Emotion Detection Using Multimodal Analysis

Jhansi Sneha Kamsali
jkamsali@gmu.edu
G01478467

Surbhi Kharche
skharche@gmu.edu
G01481072

Venkata Tejaswi Kalla
vkalla@gmu.edu
G01460709

Vedansh Adepu
vadepu@gmu.edu
G01446978

***Abstract*— This project designs a robust Emotion Detection Module using a multimodal approach that integrates facial expressions, physiological signals, and sleep pattern data. Unlike traditional systems that rely on a single modality and often falter in real-world scenarios, this approach leverages multiple inputs to provide a more accurate and holistic understanding of emotions. By employing a late fusion strategy, the system processes these inputs independently, combining their outputs at the decision level to improve robustness and adaptability. This enables it to overcome challenges like poor lighting, occlusions, and the inability to capture long-term emotional trends. Applications include personalized media recommendations, mental health support, stress management, and interactive learning environments. This project serves as a foundation for next-generation affective computing systems designed to enhance human-computer interaction.**

I. INTRODUCTION

1.1 Problem Definition

Conventional emotion detection systems primarily rely on facial recognition, which, while effective for basic expressions, often fails in real-world conditions due to poor lighting, occlusions, or subtle emotional states. These systems also lack the depth to assess long-term emotional states or overall well-being. This limitation becomes critical in applications requiring nuanced emotion detection, such as mental health monitoring or personalized recommendations. This paper presents a multimodal Emotion Detection Module integrating facial expressions, physiological signals, and sleep patterns. Each modality contributes unique insights: visible emotions via facial recognition, internal states through physiological signals, and long-term emotional context from sleep metrics. Outputs are combined using a late fusion strategy to achieve high accuracy and robustness in diverse conditions.

1.2 Motivation

Human emotions profoundly influence decision-making, productivity, relationships, and overall well-being. Yet, existing technologies fall short of understanding this complexity, leaving a significant gap in human-computer interaction. Traditional systems typically rely on single modalities, such as facial recognition or physiological signals, which are often contextually blind. For instance, facial recognition systems may misinterpret subtle emotions under poor lighting, while physiological signals alone fail to account for behavioral or psychological context. This project's urgency lies in bridging this gap by integrating multimodal data—facial expressions, physiological metrics, and sleep patterns—to form a comprehensive emotional profile. Such integration enables a dynamic understanding of both transient emotions and long-term trends, facilitating real-world applications in mental health, personalized learning, adaptive systems, and stress management.

II. METHODOLOGY

The proposed system incorporates facial recognition, physiological signals, and sleep pattern data in a four-stage pipeline:

- data collection
- preprocessing
- model training
- output fusion using a late fusion strategy.

Each modality is processed independently to maximize its contributions while addressing limitations. The final system provides a unified emotional state prediction that adapts to real-world conditions.

2.1 Data Collection

The collection of data is a rather complex process, given the inclusion of three different modalities of data: facial expression, physiological signals, and sleep condition. Each source gives different but vital information for understanding the emotional state.

2.1.1 Facial Emotion Data: The facial dataset utilized in this system is FER-2013, a widely recognized resource for training deep learning models in emotion recognition.

- FER-2013 contains over 35,000 grayscale images labeled with seven emotion categories: happiness, sadness, anger, surprise, fear, disgust, and neutrality.
- The dataset captures a variety of scenarios, including variations in pose, lighting, and occlusions, ensuring robustness in real-world applications.
- FER-2013 enables the system to recognize facial expressions in real-time using live webcam inputs or from preprocessed datasets, ensuring flexibility and adaptability across different use cases.

By leveraging the FER-2013 dataset, this system achieves reliable and versatile emotion detection, supporting applications in both controlled and dynamic environments.

Fig. 1. FER-2013 DATASET

2.1.2 Physiological Signals Data: Physiological signals, such as heart rate variability (HRV), respiration rate, and other biometric indicators, are sourced from the WESAD dataset, a widely recognized resource for emotion recognition research. This dataset provides high-quality data from wearable devices and is labeled with four emotional states: stress, amusement, neutral (baseline), and meditation.

- The WESAD dataset includes electrocardiogram (ECG), respiration signals, and accelerometer readings, capturing a comprehensive range of physiological responses.
- Emotional states such as stress, amusement, neutral, and meditation are annotated, enabling the system to train and validate models effectively.
- The dataset supports advanced models like LSTMs, which are designed to analyze changes in signals over time.
- Features such as heart rate variability and breathing patterns are extracted to help the system distinguish between internal emotional states, such as stress or relaxation.

