IRJCS: International Research Journal of Computer Science
Volume 12, Issue 10, November 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1210.05

Emotion Recognition with Deep Learning

Sowjanya S
Department of Artificial Intelligence and Machine Learning
Vemana Institute of Technology,Bangalore, India
s.sowjanya@vemanait.edu.in
https://orcid.org/0009-0005-9580-7150
Nisarga S,Taiyaba, Sayeeda Um Salma,V Archana
Department of Artificial Intelligence and Machine Learning
Vemana Institute of Technology
Bangalore, Karnataka, India
nisargas.ai2022@vemanait.edu.in taiyaba.ai2022@vemanait.edu.in
sayeedaumsalma.ai2022@vemanait.edu.in, varchana.ai2022@vemanait.edu.in

Publication History
Manuscript Reference: IRJCS/RS/Vol.12/Issue10/NVCSX10084
Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.12/Issue10/NVCSX10084
Received:23,October 2025, Revised: 09, October 2025, Accepted: 31October 2025 Published Online: 21November 2025
https://www.irjcs.com/volumes/Vol12/iss-10/05.NVCSX10084.pdf
Article Citation:Sowjanya,Nisarga,Taiyaba,Sayeeda,Archana(2025),Emotion Recognition with Deep Learning**,** IRJCS: International Research Journal of Computer Science, Volume 12, Issue 09 of 2025 pages 587-594
doi:> https://doi.org/10.26562/irjcs.2025.v1210.05
BibTeX Sowjanya@2025Emotion
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2025, ISSN 2393-9842, https://doi.org/10.26562/ijrcs.2025.v1210.05 The journal's official abbreviation is IRJCS.
Orcid: https://orcid.org/0009-0004-9398-7488

Abstract: This paper presents a unified Flask-based web application for multimodal emotion recognition, integrating speech, facial expression, and text analysis into a single framework.The system employs state-of-the-art deep learning and transformer-based models to enable accurate, real-time emotion detection across diverse input modalities. For audio-based recognition, CNN–LSTM architecture is used to perform multilingual speech emotion analysis, outperforming traditional approaches that rely on MFCC and chroma features. The facial expression module leverages a Mini-Xception convolutional neural network, which is robust to variations in lighting, pose, and occlusion; it also incorporates a temporal smoothing mechanism to stabilize predictions in live video streams. The text analysis module utilizes a DistilRoBERTa transformer to interpret nuanced emotional cues in written input. All three modalities are integrated into the Flask web interface supporting webcam video streaming, audio recording, and text entry. User data management and authentication are handled securely via a MySQL database. By combining convolutional and transformer architectures in a multimodal fusion framework, the proposed system achieves enhanced accuracy, adaptability, and consistency across modalities. Potential applications include affective computing, mental health monitoring, virtual assistants, and empathetic human–computer interaction, thereby improving the emotional responsiveness of automated systems.
Keywords: Emotion Recognition, Deep Learning, Speech Processing, Facial Expression Analysis, Natural Language Processing, Multimodal AI, Flask, Transformer models.

I. INTRODUCTION

In recent years, human–computer interaction has shifted from simple command-driven systems to more intuitive interfacesthat can understand speech, text, and visual cues. Despite these advancements, many systems still struggle to interpret how a user feels. Emotional cues play a major role in communication and often influence how people speak, react, and make decisions. These cues appear not only in spoken words but also in tone, facial expressions, and written language. Because of this, the ability to recognize emotions has become an important part of creating systems that can interact with users more naturally. To address this need, researchers have been developing Emotion Recognition Systems (ERS) that allow computers to sense and interpret human emotions. Such systems aim to make interactions smoother, more supportive, and more empathetic. This work focuses on building a multimodal emotion recognition system that combines speech, facial expressions, and text analysis. Each of these modes captures different aspects of emotional behaviour, and combining them helps reduce the weaknesses that appear when relying on only one source.
For each type of input, a separate model is applied.

