STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, May 2026, pp 4237–4254.
Published online in International Academic Press (www.IAPress.org)

Decision-Level Fusion for Facial and Speech Emotion Recognition: A CNN-Based Web Application

Hind Mestouri*, Abdelilah Jraifi, Kamal Baraka

University of Caddi Ayyad (UCA), National Schools of Applied Sciences of Safi, Laboratory of Mathematical Computer and Communication System, Morocco

Abstract This paper presents a real-time web-based emotion recognition system based on unimodal deep learning models for facial and speech analysis, combined through decision-level score aggregation. Facial emotion recognition is performed using convolutional neural networks (CNNs), while speech emotion recognition relies on a CNN–BiLSTM architecture to capture both spatial and temporal speech patterns. These models are chosen for their effectiveness and low computational cost, making them suitable for web-based deployment. The facial model is trained on the FER2013 dataset, and the speech model is trained on the RAVDESS corpus using MFCC-based audio features. Rather than performing multimodal representation learning, this work demonstrates decision-level fusion by aggregating unimodal prediction scores to improve robustness when combining facial and speech information. Experimental results show competitive recognition performance and support the applicability of the proposed system for human-computer interaction in real-time and web-based affective applications.

Keywords Emotion Recognition; Decision-Level Fusion; Facial Expression Analysis; Speech Emotion Recognition; CNN ; Web-Based Systems

AMS 2010 subject classifications 68T07, 68T45, 62H35.

DOI: 10.19139/soic-2310-5070-misi25:655676

1. Introduction

Automatic recognition of human emotions has emerged as a fundamental research area in artificial intelligence (AI), particularly within domains involving human-computer interaction, cognitive modeling, and affective computing. The ultimate objective is to enable machines to detect, interpret, and respond to emotional states based on observable signals such as facial expressions, vocal intonation, body language, and gestures. With the growing integration of intelligent systems in real-world environments, emotion recognition has proven valuable in healthcare, smart homes, education, marketing, and assistive technologies, where understanding user emotions can significantly enhance personalized interactions [1].

Historically, early emotion recognition systems relied on handcrafted feature extraction and classical machine learning techniques. In facial emotion analysis, geometric and appearance-based features (such as inter-landmark distances or texture descriptors) were commonly used and classified using methods such as Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) [1, 2]. Similarly, speech-based emotion recognition exploited prosodic and spectral features, including pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs), combined with traditional classifiers.Although these methods laid the foundation for automated emotion

*Correspondence to: Hind Mestouri (Email: h.mestouri@uca.ac.ma). Department of Computer Science, Networks and Telecommunications, at the National Schools of Applied Sciences (ENSA) in Safi, University of Caddi Ayyad, Route Sidi Bouzid, BP 63, 46 000 Safi, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

recognition, they suffered from limited generalization, required substantial domain-specific tuning, and performed poorly in unconfined real-world scenarios [2]. The advent of deep learning has revolutionized the field, particularly through Convolutional Neural Networks (CNNs), which have enabled end-to-end models that automatically learn hierarchical, discriminative representations directly from raw data [3]. CNN-based models have achieved state-of-the-art performance in benchmark facial emotion datasets such as FER2013 and CK+ [4, 5]. Despite their widespread use, these datasets are limited by controlled acquisition conditions and may not fully represent real-world emotional variability. Similarly, in speech emotion recognition, the use of time-frequency representations (e.g., spectrograms) combined with deep architectures has significantly improved recognition accuracy over traditional acoustic models [6].

More recently, multimodal emotion recognition, which integrates both visual and auditory modalities, has emerged as a promising research direction. Empirical studies have shown that the fusion of complementary signals enhances robustness and improves the accuracy of classification. Advanced approaches include hybrid deep networks and transformer-based fusion models, which jointly learn spatial and temporal features across modalities [7, 8]. Despite these advances, many existing systems remain computationally intensive and challenging to deploy in real-world settings. This creates a gap between research prototypes and practical applications. To address this, we propose a lightweight web-based application for multimodal emotion recognition, using CNN-based architectures for facial and speech analysis. The system is trained and evaluated on publicly available datasets, including RAVDESS, and supports real-time emotion classification. The goal is to provide an accessible, efficient, and user-friendly platform for human-centered AI applications, bridging the gap between academic research and everyday usability.

Many state-of-the-art multimodal fusion approaches rely on heavy architectures and temporally synchronized audio–visual data, which limits their applicability in real-time web environments. In contrast, this work prioritizes lightweight unimodal models and decision-level fusion to ensure low latency and deployment feasibility. The main contributions of this work are as follows: we propose a lightweight CNN-based facial emotion recognition model suitable for real-time web deployment; we introduce a CNN–BiLSTM architecture for speech emotion recognition using MFCC-based audio representations; we define a clear decision-level late fusion strategy that aggregates unimodal predictions at inference time without relying on joint multimodal learning; we implement a complete web-based system enabling real-time emotion inference; and we conduct an experimental evaluation under realistic usage conditions.

2. Methodology

2.1. System Overview

The proposed system is a web-based emotion recognition application that analyzes human emotions in real time from facial images and speech signals. Two independent unimodal deep learning models are used, each trained and evaluated separately in its respective modality, without joint multimodal learning or aligned audio–visual data. The system consists of two parallel processing pipelines for visual and audio inputs. Each pipeline performs emotion recognition independently, and their output can be optionally combined at inference time through decision-level score aggregation to enhance robustness. The overall architecture is illustrated in Figure 1, where two separate processing streams handle the input modalities.

• Audio Processing Pipeline: The audio pipeline begins by capturing speech signals either via microphone input or audio file upload. The raw waveform is transformed into time-frequency representations, including Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms. These feature maps are fed into a deep neural network composed of multiple convolutional layers followed by a bidirectional Long- and Short-Term memory layer (BiLSTM). This configuration enables the model to capture both spatial patterns and temporal dependencies, which are crucial for speech emotion recognition.

• Image Processing Pipeline: The visual pipeline processes grayscale facial images obtained from either real-time webcam capture or uploaded image files. A CNN model trained on publicly available facial emotion datasets

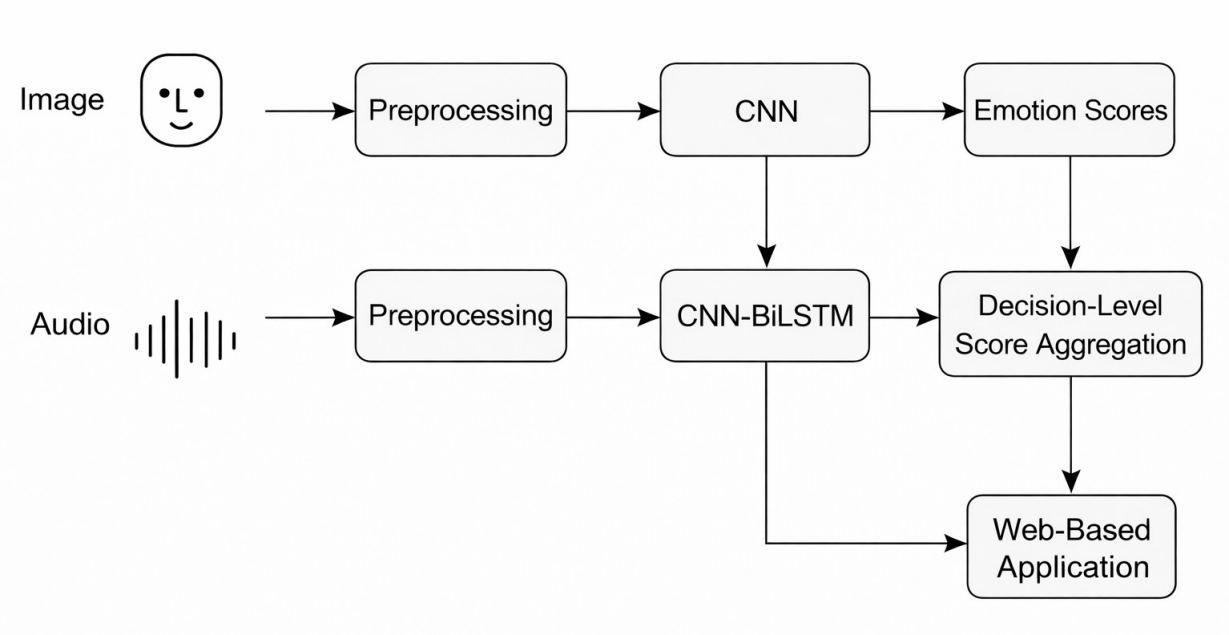

Figure 1. Architecture of the proposed web-based emotion recognition system with decision-level score aggregation.

is used to automatically detect discriminative facial features, such as eye and mouth movements, and map them to predefined emotion categories.