By leveraging the WESAD dataset and simulated data, the system gains the ability to process physiological signals accurately, making it robust, adaptable, and ready for potential real-time applications. In a real-time system, after the model is trained using the WESAD dataset, physiological data would be collected from wearable devices used by the end user. For the purpose of this project, simulated data for ECG and respiratory rate is utilized to mimic real-world scenarios.

2.1.3 Sleep Patterns: The Sleep-EDF database is utilized for sleep data analysis, offering a detailed resource for understanding long-term emotional well-being.

- It includes polysomnographic (PSG) files, hypnogram annotations, and metadata such as subject age and gender, ensuring comprehensive input for analysis.
- Metrics such as the ratio of REM sleep to total sleep, the proportion of deep sleep, and overall sleep quality are derived to provide insights into sleep patterns.
- These features help establish a broader understanding of emotional states over time, linking sleep patterns to long-term emotional well-being.

By leveraging the Sleep-EDF database and simulated data, the system combines short-term indicators with long-term sleep metrics, offering a robust framework for analyzing emotional states. In a real-time system, sleep data would be obtained from user sleep-tracking applications or wearable devices. For the purpose of this project, simulated sleep data is used to emulate real-world scenarios.

2.2 Preprocessing

Preprocessing remains one of the quintessential activities in the whole pipeline because it assures that consistency can be attained for improving model performance and also effectively prepares the inputs for emotion detection. Every modality will now undergo some specific preprocessing according to its characteristics.

2.2.1 Preprocessing of Physiological Data: The WESAD (Wireless Sensor-based Affect and Depression Detection) dataset is a valuable resource for research in affective computing and mental health. It contains physiological signals like ECG and respiration, along with corresponding emotion labels. To effectively train a machine learning model on this dataset, rigorous data preprocessing is essential.

- **Data Extraction:** Chest sensor signals (ECG and Respiration) and corresponding emotion labels are extracted from .pkl files.
- **Signal Segmentation:** Signals are divided into fixed-length windows (e.g., 256 samples) with a 75% overlap to preserve context and generate more samples.
- **Label Assignment:** Each window is assigned the most frequent emotion label, ensuring the dominant emotional state is represented.
- **Normalization:** Features are scaled using the RobustScaler to handle outliers and ensure consistent input.
- **Filtering Valid Labels:** Only windows with valid labels (Baseline, Stress, Amusement, Meditation) are retained.
- **Data Splitting:** The dataset is divided into training (70%), validation (15%), and testing (15%) subsets for model evaluation.
- **Real-Time Input Preprocessing:** For real-time input, data from wearable devices is segmented, normalized using the same scaler as the training data, and formatted into the required input shape for the model.
- **Tensor Preparation:** Data is formatted into PyTorch tensors for input to the LSTM model.

2.2.2 Facial Data Preprocessing: The FER-2013 dataset consists of grayscale images labelled with seven emotions. The following steps are applied:

- **Normalization:** Pixel values are scaled to [0, 1] by dividing by 255, ensuring consistent and stable training inputs.
- **Data Augmentation (Training Data Only):** Random transformations such as rotation (up to 40 degrees), horizontal/vertical shifts (up to 20%), shearing, zooming (within 20%), horizontal flipping, and brightness adjustments (80 to 120) increase diversity and reduce overfitting.
- **Resizing:** Images are resized to 48x48 pixels for uniformity and compatibility with the model.

- **Color Conversion:** Grayscale images are converted to RGB mode (three channels) to meet VGG16 model requirements.
- **Batch Preparation:** Augmented training images and rescaled testing images are organized into batches using ImageDataGenerator.
- **Webcam Input Processing:**
  - For real-time input, frames from the webcam are converted to grayscale for face detection.
  - Detected face regions are resized to 48x48 pixels. These regions are normalized (scaled to [0, 1]) and stacked into RGB mode to align with the model's input requirements.
  - The processed region is passed to the trained model for emotion prediction.

These preprocessing steps optimize the FER-2013 dataset, improving model compatibility, diversity, and performance for robust emotion detection.

2.2.3 Sleep Data Preprocessing: The Sleep-EDF Database is a publicly available dataset containing polysomnography (PSG) recordings and sleep stage annotations. It's a valuable resource for research in sleep medicine and related fields.

- **Data Loading and Filtering:**
  - Load EDF files using libraries like pyedflib.
  - Handle unreadable files and extract essential information such as signal labels, sampling frequency, file duration, and sleep stage annotations.
- **Feature Extraction:**
  - Calculate sleep stage proportions (Wake, Stage 1-4, REM).
  - Extract features like sleep onset latency, WASO, sleep efficiency, and sleep fragmentation indices.
- **Data Cleaning and Imputation:**
  - Handle missing values using imputation techniques.
  - Identify and remove outliers using statistical methods or domain knowledge.
- **Data Normalization:**
  - Standardization: Scale features to have zero mean and unit variance.
  - Min-Max normalization: Scale features to a specific range (e.g., 0-1).
- **Feature Engineering:**
  - Create new features from existing ones.
  - Apply time-series analysis and frequency domain analysis to extract additional features.
- **Labeling and Annotation:**
  - Assign labels based on sleep quality or disorders.
- **Data Splitting:**
  - Divide the dataset into training and testing sets (70-30 or 80-20 split).