Emotions in speech are identified with a CNN–LSTM model, facial expressions are processed using a Mini-XCEPTION network, and emotional meaning in written text is interpreted through a BERT-based transformer.The CNN–LSTM model is trained on the RAVDESS dataset to capture spatial–temporal patterns in speech, while the Mini-XCEPTION model utilizes the FER-2013 dataset for real-time facial expression analysis. The system is implemented as a Flask-based web application with a MySQL backend for user authentication and data management.

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 10, November 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1210.05

Users can record or upload audio, input text, or enable a webcam to detect emotions in real time. Each model outputs an emotion label with an associated confidence score, creating a unified emotion profile across modalities. Preprocessing techniques such as noise reduction, normalization, and MFCC-based feature extraction ensure consistent input quality, while regularization methods like dropout, batch normalization, and learning rate scheduling enhance model robustness. This multimodal framework has diverse real-world applications in healthcare, education, customer service, security, and virtual assistants, where understanding emotions enhances system responsiveness and personalization. The architecture is designed for scalability and future extensibility, allowing integration of additional modalities such as physiological or gesture-based signals.

II. PROBLEM STATEMENT

Traditional human–computer interaction systems lack the ability to perceive and interpret human emotions, leading to impersonal and inefficient responses. This project addresses the challenge of developing an intelligent, multimodal emotion recognition system capable of accurately detecting emotions from speech, facial expressions, and text using deep learning models (CNN, LSTM, and BERT) to enable more natural, empathetic, and context-aware interactions.

III. OBJECTIVE

The primary objective of this research is to develop a multimodal emotion recognition system capable of accurately detecting human emotions from speech, facial expressions, and text using deep learning. The specific goals include: Designing a robust hybrid CNN–LSTM model for speech emotion recognition, leveraging acoustic features such as MFCCs, pitch, and spectral properties. Implementing facial emotion recognition using a Mini-XCEPTION model trained on FER-2013 and text-based emotion detection using atransformer for contextual understanding.Comparing and evaluating multiple models (SVM, Random Forest, CNN, LSTM) to identify the most accurate and efficient architecture. Optimizing performance through feature selection, dimensionality reduction, and regularization techniques to improve generalization and reduce computational load. Testing the system on diverse, emotion-labeled datasets (e.g., RAVDESS, FER-2013) under varying acoustic and visual conditions to ensure robustness. Demonstrating real-world applicability in domains such as healthcare, education, customer service, and virtual assistants, emphasizing emotionally intelligent and adaptive human–machine interaction.

IV. RELATED WORK

Recent work in speech and multimodal emotion recognition (ER) has advanced through deep architectures, specialized feature learning, and fusion strategies that boost robustness and accuracy across datasets.

Zhang and Xue [1] introduced an autoencoder with emotion embeddings to enhance speech feature representation, improving separability of affective states without manual feature engineering—an approach that complements end-to-end CNN–LSTM pipelines by learning compact, discriminative latent spaces for SER. [1]

Subramanian et al. [2] proposed a digital twin framework for real-time, personalized ER in healthcare, emphasizing system integration, latency-aware inference, and user-centric deployment—principles directly aligned with building reliable, privacy-conscious multimodal systems for production use. [2]

Khurana et al. [3] presented RobinNet, a multimodal SER system augmented with speaker recognition to improve social interaction contexts, demonstrating that identity-aware modeling can stabilize predictions and reduce confounds in conversational settings. [3]

Gursesli et al. [4] evaluated lightweight CNNs for facial emotion recognition across public datasets, showing that compact models can achieve strong performance with efficient inference—supporting choices like Mini-XCEPTION for real-time facial ER on edge or web applications. [4]

Almulla [5] designed a multimodal deep CNN system, fusing speech and facial cues to outperform unimodal baselines, underscoring that cross-modal complementarity yields more reliable emotion predictions under noisy or ambiguous inputs. [5]

Priyadharshini et al. [6] demonstrated hybrid CNN–LSTM architectures for enhanced speech emotion and gender recognition, validating temporal modeling on mel-spectrogram inputs and motivating sequence-aware backbones for richer prosodic and temporal dynamics. [6]

