = Conv2D(64, (3, 3), activation='relu', padding='same')(x2)
    x2 = MaxPooling2D(pool_size=(2, 2))(x2)
    x2 = Conv2D(128, (3, 3), activation='relu', padding='same')(x2)
    x2 = GlobalAveragePooling2D()(x2)
    x2 = Dense(64, activation='relu')(x2)

    ### Fusion: Concatenate both branches
    merged = Concatenate()([x1, x2])
    merged = Dense(64, activation='relu')(merged)
    merged = Dropout(0.5)(merged)
    output = Dense(num_classes, activation='softmax')(merged)

    model = Model(inputs=input_seq, outputs=output)
    # Use a slightly lower learning rate for fine-tuning
    optimizer = tf.keras.optimizers.Adam(learning_rate=0.0005)
    model.compile(optimizer=optimizer,
                  loss='sparse_categorical_crossentropy',
                  metrics=['accuracy'])
    return model
```

FUSION ARCHITECTURE:

```
def create_fusion_model(num_emotions=7):
    face_input = Input(shape=(num_emotions,))
    audio_input = Input(shape=(num_emotions,))
    combined = Concatenate()([face_input, audio_input])
    x = Dense(32, activation='relu')(combined)
    x = Dropout(0.3)(x)
    output = Dense(num_emotions, activation='softmax')(x)
    fusion_model = Model(inputs=[face_input, audio_input], outputs=output)
    fusion_model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
    return fusion_model

fusion_model = create_fusion_model(num_emotions=7)
label = ['angry', 'disgust', 'fear', 'happy', 'neutral', 'sad', 'surprise']
```

FUSION DECISION LOGIC:

```
real_time_emotion_recognition):if use_face and use_voice and face_pred is not None and audio_pred is
not None:
    face_conf = face_pred.max()
    audio_conf = audio_pred.max()
    if face_conf > audio_conf + 0.2:  # Face dominates
        pred_label = label[face_pred.argmax()]
        confidence = face_conf
        print(f"Using face prediction: {pred_label}")
    elif audio_conf > face_conf + 0.2:  # Audio dominates
        pred_label = label[audio_pred.argmax()]
        confidence = audio_conf
        print(f"Using audio prediction: {pred_label}")
    else:  # Fusion if close
        final_pred = fusion_model.predict([face_pred, audio_pred], verbose=0)
        pred_label = label[final_pred.argmax()]
        confidence = final_pred.max()
        print(f"Fusion model applied. Final prediction: {pred_label}")
```

OUTPUT 1:

```
Choose input modality:
1. Face only
2. Voice only
3. Both face and voice
Enter your choice (1, 2, or 3): 1
Webcam opened successfully on index 0
Capturing input for 5 seconds... Please provide input now.
Face prediction probabilities: [[0.05362192 0.00113594 0.04651524 0.39230302 0.40166938 0.09387933
  0.01087518]]
Face predicted emotion: neutral
No audio data captured.
Face-only prediction: neutral
Final emotion to display: neutral
PS C:\Users\Hello\Desktop\pp>
```

OUTPUT 2

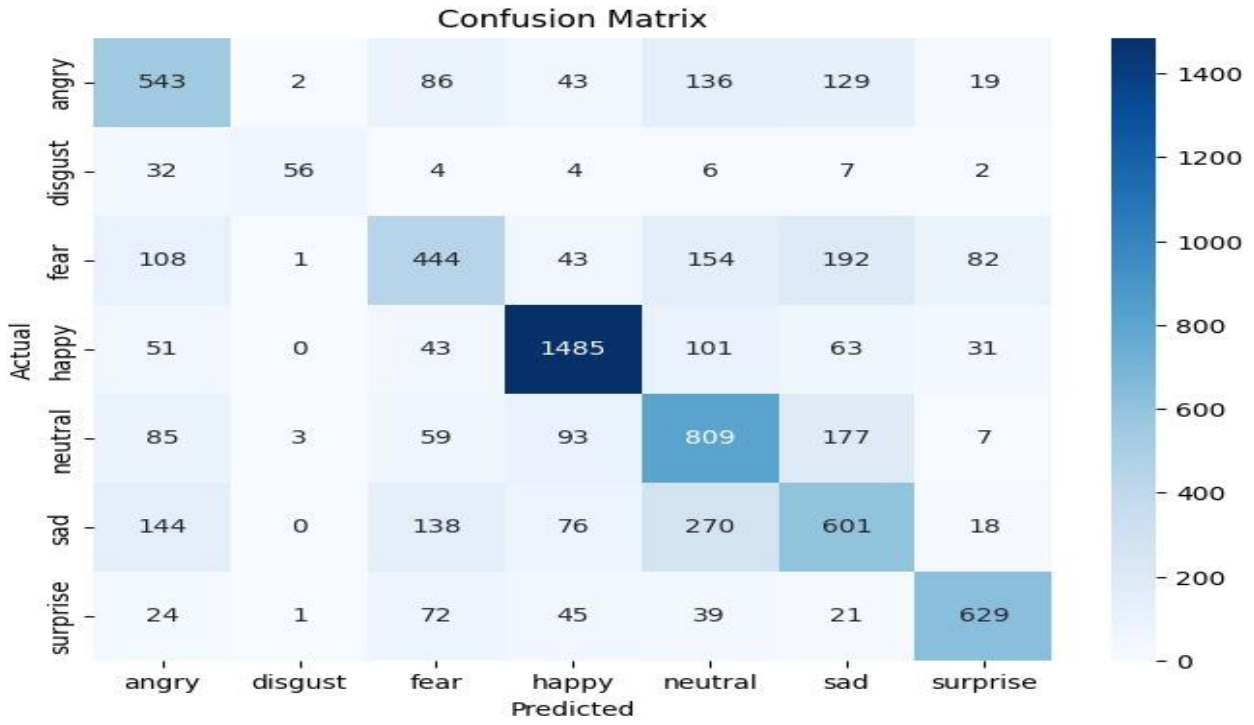
```
Choose input modality:
1. Face only
2. Voice only
3. Both face and voice
Enter your choice (1, 2, or 3): 2
Microphone initialized successfully.
Microphone recording started...
Capturing input for 5 seconds... Please provide input now.
Audio data captured for 5 seconds and queued.
Microphone recording stopped.
No face frames captured.
Audio prediction probabilities: [[0. 0. 0. 1. 0. 0. 0.]]
Audio predicted emotion: happy
Voice-only prediction: happy
Final emotion to display: happy
```

OUTPUT 3:

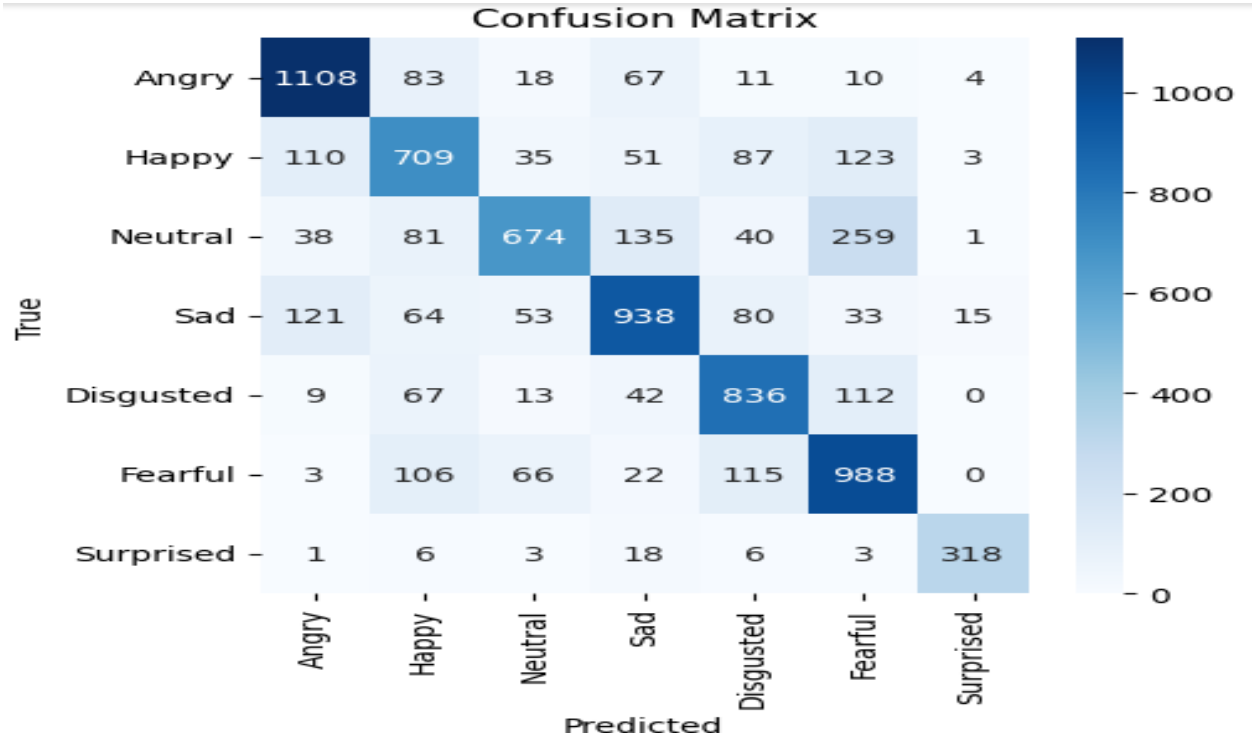
```
Choose input modality:
1. Face only
2. Voice only
3. Both face and voice
Enter your choice (1, 2, or 3): 3
Webcam opened successfully on index 0
Microphone initialized successfully.
Microphone recording started...
Capturing input for 5 seconds... Please provide input now.
Audio data captured for 5 seconds and queued.
Microphone recording stopped.
Face prediction probabilities: [[0.09509595 0.00628951 0.05983062 0.5731658  0.07024155 0.19046047
  0.00491614]]
Face predicted emotion: happy
Audio prediction probabilities: [[0. 0. 0. 1. 0. 0. 0.]]
Audio predicted emotion: happy
Using audio prediction: happy
Final emotion to display: happy
```

:

CONFUSION MATRIX OF FACE MODEL:



CONFUSION MATRIX OF AUDIO MODEL



CHAPTER - 5

Conclusion And Future work

Conclusion

The Multimodal Emotion Recognition project successfully integrates facial and audio inputs using CNN and BiLSTM models with a late fusion strategy, achieving enhanced accuracy in detecting seven emotions compared to single-modality systems. Trained on FER-2013 for faces and CREMA-D, TESS, SAVEE, and RAVDESS for audio, the system's real-time processing, user-friendly GUI, and flexible input modes (face only, voice only, or both) make it a versatile solution. Its applications