2.2.4 Late Fusion Preparation:

- Outputs from the preprocessing steps of each modality are formatted to align with the input requirements of their respective models (e.g., CNN for facial data, LSTM for physiological signals, and Random Forest for sleep data).
- Temporal coherence is maintained to support real-time data integration, ensuring that outputs from different models can be effectively combined in the late fusion stage.

2.3 Model Training

2.3.1 Facial Emotion Recognition:

- **Input (48x48x3):** The model receives a resized color image of a face, represented as a tensor of shape (1, 48, 48, 3).
- **Convolutional Layers (VGG16 Base):**
  - Convolution: A small filter slides across the input image, creating a feature map by highlighting specific patterns.
  - ReLU Activation: Sets negative values to zero, introducing non-linearity.
  - Max Pooling: Reduces spatial dimensions of feature maps by selecting the maximum value within a window, enhancing robustness.
  - Repetition: This process repeats through multiple convolutional blocks, detecting simple features like edges initially and more complex features in deeper layers.
- **Flattening:** Converts the multi-dimensional tensor from the last convolutional block into a one-dimensional vector.
- **Dense Layers (Fully Connected Layers):**
  - Dense Layer 1 (256 neurons): Learns complex relationships between features.
  - Linear Transformation: Computes a weighted sum of inputs and adds a bias.
  - ReLU Activation: Applied to the output.
  - Batch Normalization: Normalizes activations for each mini-batch.
  - Dropout (0.5): Randomly sets 50% of neurons to zero during training to prevent overfitting.
  - Dense Layer 2 (7 neurons): One neuron for each emotion.
  - Softmax Activation: Converts outputs into a probability distribution over 7 emotion classes.
- **Output (7 probabilities):** The model outputs a vector of probabilities, with each probability representing the confidence that the input image belongs to a particular emotion class. The highest probability indicates the predicted emotion.

2.3.2 Physiological Signal Processing:

- **LSTM Model with Attention Mechanism:**
  - Source: WESAD dataset (ECG for heart rate variability, respiration rates).
  - Labels: Baseline, Stress, Amusement, Meditation.
- **Model Architecture:**
  - LSTM Layers: Bidirectional LSTM captures temporal contexts from both past and future.

  - Attention Mechanism: Enhances important parts of the input sequence for better emotion classification.
- **Regularization:**
  - Dropout: Prevents overfitting.
  - Early Stop: Stops training when performance no longer improves.
- **Optimization:**
  - Optimizer: AdamW with weight decay for better generalization.
  - Learning Rate Scheduler: ReduceLROnPlateau adjusts the learning rate based on validation performance.
  - Loss Function: Cross-entropy loss.
- **Final Layer:**
  - Dense Layer: Maps learned features to one of four emotional states (Baseline, Stress, Amusement, Meditation).
  - Activation Function: Softmax for converting outputs into a probability distribution.
- **Model Evaluation:**
  - Metrics: Precision, Recall, and F1-score (using sklearn metrics) to ensure robustness before final fusion stage.

2.3.3 Sleep Pattern Analysis:

- **Random Forest Classifiers:**
  - Data Processed: Categorical and periodic data including REM sleep ratio, deep sleep duration, and sleep stage annotations from the Sleep-EDF dataset.
  - Feature Extraction:
    * Proportions of different sleep stages.
    * Sleep quality metrics.
  - Normalization: StandardScaler used to normalize features.
  - Synthetic Emotion Labels: Augmented with synthetic emotion labels generated using pre-defined rules.
  - Class Imbalance Handling: SMOTE used to address class imbalances in training data.
  - Hyperparameter Tuning: GridSearchCV optimizes tree depth and number of estimators for improved predictive performance.
  - Linking Behavioral Patterns and Emotions:
    * Behavioral Patterns: Analysis of sleep quality and duration linked to synthetic emotional states (happiness, sadness, anger).
    * Feature Importance: Plots provide explanations of sleep stages' contributions to emotion predictions.
  - Output: Probabilities for each synthetic emotion, combined in the fusion step.
  - Contribution: Ensures sleep data contributes to long-term emotional insights, supplementing real-time predictions of facial and physiological models.

2.3.4 Fusion and Unified Prediction

Implements a multi-modal emotion recognition system by combining predictions from three distinct models: facial expression analysis with a Convolutional Neural Network, physiological signal processing using a Long Short-Term Memory network with attention, and sleep pattern analysis using a pre-trained model.