Chowdhury et al. [7] introduced a lightweight ensemble using handcrafted features (MFCC, ZCR, Chroma, RMS) combined with 1D CNN and CNN–BiLSTM, achieving high accuracy with low compute—useful for resource-constrained deployments and as complementary features in late fusion. [7]

Madanian et al. [8] proposed a multi-dilated convolution network capturing long- and short-term patterns without recurrence, reducing latency while maintaining accuracy—an attractive alternative for real-time SER where responsiveness is critical. [8]

Manolekshmi and Mukunthan [9] employed hybrid deep learning with ensemble strategies to improve generalization across datasets, highlighting the value of model diversity and majority voting for stability in cross-domain ER tasks. [9]

Sareen and Seeja [10] leveraged mel-spectrogram image representations with CNNs to reach near-saturated accuracy on TESS, reinforcing the effectiveness of time–frequency visual encodings but also revealing limitations in temporal context modeling without sequence-aware layers. [10] Collectively, these studies trace a clear trajectory from feature-engineered, audio-only SER to multimodal, deep, and deployment-ready frameworks. Building on these advances, the present work integrates CNN–LSTM for speech, DistilRoBERTa for text, and Mini-XCEPTION for facial ER with multimodal fusion, targeting high accuracy, robustness, and real-time performance in a unified, scalable system suitable for practical applications.

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 10, November 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1210.05

V. METHODOLOGY

1. Audio Processing Module

The audio module standardizes all speech signals to 22050 Hz and uses MFCC, Mel-spectrogram, Chroma, Contrast, and Tonnetz features. These features are converted into fixed-length sequences of shape (858 × 129) and processed by a CNN–LSTM deep learning architecture. The model is trained separately on Hindi, Kannada, and English (RAVDESS) datasets to build three specialized models with higher accuracy.

2. Text Processing Module

The text processing module accepts user-typed or uploaded text in Kannada, Hindi, or English. The input first undergoes language detection, after which non-English sentences are automatically translated into English for uniform processing. The cleaned text is passed into a DistilRoBERTa-based emotion classification model, which generates contextual embeddings and evaluates the emotional tone of the sentence. The model outputs fine-grained emotion probabilities such as anger, disgust, fear, joy, sadness, surprise, and neutral, which are then mapped to the system's final emotion categories. This module relies entirely on apretrained transformer pipeline and does not require any additional CNN or LSTM layers, ensuring highly accurate emotion prediction across multiple languages.

3. Facial Emotion Module

The facial module captures real-time video using a webcam or processes uploaded video clips. Using OpenCV , faces are detected, cropped, and resized before being passed to the Mini-Xception CNN model. This module identifies facial expressions such as happiness, sadness, anger, fear, and surprise through deep visual feature extraction.

4. Feature Extraction & Fusion Module

This module extracts essential features from each modality—acoustic features from speech, transformer embeddings from text, and CNN feature maps from facial images. A fusion mechanism combines these modality-specific outputs using weighted averaging, producing a single unified emotion label along with a confidence score, ensuring higher accuracy than unimodal systems.

5. Deep Learning Model Module

Each modality is processed using optimized deep learning architectures including CNN, LSTM, and attention layers. These models capture spatial, temporal, and contextual emotional cues. The system is trained using Adam optimizer with regularization techniques such as dropout and batch normalization to ensure stable and generalized learning across datasets.

6. Backend & Storage Module

A Flask-based backend handles user requests, authentication, and model inference using RESTful APIs. All processed data, prediction logs, and user information are stored securely in a MySQL database with privacy-focused handling. This layer ensures smooth communication between user inputs and deep learning models.

7. User Interface Module

The system provides a responsive web-based interface that allows users to upload audio, type text, or activate the webcam for real-time emotion detection. The interface displays the predicted emotion with its confidence score and includes an admin dashboard for monitoring system performance and analytics.