After independent inference, the predicted class probabilities generated by the unimodal facial and speech models may be combined using a decision-level late fusion strategy, such as probability averaging or weighted score aggregation. This aggregation is performed only at inference time and does not involve joint training or shared feature representations. The final emotional label is obtained directly from the aggregated decision scores. The system is implemented as a user-friendly web application that supports both real-time interaction and offline analysis. Its modular design ensures flexibility, interpretability, and suitability for deployment in real-time human–computer interaction scenarios.

2.2. Datasets Description

To train and evaluate the proposed multimodal emotion recognition system, four widely used benchmark datasets were employed: FER2013, RAVDESS, JAFFE, and KDEF. These datasets provide complementary characteristics, ensuring robustness and generalization across modalities and recording conditions.

• FER2013 (Facial Expression Recognition 2013)

The FER2013 dataset contains 35,887 grayscale facial images categorized into seven emotion classes: happiness, sadness, anger, fear, surprise, disgust, and neutral. All images are resized to 48×48 pixels and were originally collected from the internet under unconstrained conditions, making FER2013 a challenging benchmark for facial expression recognition [4].

• RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)

The RAVDESS dataset consists of 1,440 recordings produced by 24 professional actors (12 male, 12 female), expressing eight different emotions (neutral, calm, happy, sad, angry, fearful, disgust, and surprised). The recordings are balanced across speech and song modalities and are widely used in speech emotion recognition studies [6].

• JAFFE (Japanese Female Facial Expression Database)

The JAFFE dataset contains 213 grayscale images of 10 Japanese female subjects, each depicting seven basic emotions: happiness, anger, sadness, fear, surprise, disgust, and neutrality. Each image is sized 256×256 pixels

and has been annotated by multiple human raters to validate the emotional content. The dataset is widely used for evaluating facial expression recognition systems under controlled conditions [9].

• KDEF (Karolinska Directed Emotional Faces)

The KDEF dataset consists of 4,900 color images of 70 individuals (35 male and 35 female), each expressing seven emotional states: happiness, anger, sadness, fear, surprise, disgust, and neutral. Images are captured from five different head angles (frontal, ±45°, and profile views), providing a useful benchmark for systems that must generalize across pose variation [10].

In this study, FER2013 was used as the primary training dataset for the facial emotion recognition pipeline, while RAVDESS was employed to train and evaluate the speech emotion recognition module. JAFFE and KDEF were reserved exclusively for independent testing and external validation of the facial model, enabling an unbiased assessment of cross-dataset generalization under varying recording conditions and demographic distributions. All image datasets underwent identical preprocessing, including grayscale conversion, resizing to 48×48 pixels, and normalization. For speech data, spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) were extracted.

• Emotion Class Harmonization

To address differences in emotion taxonomies between FER2013 and RAVDESS, a unified set of seven emotions was defined: happy, sad, angry, fear, surprise, disgust, neutral. Non-overlapping classes were removed or merged, with the calm class in RAVDESS combined with neutral. All reported figures, tables, and results adhere to this standardized taxonomy, which ensures consistent multimodal evaluation and interpretation.

• Facial Emotion Recognition Pipeline

For the visual modality, grayscale facial images resized to 48×48 pixels are processed by a CNN comprising four 3×3 convolutional layers with ReLU activation and 2×2 max-pooling, followed by two fully connected layers (128 and 64 neurons) with ReLU activation and dropout (0.5). A Softmax layer outputs probabilities over the seven emotion classes. The network is trained using categorical cross-entropy and optimized with Adam (learning rate = 0.001). Data augmentation (horizontal flipping and random cropping) and early stopping are applied to reduce overfitting. The model contains approximately 1.2 million parameters (5.3 MB) and was trained for 50 epochs with a batch size of 64, balancing computational efficiency and real-time applicability while ensuring reproducibility.

Figure 2 illustrates the actual CNN architecture used in this study. The model receives 48×48 grayscale facial images and processes them through four convolutional blocks (3×3 kernels, ReLU activations), each followed by 2×2 max-pooling. The feature maps are then flattened and passed through two fully connected layers with 128 and 64 units, before the final Softmax layer that predicts the seven standardized emotion categories.

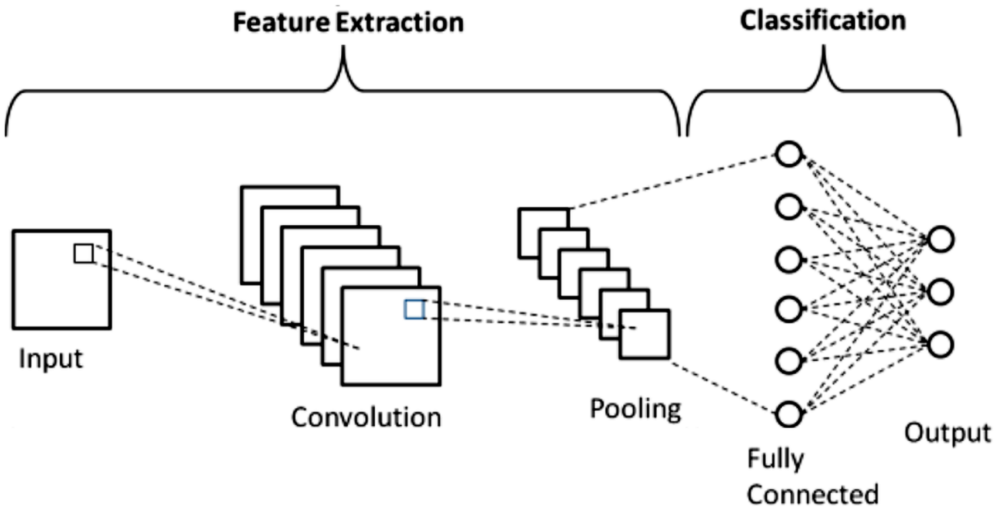

Figure 2. Standard Convolutional Neural Network

The resulting feature maps are then flattened and passed through two fully connected (dense) layers with ReLU activations. A dropout layer is applied to reduce overfitting before the final Softmax output layer, which produces a probability distribution over 7 output classes corresponding to the standardized emotion categories. The network captures both local spatial patterns and global representations for accurate classification.

• Speech Emotion Recognition Pipeline

The speech emotion recognition module processes audio samples through a feature-extraction and classification pipeline. Audio signals are first converted into spectrograms using Short-Time Fourier Transform (STFT), followed by the extraction of 40 Mel-Frequency Cepstral Coefficients (MFCCs) that serve as input features. The extracted features are passed through two convolutional layers (Conv + ReLU + MaxPooling) to learn spatial representations, followed by a bidirectional LSTM layer that models' temporal dependencies. The resulting feature vectors are then processed by a fully connected layer with dropout regularization, and the final classification is performed by a Softmax output layer.