Data Flow and Processing::

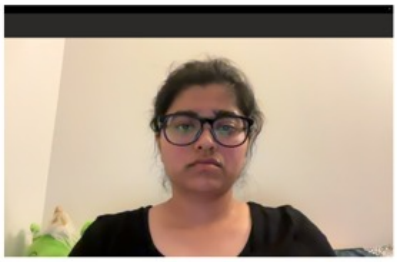
1) **Input Acquisition:** The system gathers input from three sources:
   - *Facial Expressions:* A live webcam feed captures an image of the user's face.
   - *Physiological Signals:* The user is prompted to input ECG (electrocardiogram) and respiration rate data as comma-separated values.
   - *Sleep Metrics:* The user provides six sleep-related metrics: proportion of wake time (prop_W), proportions of NREM sleep stages (prop_1 to prop_4), and proportion of REM sleep (prop_R), also as comma-separated values.
2) **Individual Model Predictions:**
   - *Facial Expression Analysis:* Converts the captured image to grayscale, detects faces using a Haar cascade classifier, resizes the facial region to 48×48 pixels, and processes it with a pre-trained Keras CNN (`face_emotion_model`) to generate probabilities over seven emotions (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral).
   - *Physiological Signal Processing:* Converts input ECG and respiration values into a PyTorch tensor with a batch dimension, processes it with a pre-trained LSTM network (`lstm_model`) with attention to capture temporal dependencies, and outputs probabilities over four emotions (Baseline, Stress, Amusement, Meditation).
   - *Sleep Pattern Analysis:* Converts input sleep metrics into a NumPy array, standardizes using a `StandardScaler`, and processes with a pre-trained `joblib` model (`sleep_model`) to generate probabilities over five emotions (happiness, sadness, anger, surprise, fear).
3) **Emotion Mapping:** To reconcile the different emotion classes predicted by each model, a mapping is applied. This maps the specific emotions outputted by each model to a common set of seven emotions - Happy, Sad, Neutral, Angry, Relaxed, Stressed, Bored. The probabilities for mapped emotions are summed and then normalized by the total probability across all mapped emotions for each modality.
4) **Probability Fusion:** The probability distributions from the three models, after mapping, are combined using a weighted average. The weights for each model are defined in the `weights` list. This fusion step combines the information from all three modalities into a single, more robust emotion prediction.

5) **Final Emotion Prediction:** The emotion with the highest combined probability is selected as the final predicted emotion.
6) **Output Display:** The individual model predictions, the combined probabilities, and the final predicted emotion are printed to the console.

III. Results :

Facial Emotion Recognition Module

Input:



```
Initializing webcam...
Press 'c' to capture a photo.
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 70ms/step
```

Fig. 2. Real-time emotion detection system interface showing webcam initialization and photo capture for analysis

Output

```
--- Emotion Detection Results ---

Face Emotion Results:
  Happy: 35.28%
  Sad: 0.00%
  Neutral: 0.00%
  Angry: 0.00%
  Relaxed: 0.00%
  Stressed: 59.13%
  Bored: 5.59%
```

Fig. 3. Real-time face emotion detection results showing confidence percentages for each emotion

Physiological Signals Emotion

Simulated Input:

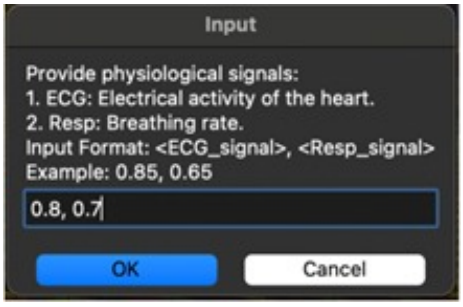


Fig. 4. Input interface for providing physiological signals, including ECG and Respiration values

Output:

```
LSTM Emotion Results:
  Happy: 4.65%
  Sad: 0.00%
  Neutral: 0.00%
  Angry: 0.00%
  Relaxed: 5.70%
  Stressed: 66.26%
  Bored: 23.39%
```

Fig. 5. LSTM emotion detection results showing predicted emotions with confidence percentages

Sleep Pattern-Based Emotion Detection Module

Simulated Input:

Fig. 6. Input interface for providing six sleep metrics, including Wake, NREM stages, and REM sleep proportions

Output:

```
Sleep Emotion Results:
  Happy: 100.00%
  Sad: 0.00%
  Neutral: 0.00%
  Angry: 0.00%
  Relaxed: 0.00%
  Stressed: 0.00%
  Bored: 0.00%
```

Fig. 7. Sleep emotion detection results displaying predicted emotions with confidence percentages