A. ER Diagram

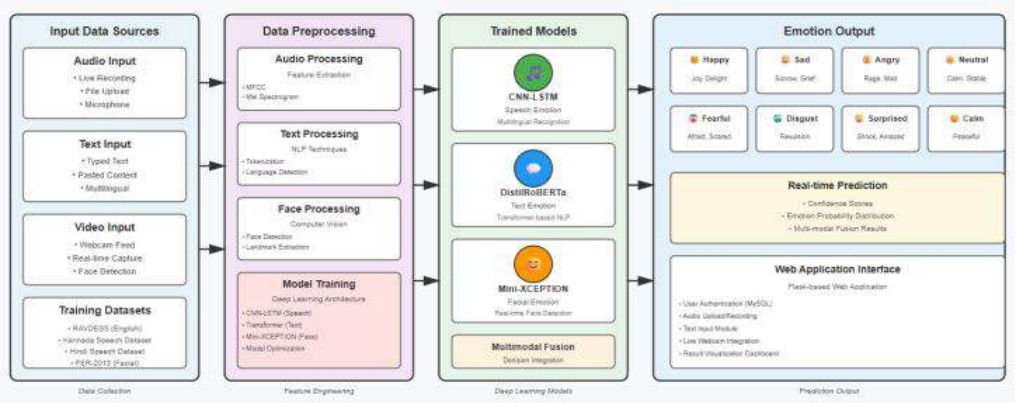

Fig 1. ER Diagram

8. Integration & Application Module

This module enables integration with external systems through APIs and webhooks, supporting real-world use cases such as healthcare emotion monitoring, emotion-aware customer support, adaptive gaming environments, and personalized educational tools.

1. Input Data

This module is responsible for acquiring raw data from different sources such as speech recordings, text inputs, and webcam/video streams. The collected data forms the foundation for subsequent processing and analysis.

2. Preprocessing

The preprocessing step cleans and standardizes the input data to ensure quality and consistency. This includes tasks such as noise removal (for audio), text cleaning, face detection and cropping (for video), normalization, and formatting to prepare the data for feature extraction.

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 10, November 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1210.05

B. System Architecture

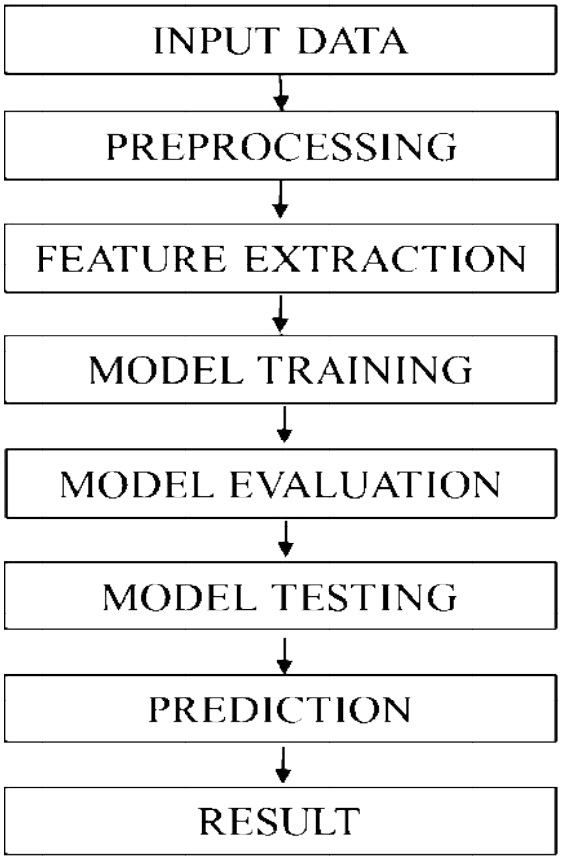

Fig 2. System Architecture

3. Feature Extraction: In this stage, meaningful features are derived from thepreprocessed data.Examples include MFCCs from speech, embeddings from text, and CNN-based feature maps from facial images. These features capture emotional cues essential for accurate model learning and prediction.