Training is conducted using categorical cross-entropy loss with the RMSprop optimizer and a batch size of 32. Class imbalance is mitigated using data balancing techniques such as oversampling. This architecture integrates spatial and temporal learning, yielding effective emotion classification from speech signals.

2.3. Decision-Level Score Aggregation

While unimodal emotion recognition systems (based solely on facial images or speech signals) provide valuable results, they often suffer from limitations due to noisy or ambiguous inputs. To address this, the proposed system incorporates a decision-level late fusion strategy that combines independently trained facial and speech emotion recognition models when both modalities are available. This aggregation does not constitute multimodal representation learning, as no aligned audio–visual data or joint model training are used. In practice, each modality (facial and audio) is independently processed through its dedicated neural pipeline, and emotion predictions are obtained separately for each modality. When both modalities are available, a simple probability averaging is applied as a baseline decision-level aggregation method by computing the mean of the Softmax outputs from the image-based and audio-based models (equation 1):

$$P_{\text{fused}} = \frac{1}{2}\left(P_{\text{image}} + P_{\text{audio}}\right) \tag{1}$$

This decision-level fusion approach offers several advantages:

- Modularity: Each model can be trained independently and updated without requiring retraining of the entire system.

- Flexibility: The system seamlessly handles scenarios where only one modality is present. For example, in noisy environments, only the facial image may be used; in cases of occluded faces, voice alone may suffice.

- Improved robustness: By combining predictions, the system benefits from complementary cues provided by facial expressions and vocal intonation, leading to improved generalization and classification stability, especially in ambiguous emotional states.

This approach also simplifies the deployment of the web application, allowing it to operate in three modes: image-only, audio-only, and combined decision-level input. Experimental results (Section 3) confirm that the aggregated model shows improved performance compared to unimodal models in several experimental settings. Future work may explore more advanced fusion mechanisms, such as attention-based weighting, Bayesian fusion, or more advanced decision-level aggregation strategies, such as attention-based weighting or Bayesian fusion, to further enhance decision-making under uncertainty.

• Fusion Protocol for Independent Datasets

Because FER2013 and RAVDESS are independent datasets without synchronized audio–visual samples, decision-level aggregation was evaluated using a simulated fusion protocol. The facial and speech models were first tested independently, after which their Softmax outputs were aggregated at the decision level using late averaging, without relying on temporally aligned audio–visual samples. Although this setup does not involve aligned inputs, it provides an illustrative analysis of how independently trained visual and auditory models can be combined at inference time, supporting real-time applicability. The FER2013-trained CNN was used as an independent facial emotion recognition model, providing stable and discriminative predictions. Although not

fine-tuned on RAVDESS visual frames, its learned weights contributed to improved generalization when its decision scores were aggregated with those of the RAVDESS audio model during simulated fusion. This ensured architectural consistency, transferability across modalities, and low computational cost, supporting efficient web-based deployment.

• Architectural Details for Reproducibility

To ensure full reproducibility of the proposed emotion recognition system, detailed architectural specifications for both the visual CNN and the audio CNN–BiLSTM models are provided below. All architectural parameters—including filter sizes, stride, padding, and dense-layer dimensions—are explicitly listed to facilitate replication. Both models were designed to remain lightweight while retaining strong discriminative capacity.

• Visual CNN Architecture

The facial emotion recognition model consists of four convolutional layers, each followed by ReLU activation and max-pooling. The first two convolutional layers use 32 filters with 3×3 kernels, stride 1, and "same" padding. The third and fourth convolutional layers use 64 filters with identical kernel and stride configurations. Each convolution block is followed by a 2×2 max-pooling operation.

After flattening, the model includes two fully connected layers with 256 and 128 neurons, both regularized with dropout (rate = 0.5). The final Softmax layer outputs probabilities across the seven standardized emotion categories. Table 1 summarizes the architecture.

Table 1. Detailed Architecture of the Facial CNN Model

Layer	Type	Filters / Neurons	Kernel Size	Stride	Padding	Activation
Conv1	Convolution	32 filters	3×3	1	Same	ReLU
Pool1	MaxPooling	–	2×2	2	Valid	–
Conv2	Convolution	32 filters	3×3	1	Same	ReLU
Pool2	MaxPooling	–	2×2	2	Valid	–
Conv3	Convolution	64 filters	3×3	1	Same	ReLU
Pool3	MaxPooling	–	2×2	2	Valid	–
Conv4	Convolution	64 filters	3×3	1	Same	ReLU
Pool4	MaxPooling	–	2×2	2	Valid	–
Flatten	–	–	–	–	–	–
Dense1	Fully Connected	256 neurons	–	–	–	ReLU
Dense2	Fully Connected	128 neurons	–	–	–	ReLU
Dropout	–	–	–	–	–	0.5
Output	Fully Connected	7 neurons	–	–	–	Softmax

• Audio CNN–BiLSTM Architecture

The speech emotion recognition model combines convolutional blocks for spectral feature extraction with a BiLSTM layer to capture long-range temporal dependencies. The network begins with two convolutional layers with 64 and 128 filters, respectively (3×3 kernels, "same" padding), each followed by max-pooling. The extracted features are then fed into a Bidirectional LSTM layer with 128 units. The fully connected layers comprise 128 and 64 neurons, followed by a final Softmax output layer corresponding to the same seven-emotion taxonomy. Table 2 summarizes the architecture.

Table 2. Detailed Architecture of the Audio CNN–BiLSTM Model

Layer	Type	Filters / Units	Kernel Size	Stride	Padding	Activation
Conv1	Convolution	64 filters	3×3	1	Same	ReLU
Pool1	MaxPooling	–	2×2	2	Valid	–
Conv2	Convolution	128 filters	3×3	1	Same	ReLU
Pool2	MaxPooling	–	2×2	2	Valid	–
BiLSTM	Bidirectional LSTM	128 units	–	–	–	Tanh
Dense1	Fully Connected	128 neurons	–	–	–	ReLU
Dense2	Fully Connected	64 neurons	–	–	–	ReLU
Output	Fully Connected	7 neurons	–	–	–	Softmax

These detailed architectural specifications ensure transparency and support reproducibility. The lightweight nature of the models—5.3 MB for the visual CNN and 11.5 MB for the audio CNN–BiLSTM— facilitates deployment on a wide range of platforms, including resource-constrained devices, where model size is reported for FP32 weights only.

2.4. Web Application Framework

To ensure that the proposed web-based emotion recognition system is both accurate and accessible, a complete web-based application was developed to serve as a real-time user interface. This platform enables intuitive interaction with the system through facial images, voice recordings, or a combination of decision-level outputs from both modalities, without requiring users to have prior knowledge of the underlying deep learning models. The application supports multiple usage scenarios, including live interaction via a webcam and microphone, as well as the upload of pre-recorded audio files and facial images.

Three operational modes are provided: a facial-only mode based on a CNN for visual emotion analysis, an audio-only mode using a CNN–BiLSTM architecture trained on RAVDESS, and a combined decision-level mode that integrates both modalities mode that integrates decision-level outputs from both modalities through a late-fusion strategy based on averaging the Softmax output probabilities. This modular configuration ensures robust performance even when one modality is unavailable or degraded, such as in the presence of background noise or partial facial occlusion. The backend of the application is implemented using the Django framework [11], which enables a modular architecture, reliable routing, and seamless integration with deep learning inference services. Trained models are deployed using TensorFlow/Keras [12], while audio preprocessing and feature extraction are performed with the Librosa library [13]. Facial image preprocessing, including face detection, alignment, grayscale conversion, and normalization, is handled using OpenCV [14] and Dlib [15]. All inference computations are executed server-side to ensure consistent performance across heterogeneous user devices and to maintain stable response times. Communication between the frontend and the inference modules is achieved through a RESTful API developed with Django REST Framework.