Multi Modal Integrated Model

Technique:

```
# Combine and determine final emotion
weights = [0.6, 0.2, 0.2]
combined = combine_emotion_values([mapped_face, mapped_lstm, mapped_sleep], weights)
final_emotion = max(combined, key=combined.get)

# Display results
display_results(mapped_face, mapped_lstm, mapped_sleep, combined, final_emotion)

return final_emotion  # Returning the final emotion
```

Fig. 8. Code snippet showing the fusion process where facial, LSTM, and sleep emotion predictions are combined using weighted averaging to determine the final emotion

Output:

```
Combined Results:
  Happy: 42.10%
  Sad: 0.00%
  Neutral: 0.00%
  Angry: 0.00%
  Relaxed: 1.14%
  Stressed: 48.73%
  Bored: 8.03%
```

Fig. 9. Combined emotion detection results showing the final weighted fusion of facial, LSTM, and sleep-based predictions, with ”Stressed” as the dominant emotion.

```
Final Detected Emotion: Stressed
```

```
Detected Emotion: Stressed
```

Fig. 10. Final detected emotion result showing ”Stressed” as the dominant emotion

Performance Evaluation

A. Output:

Record	Ground Truth	Predicted Emotion
1	Happy	Happy
2	Stressed	Stressed
3	Happy	Stressed
4	Happy	Stressed
5	Bored	Stressed
6	Bored	Stressed
7	Stressed	Stressed
8	Stressed	Stressed
9	Stressed	Stressed
10	Stressed	Stressed
11	Neutral	Stressed
12	Happy	Stressed
13	Happy	Stressed
14	Stressed	Stressed
15	Stressed	Happy
16	Happy	Stressed
17	Stressed	Stressed
18	Stressed	Stressed
19	Stressed	Stressed
20	Happy	Stressed
21	Happy	Stressed
22	Happy	Stressed
23	Stressed	Stressed
24	Stressed	Stressed

Final Accuracy: 0.50

Fig. 11. able showing the comparison of ground truth and predicted emotions for each record, with a final accuracy of 0.50 for Conventional Face Recognition Model

- The primary metric used to evaluate the model’s performance is accuracy. It measures the proportion of correctly classified records. In the given example, the final accuracy is reported as 0.50, which indicates that the model correctly predicted the emotion in 50% of the test cases.
- Evaluation Methodology: The code reads the test data from a CSV file containing ”Ground Truth Emotion,” and an ”ImagePath”.
- For each record, facial features are extracted and emotion is predicted using the loaded face_emotion_model. The predicted emotion is then mapped to the common emotion labels. The predicted emotion is compared to the ”Ground Truth Emotion” from the CSV file.
- The accuracy is calculated by dividing the number of correct predictions by the total number of valid predictions.
- Low Accuracy: The reported accuracy of 0.50 suggests that the face recognition model is not performing well on the given test data.
- This is due to several factors:
  - Model Limitations: The pre-trained face emotion

model might not be well-suited for the specific characteristics of the faces in the test data.
  - Ground Truth Errors: The ground truth labels in the test data might contain errors or inconsistencies.

Multi Modal Emotion Detection Module:

B. Output:

Record	Ground Truth	Predicted Emotion
1	Happy	Happy
2	Stressed	Stressed
3	Happy	Stressed
4	Happy	Stressed
5	Bored	Stressed
6	Bored	Stressed
7	Stressed	Stressed
8	Stressed	Stressed
9	Stressed	Stressed
10	Stressed	Stressed
11	Neutral	Happy
12	Happy	Stressed
13	Stressed	Stressed
14	Stressed	Happy
15	Stressed	Happy
16	Stressed	Stressed
17	Stressed	Stressed
18	Stressed	Stressed
19	Stressed	Stressed
20	Happy	Stressed
21	Happy	Stressed
22	Stressed	Stressed
23	Stressed	Stressed
24	Stressed	Stressed

Final Accuracy: 0.58

Fig. 12. Comparison of ground truth and predicted emotions for each record, with a final accuracy of 0.58 for Multi modal Emotion detection

- The primary metric used to evaluate the model's performance is accuracy, which measures the proportion of correctly classified records.
- In the given example, the final accuracy is reported as 0.58, indicating that the model correctly predicted the emotion in 58% of the test cases.
- Evaluation Methodology:
  - The code reads test data from a CSV file containing columns for "Record," "Ground Truth Emotion," "ImagePath," "LSTMInput," and "SleepInput."
  - For each record, the code calls the corresponding functions to obtain emotion probabilities from each modality.
  - Emotion probabilities from each model are mapped to a common set of emotions using the emotion_mappings dictionary.
  - The mapped probabilities are combined using weighted averaging (face: 0.4, LSTM: 0.4, sleep: 0.2).
  - The emotion with the highest combined probability is selected as the final predicted emotion.
  - The predicted emotion is compared to the "Ground Truth Emotion" from the CSV file.
  - Accuracy is calculated by dividing the number of correct predictions by the total number of valid predictions (excluding cases where an error occurred during prediction).