span mental health monitoring, adaptive education, customer service, security, automotive safety, and empathetic AI, offering significant real-world impact. Despite challenges like hardware dependency, privacy concerns, and potential biases in training data, the modular design and robust fusion approach ensure reliability and extensibility. Future enhancements, such as attention-based fusion, training on diverse datasets, and optimization for low-resource devices, can further elevate its performance, paving the way for more intuitive and responsive intelligent systems in diverse domains.

Future Works

- **Emotion-Based Tourist Recommendations**: Develop a feature that recommends tourist destinations tailored to the user's detected emotion, such as calming nature spots for sadness, vibrant cultural sites for happiness, or introspective historical places for neutrality, enhancing personalized travel experiences.

- **Advanced Fusion Techniques**: Incorporate attention mechanisms or transformer-based fusion to dynamically weigh facial and audio contributions, improving accuracy in complex scenarios.

- **Diverse Dataset Training**: Train models on more inclusive datasets to address biases, ensuring robust performance across varied demographics, accents, and cultural expressions.

- **Additional Modalities**: Integrate text (e.g., sentiment analysis) or physiological signals (e.g., heart rate) to enhance emotion detection in multi-modal contexts.

- **Real-Time Optimization**: Optimize model inference for low-resource devices (e.g., mobile phones) using lightweight architectures like MobileNet to broaden deployment.

- **Robustness to Noise**: Implement noise reduction for audio and advanced face detection (e.g., MTCNN) to improve performance in challenging environments environments like crowded or poorly lit settings.

- **Fusion Model Training**: Train the fusion model on paired face-audio datasets to learn cross-modal patterns, replacing the current heuristic-based approach with data-driven fusion.

REFERENCES :

1.A BiLSTM–Transformer and 2D CNN Architecture for Emotion Recognition from Speech. By - Sera Kim 1 and Seok-Pil Lee 2

2. Development of Convolutional Neural Network Models to Improve Facial Expression Recognition Accuracy - ISSN: 2338-3070, DOI: 10.26555/jiteki.v10i2.28863