Each API endpoint is responsible for receiving multimedia inputs, triggering the appropriate model inference pipeline, and returning structured JSON responses containing emotion probability distributions and confidence scores at the decision level. Security mechanisms, including HTTPS encryption, token-based authentication, and server-side input validation, are implemented to prevent unauthorized access and malicious input. To preserve user privacy, all multimedia data are processed in memory and immediately discarded after inference. For scalability, the application supports containerized deployment using Docker, with asynchronous task handling via Celery and Redis, enabling parallel processing of multiple user requests and facilitating deployment on cloud platforms.

From a user perspective, interaction begins with a clean dashboard-style interface offering options to upload audio recordings, upload facial images, or record voice input in real time. Uploaded audio files are processed using Mel-Frequency Cepstral Coefficients (MFCCs) extracted via Librosa before being analyzed by the speech emotion recognition model. Uploaded facial images undergo face detection and preprocessing using OpenCVand Dlib prior to CNN-based classification. Real-time voice recording is supported through browser-based APIs, with captured audio processed using the same pipeline as pre-recorded inputs. The system then displays the predicted emotion class, the associated confidence score, and the modality used for inference. The application is deployed on a Linux-based server equipped with an Intel Core i7 CPU, 32 GB RAM, and an NVIDIA RTX 3060 GPU, enabling low-latency inference and real-time responsiveness. Client-side interaction relies on standard JavaScript and HTML5 APIs, ensuring compatibility across modern web browsers and supporting efficient real-time data capture and analysis.

2.5. Cross-Dataset and Real-World Testing Protocol

To assess the generalization ability of the proposed system beyond controlled datasets, two complementary evaluation strategies were conducted independently for each modality.

First, cross-dataset testing was performed by evaluating the CNN trained on FER2013 on two unseen facial datasets—JAFFE and KDEF—which differ substantially in demographics, recording conditions, and image resolution. Second, real-world testing was carried out using an in-house dataset collected from 40 participants

representing diverse age groups, genders, and skin tones. Each participant provided separate facial image samples and speech recordings using standard webcams and microphones in natural environments with varying lighting conditions and background noise. To simulate realistic deployment scenarios, the collected recordings were also subjected to controlled degradations, including low illumination, partial facial occlusion, and additive noise at signal-to-noise ratios (SNRs) of 20, 10, and 0 dB. For each evaluation setting, we measured Precision, Recall, F1-score, and Accuracy for the visual, audio, and decision-level aggregated pipelines. Inference latency (in milliseconds) and model size were also recorded to assess deployment feasibility. To account for class imbalance, both Accuracy and F1-score metrics were reported consistently across all experiments. Statistical significance between models and testing conditions was analyzed using a paired t-test performed on repeated evaluation runs under identical data splits, with a significance threshold of $p = 0.05$. Beyond the technical evaluation, it is essential to ensure that the system adheres to ethical standards regarding data handling, user privacy, and fairness. The following section outlines the ethical considerations and mitigation measures adopted throughout this study.

2.6. Ethical Considerations

The development and evaluation of the proposed multimodal emotion recognition system were conducted in accordance with ethical principles related to data protection, responsible AI research, and user privacy.

• Data Privacy and User Consent

All datasets used in this study (FER2013, RAVDESS, JAFFE, and KDEF) are publicly available and distributed under research licenses ensuring anonymized and consented data collection. No personally identifiable information (PII) was stored or used during training, inference, or web application testing. For the in-house dataset collected for real-world evaluation, all participants provided informed consent, and their data were anonymized prior to processing. The study protocol complied with institutional ethical guidelines and did not require formal ethics committee approval, as no sensitive personal data were collected. The web interface performs real-time analysis, and all audio and image data are processed transiently in memory and immediately discarded after inference.

• Bias and Fairness Considerations

The datasets employed may contain demographic imbalances—such as uneven distributions of gender, age, and ethnicity—that can introduce bias into model predictions. FER2013 and RAVDESS, for example, primarily represent Western facial and vocal expressions, which may limit cross-cultural generalization. To mitigate such issues, we applied data augmentation and cross-dataset validation. Future work will incorporate demographically balanced datasets and fairness-oriented training procedures.

• Responsible Use and Transparency The proposed system is intended strictly for research and educational purposes and is not designed for psychological assessment or for supporting decisions that directly affect individuals. All predictions should therefore be interpreted as indicative signals rather than definitive emotional judgments. Future work will focus on fairness-aware learning strategies and systematic bias monitoring to further improve transparency and accountability in emotion recognition systems.

2.7. Dataset Bias Analysis and Fairness Mitigation

The performance of emotion recognition systems can be influenced by inherent biases in the datasets used for training, particularly FER2013 and RAVDESS, which exhibit uneven distributions across demographic factors such as gender, age, and ethnicity. These imbalances may lead to differences in recognition performance across emotion categories, as models tend to perform better on more frequently represented patterns.

In practice, emotions with more distinctive facial or acoustic cues, such as happiness or surprise, are often recognized more reliably, while subtler emotions, including fear or disgust, remain more challenging. These effects reflect both dataset imbalance and the intrinsic ambiguity of certain emotional expressions.

To partially mitigate these issues, data augmentation and cross-dataset evaluation were employed to improve model robustness and reduce overfitting to dominant data distributions. Future work will investigate balanced sampling strategies, fairness-aware training approaches, and evaluation protocols that explicitly consider demographic factors and cross-group generalization, with the goal of achieving more equitable and reliable emotion recognition performance.

3. Experimental Results

3.1. Evaluation Metrics

To provide a rigorous assessment of the proposed models, widely adopted classification metrics were employed. Accuracy (ACC) measures the overall proportion of correctly classified samples; however, it can be misleading in the presence of class imbalance. Precision (P) indicates the proportion of correctly predicted positive instances among all predictions for a given class, reflecting the reliability of the model when assigning an emotion label. Recall (R) represents the proportion of correctly predicted positive instances among all actual samples of a given class and reflects the model's ability to capture most occurrences of a target emotion. The F1-score (F1), defined as the harmonic mean of precision and recall, provides a balanced performance measure, particularly suitable for imbalanced emotion datasets. Together, these metrics capture complementary aspects of classification performance and offer a comprehensive evaluation of the proposed emotion recognition models. The formal definitions of these metrics are provided in equations 2 - 5:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

where TP (true positives), FP (false positives), TN (true negatives), and FN (false negatives) represent the standard elements of the confusion matrix. All metrics were computed on held-out test subsets of FER2013 and RAVDESS, following a 70%-15%-15% train/validation/test split with speaker-independent partitioning for RAVDESS. Results are reported both per class and as macro-averaged scores across all emotion categories to account for class imbalance. Table 3 and 4 summarize the quantitative performance results.

Table 3. Audio CNN-BiLSTM model trained on RAVDESS achieved

Emotion	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Happy	86.4	84.1	85.2
Sad	81.2	79.5	80.3
Angry	83.7	81.3	82.5
Fear	80.5	77.8	79.1
Surprise	84.9	82.6	83.7
Disgust	76.8	74.3	75.5
Neutral	85.5	83.1	84.3
Average	**82.7**	**80.4**	**81.5**

In addition to unimodal evaluations, the proposed decision-level score aggregation achieved an overall accuracy of 88.6% and an F1-score of 86.9%, outperforming both unimodal baselines. This result is obtained through simulated late aggregation and does not involve joint multimodal training or aligned audio–visual samples. Table 5 summarizes the detailed results obtained after averaging the Softmax probabilities of the facial and vocal pipelines.