Performance Analysis

- Moderate Accuracy: The accuracy of 0.58 suggests moderate performance. While better than the initial face recognition model alone, there's still room for improvement.
- Multimodal Fusion: The use of multimodal fusion (combining face, LSTM, and sleep predictions) likely contributes to the improved accuracy compared to the face recognition model alone.

IV. Applications

That would be a great promise for multimodal emotion detection wherein facial expressions, physiological signals, and sleep bring out an overall record in different dimensions. We applied the Multi Modal Emotion detection system to enhance Spotify's Music Recommendation System.

Enhanced Emotion-Based Music Recommendation System

- Emotion Detection and Refinement: Utilize a multimodal emotion detection system that combines facial expressions, physiological signals, and sleep patterns for accurate emotional assessment.
- Music Recommendation Logic:
  - Emotion-to-Genre Mapping: Extend the emotion-to-genre dictionary with nuanced mappings:
    - Happy: Pop, upbeat dance, tropical, indie pop
    - Sad: Acoustic, folk, indie folk, lo-fi, ballads
    - Neutral: Indie, alternative, chillwave, jazz
    - Angry: Rock, metal, punk, grunge, hard rock
    - Relaxed: Ambient, chill, lo-fi, classical, jazz
    - Stressed: Ambient, classical, nature sounds, instrumental, deep house
    - Bored: Electronic, experimental, avant-garde, noise
  - Spotify API Integration:
    - Authenticate with Spotify API using provided credentials.
    - Search for tracks based on selected genres using Spotify API.
    - Create a new Spotify playlist with a relevant name (e.g., "Happy Vibes," "Chill Vibes").
    - Add selected tracks to the newly created playlist.
  - User Interaction and Feedback Loop:
    - User Interface: Provide a UI (e.g., a mobile app) for user interaction.
    - User Actions: View and listen to the recommended playlist.

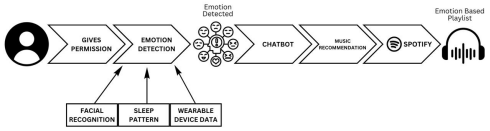


Fig. 13. End-to-end system workflow illustrating emotion detection using facial recognition, sleep patterns, and wearable device data. Detected emotions are processed by a chatbot for personalized music recommendations via Spotify, resulting in an emotion-based playlist.

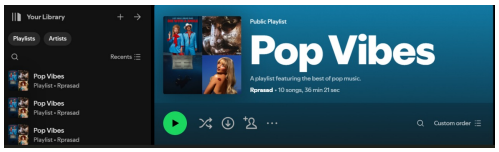


Fig. 14. Emotion-based playlist generated on Spotify, featuring songs tailored to the detected emotional state of the user.

V. Early Fusion vs. Late Fusion in Multimodal Emotion Recognition

Early Fusion

- Combines different modalities (e.g., facial expressions, physiological signals, sleep data) at the input level by concatenating feature vectors before feeding them into a single model.
- **Advantages:**
  - By fusing data early, the model can learn complex interdependencies and potentially capture more nuanced information.
  - Allows the model to exploit the strengths of each modality and compensate for weaknesses in others.
- **Disadvantages:**
  - Requires designing and training a single, complex model that handles all modalities simultaneously.
  - If one modality has significantly more or less information than others, it can dominate the learning process.
  - Adding new modalities or changing the representation of existing ones can require significant model retraining.
  - Can be computationally expensive due to the need to process all modalities together.

Late Fusion

- Processes each modality independently using separate models and then combines the predictions (e.g., probabilities or scores) at the decision level.
- **Advantages:**
  - Allows for flexibility in adding or removing modalities without affecting the core architecture of other models.
  - Each model deals with a smaller feature space, simplifying model design and training.
  - Enables independent optimization of each model for its respective modality, potentially leading to better performance for each individual task.
  - Can be more computationally efficient than early fusion, especially when dealing with high-dimensional data.
- **Disadvantages:**
  - May not capture complex interactions or dependencies between different modalities.

Why we chose Late Fusion (in the given context):

- Modularity: The provided code utilizes pre-trained models for face emotion, physiological signal processing (LSTM), and sleep analysis. Late fusion allows for easy integration of these existing models without the need to retrain a single, large model from scratch.
- Flexibility: Adding new modalities (e.g., voice analysis) is relatively straightforward by simply adding a new model and integrating its predictions into the existing fusion mechanism.
- Reduced Complexity: Late fusion avoids the complexity of training a single model that handles all modalities simultaneously, which can be challenging and computationally expensive.

TABLE I
Comparison of Early Fusion and Late Fusion

Feature	Early Fusion	Late Fusion
Concept	Combines modalities at input level	Processes modalities independently and combines outputs
Complexity	High: requires integrating and synchronizing raw data	Lower: simpler model designs for each modality
Modality-Specific Insights	Risk of diluting unique features	Maintains unique features of each modality
Flexibility	Low: hard to adapt to missing/noisy data	High: handles missing/noisy data in one modality
Scalability	Low: adding/replacing modalities is complex	High: easy integration of new modalities
Computational Cost	High: processes all modalities together	Lower: independent processing of smaller feature spaces
Interpretability	Harder: combined model complexity	Easier: separate evaluation of each modality
Example	Blending image data with signals might blur details	Each model operates independently, preserving details

VI. Late Fusion Strategy

The late fusion strategy is critical for integrating outputs from facial recognition, physiological signals, and sleep pattern models to make a unified prediction of the emotional state. To optimize the fusion, two techniques were considered: Weighted Averaging and Neural Network Fusion. These methods aim to enhance the system's accuracy by

leveraging the strengths of each modality while addressing their limitations.

Weighted Averaging

- The outputs from each model are combined based on assigned weights.
- **Weight Tuning:** These weights are adjustable to reflect the reliability of each model under specific conditions. For example:
  - In low-light conditions where facial detection may be less reliable, greater weight can be given to physiological signals or sleep metrics.
  - The weights are tuned to optimize performance, ensuring the fusion process is flexible and robust.
- **Mathematical Representation:** The final emotion probability P_{final} is computed as:

$$P_{\text{final}} = w_1 \cdot P_{\text{face}} + w_2 \cdot P_{\text{LSTM}} + w_3 \cdot P_{\text{sleep}}$$

where w_1, w_2, w_3 are the weights and $P_{\text{face}}, P_{\text{LSTM}}, P_{\text{sleep}}$ are the probability distributions from the facial, LSTM, and sleep models, respectively.

TABLE II
COMPARISON OF WEIGHTED AVERAGING AND NEURAL NETWORK FUSION

Feature	**Weighted Averaging**	**Neural Network Fusion**
Complexity	Simple to implement and tune	More complex, requires additional neural network training
Flexibility	High; weights can be adjusted based on conditions	Moderate; requires retraining for different conditions
Computational Cost	Low	Higher due to additional neural network computations
Adaptability	Moderate; manually adjusted weights	High; learns relationships between model outputs
Robustness	Robust in varying conditions with manual adjustments	Very robust, adapts to complex interdependencies
Training Time	Shorter	Longer due to meta-model training
Scalability	Easily scalable	More challenging due to added complexity
Tuning Requirement	Manual tuning of weights	Requires training and tuning of neural network

We chose Weighted Averaging due to its simplicity, lower computational cost, and ease of implementation. It allows for quick adjustments to the weights based on specific conditions, providing a flexible and robust solution. While Neural Network Fusion offers higher adaptability and robustness, its complexity and longer training time made Weighted Averaging a more practical choice for our system.

VII. DIFFERENTIATING OUR APPROACH FROM PRIOR WORK: ADVANCING LATE FUSION STRATEGIES

Late fusion strategies are widely regarded as an effective approach in multimodal emotion recognition systems. By creating independent representations for each modality (e.g., facial expressions, physiological signals, and sleep metrics) and integrating them at the decision-making stage, late fusion ensures that the unique contributions of each modality are preserved without interfering with one another during processing. This makes it particularly well-suited for handling diverse types of data that require distinct preprocessing and modeling techniques.

Jun Yu et al. have demonstrated the robustness of late fusion in enhancing the adaptability of multimodal systems, particularly in environments with variable input quality. Their work employs strategies such as weighted averaging, which adjusts the contribution of each modality dynamically based on reliability, and neural network fusion, which optimizes predictions by learning inter-modality relationships. While effective, these methods require additional calibration and training complexity.

In the context of prior work, our project differentiates itself by focusing on the integration of real-time physiological signals and long-term sleep metrics alongside facial expressions, creating a more comprehensive framework for emotion detection. Unlike Jun Yu's focus on audiovisual data and mimicry intensity estimation, our approach bridges short-term and long-term emotional indicators. Additionally, our late fusion method emphasizes balancing real-time responsiveness with long-term trends by dynamically adjusting modality contributions based on context (e.g., prioritizing physiological signals in poor lighting conditions). This nuanced approach expands the applicability of late fusion strategies to areas like mental health monitoring and adaptive learning environments, where both immediate and extended emotional states are critical.