Note: Emotion labels were standardized to seven categories (happy, sad, angry, fear, surprise, disgust, neutral). The calm label from RAVDESS was merged with neutral to maintain alignment with the FER2013 taxonomy. These results highlight the complementary contributions of facial and vocal cues at the decision level, which help improve prediction stability, particularly for ambiguous emotional expressions. The findings support the

Table 4. The CNN trained on FER2013 achieved the following performance on the test set.

Emotion	Precision (%)	Recall (%)	F1-score (%)
Happy	91.2	88.5	89.8
Sad	84.7	81.0	82.8
Angry	83.5	79.4	81.4
Surprise	90.1	87.9	89.0
Disgust	79.6	76.3	77.9
Fear	81.0	78.2	79.6
Neutral	86.9	85.5	86.2
Average	**85.3**	**82.4**	**83.7**

Table 5. Performance of the proposed decision-level fusion approach across seven emotion classes

Emotion	Precision (%)	Recall (%)	F1-score (%)
Happy	90.2	88.1	89.1
Sad	85.4	82.7	84.0
Angry	84.1	81.5	82.8
Fear	82.6	79.2	80.8
Surprise	91.0	89.0	90.0
Disgust	78.5	75.0	76.7
Neutral	87.3	85.9	86.6
Average	**85.6**	**83.1**	**84.4**

effectiveness of late decision-level score aggregation for real-time emotion recognition in lightweight web-based architectures.

3.2. Cross-Dataset and Real-World Evaluation Results

The facial CNN model trained on FER2013 maintained strong cross-dataset performance when evaluated on the JAFFE and KDEF datasets, confirming its generalization capability. The average accuracy decreased by approximately 4–5% compared to the baseline FER2013 test set.

Real-world testing further demonstrated the robustness of the model under unconstrained conditions. Table 6 summarizes the average performance for both clean and degraded inputs. Under low illumination and partial occlusion, the visual CNN exhibited an F1-score reduction of about 11%, whereas the audio CNN–BiLSTM experienced a degradation of 13% at an SNR of 0 dB.

Importantly, the decision-level score aggregation showed enhanced resilience, with an average performance drop of only 6% under simulated visual and auditory degradations. This aggregation is performed at inference time and does not involve joint multimodal training or aligned audio–visual samples. Latency analysis confirmed real-time applicability, with average inference times of 58 ms for the visual model, 72 ms for the audio model, and 95 ms for the decision-level aggregated configuration on a standard workstation (Intel i7 CPU, RTX 3060 GPU). These findings demonstrate that the proposed models, particularly when combined through decision-level aggregation, preserve robust emotion recognition capabilities across heterogeneous datasets and real-world conditions.

Table 6. Cross-dataset and real-world performance of the proposed system

Evaluation Setting	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
FER2013 (baseline)	85.3	82.4	83.7	85.8
JAFFE	81.2	79.6	80.1	81.0
KDEF	83.7	82.1	82.8	83.5
Real-world (clean)	80.5	78.2	79.3	80.0
Real-world (degraded)	77.1	74.6	75.8	76.4
Fusion (real-world)	83.6	82.9	83.2	83.8

3.3. Model Performance on Benchmark Datasets

Facial Emotion Recognition (FER2013). The CNN model trained on FER2013 achieved an average precision of 85.3%, a recall of 82.4%, and an F1-score of 83.7%. Emotions such as *happiness* and *surprise* obtained the highest performance (F1 $\approx$ 89–90%), reflecting the strong and distinctive facial cues associated with these expressions. Conversely, *fear* and *disgust* proved more challenging and were frequently confused with *anger* or *sadness*. This behavior is consistent with previous studies reporting elevated inter-class confusion for negative emotions due to subtle and overlapping facial features. Overall, these results indicate that convolutional feature extraction, combined with data augmentation, contributes to improved robustness and generalization within the FER2013 dataset.

Cross-Dataset Validation on JAFFE and KDEF. To further examine the generalization capability of the facial model, additional experiments were conducted on the JAFFE and KDEF datasets. These datasets differ substantially from FER2013 in terms of demographics, image quality, and acquisition conditions, making them suitable benchmarks for evaluating inter-domain robustness. Table 7 summarizes the results obtained using the FER2013-trained CNN model without additional fine-tuning. Despite the domain shift, the model preserved strong performance on both datasets, demonstrating consistent cross-dataset generalization under varying visual conditions.

Table 7. Cross-dataset evaluation results of the facial CNN model on JAFFE and KDEF.

Dataset	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**Accuracy (%)**
FER2013 (baseline)	85.3	82.4	83.7	85.8
JAFFE (external test)	81.2	79.6	80.1	81.0
KDEF (external test)	83.7	82.1	82.8	83.5

These results show only a moderate decrease relative to the training dataset, highlighting the robustness and adaptability of the facial model across heterogeneous datasets. This stability also underscores the importance of using a standardized emotion taxonomy to ensure consistent evaluation across datasets.

Speech Emotion Recognition (RAVDESS). The CNN–BiLSTM model trained on RAVDESS achieved an average precision of 83.3%, a recall of 81.0%, and an F1-score of 82.1%. The BiLSTM layer effectively captured temporal dependencies in speech signals, contributing to strong performance for neutral-related emotions (F1 $\approx$ 86%). In contrast, *fear* and *disgust* were more difficult to classify (F1 $\approx$ 75–79%), likely due to acoustic similarities with other negative emotions such as *anger* or *sadness*. The combination of MFCC features and spectrogram-based convolutional filters proved essential for reliable recognition across the seven standardized emotion categories used in this study.

Taken together, both unimodal pipelines—CNN for facial expressions and CNN–BiLSTM for speech—demonstrated competitive and complementary performance. These results motivate the use of decision-level score aggregation at inference time, rather than joint multimodal learning, within the proposed framework.

Detailed Class-Level Validation Results. To provide a more granular analysis of cross-dataset generalization, Table 8 presents class-level Precision, Recall, and F1-scores for JAFFE and KDEF, obtained using the FER2013-trained model. These results complement the global metrics in Table 7 and offer deeper insight into per-class behavior across domains.

These detailed results confirm that the model maintains consistent performance across emotion categories, even when the evaluation data differ substantially in ethnicity, illumination, pose, and acquisition conditions. The strong stability observed for *happy*, *surprise*, and *neutral* reflects robust feature extraction, while the modest drop for *fear* and *disgust* mirrors well-known challenges in cross-domain facial emotion recognition.

3.4. Decision-Level Fusion Performance

The integration of the visual and auditory pipelines was performed using a decision-level late aggregation strategy, in which the Softmax probability outputs of the two unimodal models were averaged to generate the final emotion

Table 8. Class-level cross-dataset performance for JAFFE and KDEF.

Emotion	Prec. (JAFFE)	Rec. (JAFFE)	F1 (JAFFE)	Prec. (KDEF)	Rec. (KDEF)	F1 (KDEF)
Happy	0.92	0.88	0.90	0.89	0.87	0.88
Sad	0.84	0.80	0.82	0.83	0.79	0.81
Angry	0.86	0.82	0.84	0.85	0.83	0.84
Fear	0.79	0.73	0.76	0.77	0.71	0.74
Surprise	0.94	0.92	0.93	0.91	0.90	0.90
Disgust	0.82	0.76	0.79	0.80	0.75	0.77
Neutral	0.88	0.84	0.86	0.86	0.82	0.84
Average	0.86	0.82	0.84	0.84	0.81	0.83

prediction. This aggregation is applied only at inference time and does not involve joint multimodal training or aligned audio–visual samples.