VIII. CHALLENGES

1) **Wearable Device Data Access:** Real-time access to physiological data from wearable devices required expensive SDKs and proprietary APIs. Limited access and high costs restricted the use of real-time data for development.
2) **Sleep Data Integration:** Integration with third-party sleep tracking apps faced challenges due to licensing and data access restrictions. The lack of seamless data interoperability hindered the incorporation of real-world sleep metrics.
3) **Data Variability and Bias in Facial Recognition:** Variability in facial data caused by poor lighting, occlusions, and cultural differences reduced model accuracy. A lack of diverse demographic representation

in datasets like FER-2013 introduced biases, limiting generalizability across populations.

4) **Sequential Nature of Physiological Data:** Individual differences in heart rate and respiration signals introduced variability, complicating emotion classification. Handling the sequential nature of physiological signals required careful consideration of data patterns.
5) **Scarcity of Emotion-Labeled Sleep Data:** The lack of real-world emotion labels for sleep patterns added uncertainty to emotion classification. The reliance on synthetic labels further complicated model validation.
6) **Multi-Modal Fusion Complexity:** Integrating diverse data streams in real-time posed synchronization and computational challenges. Determining optimal weights for each modality in late fusion added another layer of complexity.
7) **Resource-Intensive Model Training:** Training deep learning models demanded significant hardware resources and computational time. These requirements made development resource-intensive and constrained scalability.

IX. Conclusion

Our multimodal emotion detection system is a step forward in improving human-computer interaction and personalized applications. By integrating facial expressions, physiological signals, and sleep patterns, the system offers a more comprehensive framework for understanding emotions compared to single-modality approaches. It bridges the gaps in emotional interpretation by combining diverse data sources, resulting in insights that are richer and more contextually aware.

While the project faces challenges such as limited real-time data access, computational constraints, and variability in data quality, it highlights the potential of such systems. The late fusion methodology ensures each data modality contributes meaningfully to the final prediction, enhancing the system's ability to provide emotions that are not just accurate but also more contextually relevant. Although we did not see dramatic improvements in raw accuracy metrics due to these limitations, the emotions detected are more robust and better reflect real-world scenarios.

This system has potential applications in areas like mental health monitoring, adaptive learning environments, human-robot interaction, and emotion-driven media recommendations. By addressing both short-term emotional changes and long-term patterns, it lays a solid foundation for emotionally intelligent systems that adapt to real-life challenges. Future work will aim to improve real-world integration, scalability, and efficiency to further enhance its impact.

In summary, this project demonstrates the potential for creating emotion detection systems that go beyond simple accuracy to deliver deeper, more empathetic insights. It opens the door for advancing intuitive, emotion-aware technologies that can transform user experiences across various fields.

X. Team Contributions

The successful completion of this project was made possible through the collective efforts of the team. Each member contributed uniquely to different aspects of the project, as outlined below:

- **Jhansi:**
  - Project proposal and initial framework.
  - Research and design of the system architecture.
  - Design, Development, and integration of multiple models.
  - Design and implementation of the Music Recommendation Module.
  - Testing and validation of the complete system.
  - Preparation of PowerPoint slides, code demonstration, and project report writing.
- **Tejaswi:**
  - Research and design of the system architecture.
  - Contribution to the project proposal.
  - Design and development of the Physiological Data Analysis Module.
  - Support in creating the PowerPoint presentation.
- **Vedansh:**
  - Research and design of the system architecture.
  - Design and development of the Sleep Pattern Analysis Module.
- **Surbhi:**
  - Development of the Facial Detection Module.
  - Creation to the PowerPoint presentation.
  - Explanation of the project video.
  - Drafting the initial project report.

The team worked collaboratively to ensure that the project met its objectives, overcoming challenges and delivering a comprehensive solution. Each member's contributions were instrumental in achieving the project's success.

References

[1] Amil Khanzada and Others. Email thread summary dataset. Available: https://arxiv.org/pdf/2004.11823.

[2] Jukka Kortelainen and Others. Multimodal emotion recognition by combining physiological signals and facial expressions. 2012. Available: https://moscow.sci-hub.se/2154/1ddd7643e32d49e0f225a4300456267a/kortelainen2012.pdf.

[3] Aneesh Srivastava and Others. Facial emotion-based music recommender system using CNN. Available: https://ieeexplore.ieee.org/document/10085294.

[4] Barathi Subramanian and Others. Digital twin model: A real-time emotion recognition system for personalized healthcare. 2022. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9840353.

[5] Jun Yu and Others. Efficient feature extraction and late fusion strategy for audiovisual emotional mimicry intensity estimation. 2024. Available: https://openaccess.thecvf.com/content/CVPR2024W/ABAW/papers/Yu_Efficient_Feature_Extraction_and_Late_Fusion_Strategy_for_Audiovisual_Emotional_CVPRW_2024_paper.pdf.