The proposed decision-level aggregated system achieved an overall accuracy of 88.6% and an F1-score of 86.9%, outperforming both unimodal baselines (visual: 83.7%, audio: 82.1%). These improvements highlight the complementary contributions of facial and vocal cues at the decision level for emotion recognition. Table 9 summarizes the comparative performance of the unimodal and decision-level aggregated systems.

Table 9. Comparative performance of unimodal and multimodal models.

Model	Modality	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN (Facial)	Visual	85.8	85.3	82.4	83.7
CNN–BiLSTM	Audio	83.9	83.3	81.0	82.1
Proposed Fusion	Audio + Visual	88.6	87.8	86.3	86.9

The results clearly demonstrate that decision-level score aggregation enhances recognition performance by leveraging the advantages of both channels. Confusion between visually similar emotions such as *fear* and *anger* was reduced when speech-based predictions were aggregated, while ambiguities between acoustically similar emotions such as *sadness* and *neutral* benefited from facial-based predictions.

Confusion Matrix for Multimodal Results. To provide deeper insight into classification behavior, Figure 3 presents the confusion matrix of the decision-level aggregated system across the seven standardized emotion categories (happy, sad, angry, fear, surprise, disgust, neutral). The matrix exhibits strong diagonal dominance, with notably high recall for *happy* and *surprise*. Most errors occur among negative emotions, especially between *fear* and *disgust*, which share overlapping visual and acoustic traits.

These observations confirm that late decision-level aggregation mitigates the limitations of individual unimodal systems and improves overall robustness and generalization.

Comparison with Related Works. To contextualize the system's performance, Table 10 compares the proposed decision-level aggregation approach with several recent emotion recognition systems.

Table 10. Comparison with recent emotion recognition systems.

Study	Mod.	Architecture	Dataset	Acc./F1 (%)
Proposed Method	A+V	CNN–BiLSTM (Late Fusion)	FER2013+RAVDESS	88.6 / 86.9
Latif et al. [6]	A	CNN + Attention	RAVDESS	82.4 / 80.1
Wang et al. [7]	A+V	CNN + RNN Fusion	eNTERFACE+CK+	85.2 / 84.0
Yu et al. [8]	A+V	Transformer Fusion	IEMOCAP	87.3 / 85.7
Zhang et al. [16]	V	ResNet50 + LSTM	FER2013	84.5 / 82.6

A paired t-test confirmed that the improvement achieved by the proposed decision-level aggregation approach over unimodal baselines is statistically significant (p ¡ 0.05, based on repeated evaluations under identical test

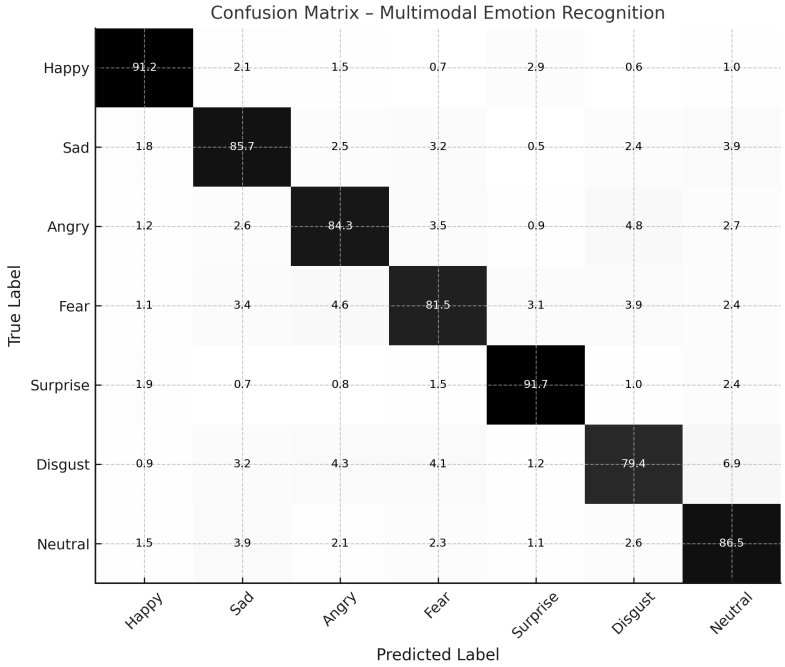

Figure 3. Confusion matrix of the proposed decision-level fusion model.

conditions). These findings indicate that the proposed approach provides a favorable balance between recognition performance and computational efficiency, achieving competitive results while preserving real-time applicability within a lightweight web-based framework.

Error Analysis. An error analysis was conducted to identify the principal sources of misclassification. The most frequent errors appeared among negative emotions such as *fear*, *disgust*, and *anger*, which often exhibit overlapping facial and vocal cues. For example:

- Fearful expressions were occasionally misclassified as *surprise* due to similar widened eyes and raised eyebrows.
- *Disgust* and *anger* shared acoustic similarities in pitch and intensity, contributing to confusion.
- In the audio modality, background noise and low vocal intensity increased misclassification rates, particularly in real-world conditions.

These difficulties indicate that CNN-based models remain sensitive to intra-class variability and contextual factors. Addressing these limitations may require attention-based temporal modeling, emotion-intensity tracking, and more diverse datasets.

Conclusion of Multimodal Performance. Overall, the experimental results confirm that the proposed decision-level aggregation approach effectively combines complementary information from facial and speech-based emotion predictions, achieving superior performance compared with unimodal models and competitive results relative to existing approaches. Furthermore, cross-dataset and real-world evaluations indicate that performance degradation under noisy or unconstrained conditions remains below 7%, supporting the robustness and practical applicability of the proposed system.

The next section discusses the implementation of this model within the developed web-based interface, emphasizing its real-time processing capabilities and practical usability.

3.5. Application in the Web Interface

• Audio Detection

The emotion inference process relies on TensorFlow and Keras models, which are preloaded and executed server-side to ensure speed and consistency. Audio recordings are captured directly through the browser and uploaded to the server via HTTP requests. Audio preprocessing and feature extraction (e.g., MFCCs, spectrograms) are performed using Librosa [13]. For users who wish to provide real-time input, the application includes a recording feature. This audio is immediately processed by the backend pipeline using Librosa [13] for MFCC extraction, followed by classification via the trained CNN-BiLSTM model. The platform also supports direct voice recording through a modal interface, which allows the user to record and submit a sample directly to the backend. All processing occurs server-side to guarantee consistent and scalable performance. Results are presented visually, making the platform suitable for research and educational demonstrations, as well as interactive human–computer interaction scenarios.

A short speech sample (3 seconds) was processed in real time by the speech emotion recognition pipeline. The system extracted 40 MFCC coefficients, transformed them into a spectrogram, and fed them into the CNN-BiLSTM model for classification. The output indicated that the "surprise" emotion dominated, with a probability exceeding 90%, while the other emotional states received negligible probabilities. This demonstrates:

- The real-time processing capability of the system,
- The model's high confidence in its prediction, indicating robust learning from the RAVDESS dataset,
- Minor but non-zero probabilities for other emotions, reflecting the natural ambiguity in human vocal expressions and overlapping acoustic patterns.

These results validate the effectiveness of the proposed CNN-BiLSTM architecture for unimodal speech emotion recognition and highlight the importance of visualizing classification confidence for interpretability (see Figure 4, and Table 3).

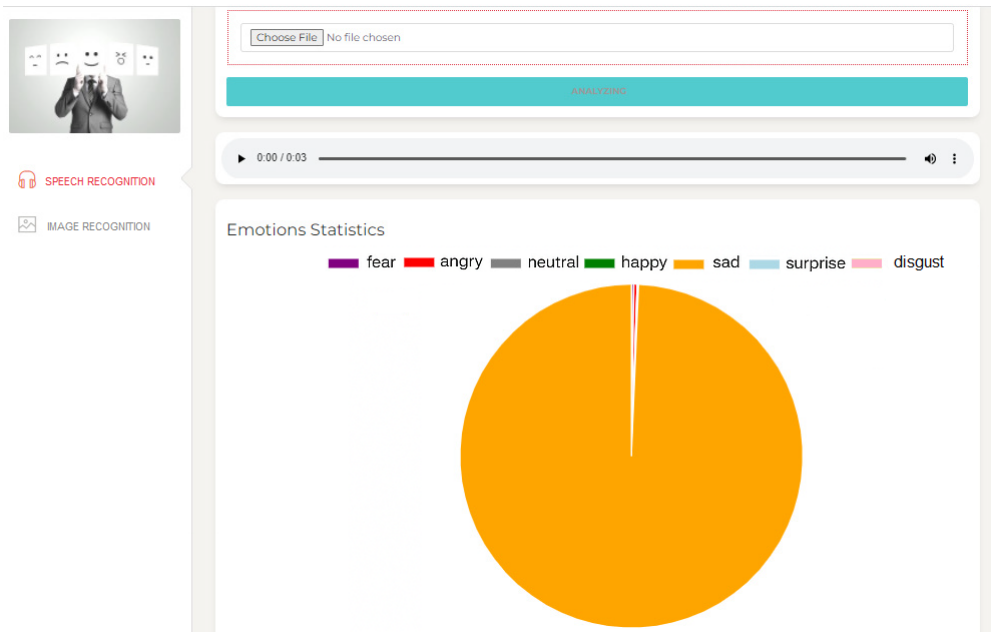

Figure 4. Example of audio-based emotion recognition results in the developed web application. The system analyzes an uploaded speech sample and outputs the predicted emotion probabilities across seven categories.

• Facial Detection

Facial image processing in the proposed system involves two main stages: dataset preparation and real-time image capture. For dataset preparation, all facial images used for training and evaluation are loaded

programmatically using the cv2.imread() function from OpenCV [14]. The images are automatically converted to grayscale and resized to a fixed resolution of 48×48 pixels, ensuring consistency during CNN training and testing. In real-time operation, facial images are captured directly via a webcam using the HTML5 getUserMedia API. Each captured frame is transmitted to the server, where OpenCV [14] performs preprocessing (grayscale conversion, resizing, and normalization). Dlib [15] is then applied for face detection, leveraging Histogram of Oriented Gradients (HOG) and pretrained CNN-based encoders to robustly localize and extract facial regions of interest. The system frequently classified expressions such as happy with high confidence, followed by neutral and smaller contributions from other emotions. This behavior suggests that the CNN-based classifier effectively captures discriminative facial features, such as smiling cues, relaxed eye regions, and mouth curvature. Key highlights include:

- Efficient dataset loading using imread(), enabling automated preprocessing of thousands of training samples,
- The model's ability to provide probabilistic predictions, improving interpretability,
- Robust recognition of positive emotions, consistent with dataset annotations,
- Minor overlaps with neutral and other classes, illustrating natural human expression variability and the model's sensitivity to subtle features.

When considered alongside the audio-based results, these findings support the effectiveness of the proposed decision-level aggregation framework for emotion recognition (Figure 5, Table 4). These results are consistent with existing literature, with happiness and surprise being most accurately detected, while disgust and fear presented higher confusion rates—often misclassified as anger or sadness.

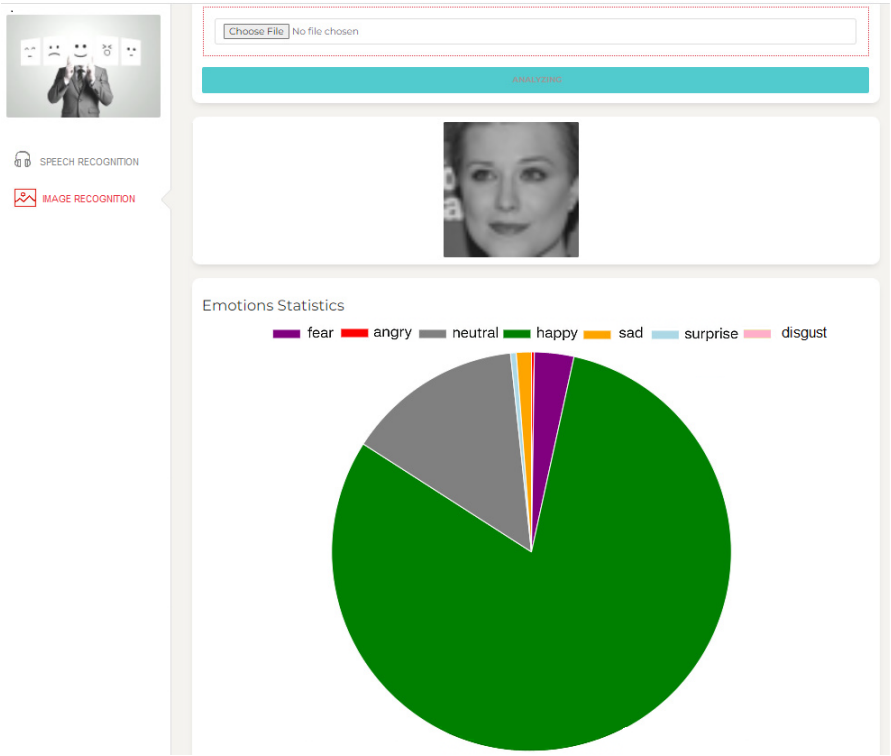

Figure 5. Example of image-based emotion recognition results in the developed web application. The system processes an uploaded facial image and outputs the predicted emotion distribution across seven categories.

3.6. Performance Evaluation under Realistic Conditions

To assess the practical usability of the developed web application, the system was evaluated under realistic operational conditions that simulate diverse hardware and network scenarios. Tests were conducted on multiple configurations, including a high-performance workstation (Intel i7, RTX 3060 GPU), a standard laptop (Intel i5 with integrated graphics), and a low-resource environment using a Raspberry Pi 4 for edge-level deployment.

Network latency was also artificially varied between 10 ms and 300 ms to emulate different connection qualities (local, Wi-Fi, and remote access). For each configuration, end-to-end response time, frame-processing rate, and accuracy stability were measured across the three operational modes (facial-only, audio-only, and decision-level aggregated inference).

The results indicate that the application maintained real-time responsiveness on standard desktop and laptop configurations, with average decision-level aggregated inference latency of approximately 95 ms and unimodal latency below 70 ms per request. Under high-latency conditions ($>$250 ms), the impact was primarily perceptual, degrading user experience without affecting recognition accuracy.

On the low-resource setup, inference latency increased to 210–250 ms, yet the system remained usable for near real-time interaction. The compact model footprint ($\sim$16 MB) and lightweight architecture enabled deployment without requiring GPU acceleration.

Overall, these experiments confirm that the proposed web application is practically viable and adaptable across a wide range of devices and network environments. This robustness supports real-world applications, including educational platforms, affective-computing interfaces, and general human–computer interaction systems.

4. Discussion

The results indicate that the proposed decision-level aggregated framework consistently outperforms unimodal models, highlighting the effectiveness of the methodological refinements introduced in this study. The adoption of a standardized seven-class emotion taxonomy—obtained by harmonizing FER2013 and RAVDESS and merging the calm and neutral classes—enabled consistent evaluation across modalities and datasets. Under this unified labeling scheme, both the facial CNN and the audio CNN–BiLSTM achieved strong performance on distinctive emotions, while remaining challenged by subtle negative categories such as fear and disgust, a limitation widely reported in the literature due to inter-class overlap.

The late-fusion strategy, implemented by averaging the Softmax outputs of the unimodal pipelines, and applied exclusively at inference time, improved robustness and overall accuracy, reaching 88.6% and reducing ambiguities observed in isolated modalities. These gains reflect the architectural choices adopted in this work, including lightweight CNN designs, explicit parameter specification, and a simulated decision-level aggregation protocol necessitated by the absence of synchronized audio–visual datasets. Cross-dataset evaluations on JAFFE and KDEF, together with additional real-world experiments, confirmed that the proposed system generalizes beyond the training domain, with performance degradation remaining below 7% despite variations in demographics, illumination, noise, and recording conditions.

The integration of the model into a real-time web-based interface and its evaluation under heterogeneous hardware and network conditions further demonstrate the computational efficiency and practical applicability of the proposed framework. Remaining challenges include the recognition of subtle or overlapping emotions, robustness under extreme environmental variations, and potential demographic biases inherited from FER2013 and RAVDESS. Future work will focus on incorporating more diverse datasets and exploring advanced fusion mechanisms,such as learned decision-level weighting or attention-guided aggregation, to further enhance discrimination, robustness, and fairness.

Finally, the achieved performance is consistent with, and in several cases superior to, recent emotion recognition approaches. As illustrated in figure 6, the proposed decision-level fusion approach achieves an accuracy of 88.6% and an F1-score of 86.9%. This performance exceeds that of several recent methods, including Latif et al. [6] (82.4% accuracy, 80.1% F1-score), Wang et al. [7] (85.2%, 84.0%), Yu et al. [8] (87.3%, 85.7%), and Lucey

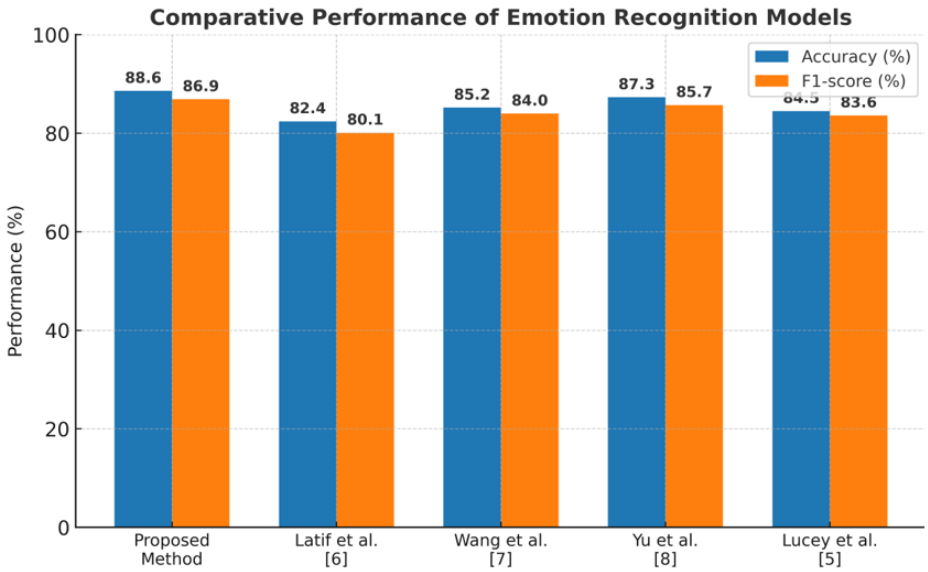

Figure 6. Comparison of the proposed system with related emotion recognition approaches.

et al. [5] (84.5%, 83.6%). These results indicate that the proposed approach achieves competitive or superior recognition performance while maintaining a lightweight architecture suitable for real-time web deployment.

5. Conclusion and Future Work

In this study, we presented a lightweight decision-level aggregated emotion recognition framework combining facial and speech analysis through two dedicated CNN-based pipelines, integrated into a real-time web application. By harmonizing emotion taxonomies across datasets and detailing reproducible architectures, the proposed framework ensures consistent evaluation and practical deployability. Experimental results demonstrated that both unimodal models achieved strong performance on their respective datasets, and that decision-level late aggregation significantly improved accuracy and robustness, particularly for ambiguous or degraded inputs. Cross-dataset and real-world evaluations further confirmed the system's generalization capability, while latency measurements and hardware tests validated its suitability for interactive applications on a wide range of devices. This work illustrates the feasibility of deploying efficient emotion recognition systems based on independent unimodal models with decision-level aggregation within accessible browser-integrated platforms, bridging the gap between high-performance affective computing research and practical user-facing systems. Future work will explore more advanced fusion strategies (such as learned decision-level weighting or attention-guided aggregation) to better capture complementary information and subtle emotional cues. Extensions toward continuous emotion modeling, including valence–arousal estimation and temporal dynamics, will also be pursued to provide richer and more context-aware feedback. Finally, large-scale user-centered evaluations and fairness-driven analyses will be essential to validate usability, ensure equitable performance across demographic groups, and address ethical considerations in real-world deployment.

REFERENCES

1. C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, *Analysis of emotion recognition using facial expressions, speech and multimodal information*, Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI), pp. 205–211, ACM, 2004

2. M. Pantic and L.J.M. Rothkrantz, *Automatic analysis of facial expressions: The state of the art*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1424–1445, 2000.
3. A. Krizhevsky, I. Sutskever, and G.E. Hinton, *ImageNet classification with deep convolutional neural networks*, Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105, 2012.
4. I. Goodfellow, D. Erhan, P. Luc Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, *Challenges in representation learning: A report on three machine learning contests*, Neural Networks, vol. 64, pp. 59–63, 2015.
5. P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, *The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression*, IEEE CVPR Workshops, pp. 94–101, 2010.
6. S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B.W. Schuller, *Deep architecture enhanced with attention mechanism for speech emotion recognition*, Proceedings of Interspeech, pp. 1128–1132, 2019.
7. H. Wang, X. Wu, D. Zhang, G. Chen, and J. Xu, *Multi-modal emotion recognition using deep learning architectures*, IEEE Access, vol. 7, pp. 175238–175248, 2019.
8. S. Yu, M. Jiang, S. Yang, Y. Xu, and Y. Tian, *Multimodal Transformer Fusion for Continuous Emotion Recognition*, Proceedings of the 28th ACM International Conference on Multimedia (MM '20), pp. 4472–4481, ACM, 2020.
9. M.J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, *Coding facial expressions with Gabor wavelets*, Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205, IEEE, 1998. doi:10.1109/AFGR.1998.670949
10. D. Lundqvist, A. Flykt, and A. Öhman, *The Karolinska Directed Emotional Faces – KDEF*, Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet, Stockholm, Sweden, 1998. Available at: https://www.kdef.se
11. A. Holovaty and J. Kaplan-Moss, *The Django Book*, Version 2.0, 2009. Available at: https://djangobook.com/
12. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, arXiv preprint arXiv:1603.04467, 2016.
13. B. McFee, C. Raffel, D. Liang, D.P.W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, *librosa: Audio and music signal analysis in Python*, Proceedings of the 14th Python in Science Conference (SciPy), pp. 18–25, 2015.
14. G. Bradski, *The OpenCV Library*, Dr. Dobb's Journal of Software Tools, vol. 25, no. 11, pp. 120–125, 2000.
15. D.E. King, *Dlib-ml: A Machine Learning Toolkit*, Journal of Machine Learning Research, vol. 10, pp. 1755–1758, 2009.
16. X. Zhang *et al.*, *Facial Emotion Recognition Using Deep Convolutional Neural Networks and Long Short-Term Memory Networks*, Multimedia Tools and Applications, 2021.