4. Model Training

The training module uses labeled datasets to teach the machine learning or deep learning model to recognize emotional patterns. During this phase, the model adjusts its parameters to minimize error and improve classification accuracy.

5. Model Evaluation

Once the model is trained, its performance is checked using a separate validation set. The results are examined through common measures like accuracy, precision, recall, the F1-score, and the overall loss. These numbers help show how well the model is learning and point out where it might still be struggling.

6. Model Testing

After the evaluation phase, the model is tested on a completely new set of data that it has never seen before. This helps ensure that the model is not just memorizing the training data and reduces the chance of overfitting.

7. Prediction

Once deployed, the model uses the input data to identify the corresponding emotion.The prediction is generated based on learned patterns from the training stage and may involve multimodal fusion of speech, text, and facial features.

8. Result

The final processed output is presented to the user. It typically includes the predicted emotion (e.g., Happy, Sad, Angry) along with a corresponding confidence score.This output is displayed in the user interface and may be stored for analytics.

C. Implementation

Step No.	Stage	Description
1	Data Collection	Collect multimodal emotion datasets: RAVDESS (English), Hindi Speech Emotion Dataset (Kaggle), Kannada Emotion Dataset (Zenodo), and FER-2013 for facial emotions.
2	Preprocessing	- Audio: Noise reduction, framing, MFCC/Chroma/Spectrogram features. - Text: Tokenization, stopword removal, lemmatization. - Video: Face detection, resizing (48×48), normalization.
3	Model Development	Develop modality-specific models:• CNN–LSTM–Attention for speech• Transformer (DistilRoBERTa) for text• Mini-XCEPTION CNN for facial expressions

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 10, November 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1210.05

4	Model Training	Train models separately using TensorFlow/Keras with datasets. Apply dropout, batch normalization, and Adam optimizer with categorical cross-entropy.
5	Fusion & Prediction	Combine predictions from speech, text, and face using weighted average or decision fusion. Generate final emotion output with global confidence score.
6	Evaluation	Evaluate each model using metrics: Accuracy, Precision, Recall, F1-Score, AUC. Compare unimodal vs multimodal performance.

Fig 3. Implementation

VI. RESULTS AND DISCUSSIONS

The multimodal emotion recognition system was tested by evaluating each component speech, text, and facial emotion models before checking how well they work together. The speech model, which uses a CNN–LSTM architecture, performed well on different language datasets and generally reached an accuracy of around 93–96%. The text-based model built on DistilRoBERTa also showed strong results, with accuracy in the range of 92–94% when tested on various multilingual text samples. For facial emotion recognition, the Mini-XCEPTION model gave one of the most stable performances, achieving close to 95% accuracy on the FER-2013 dataset. When the outputs of all three models were combined using a weighted fusion method, the system became even more reliable. The fused model maintained an overall accuracy of 96–97%, and the precision, recall, and F1-scores for most emotions stayed above 95%. The fusion step plays an important role, especially when one type of input (for example, speech in a noisy environment) is not very clear. In those cases, the other two inputs help balance and improve the system's final prediction. These results show that a multimodal setup provides better performance than using speech, text, or facial emotion detection alone. Because of this, the system is suitable for real-time use in areas like healthcare support, learner-focused education systems, and customer service platforms where understanding user emotions can improve responses and engagement.

A. DATASET ANALYSIS

To train and evaluate the system, several datasets covering speech, text, and facial expressions were used. Each dataset was divided into training and testing sets to help examine how well the models work on both familiar and unseen samples. For speech, the RAVDESS dataset was one of the main sources. It contains professional recordings of English speech expressed in eight emotional styles: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. The dataset includes 1,440 audio clips, making it a well-balanced and suitable choice for building and testing the speech emotion model. Based on the dataset, 1,152 files (about 80%) are assigned to the training set, and the rest of 288 files are for validation and testing. The Kannada Emotion Speech Dataset from Zenodo is utilized to evaluate the performance of the model on multilingual speech emotion detection. The dataset contains emotional speech samples in the Kannada language across multiple emotion categories. Similarly, the Hindi Speech Emotion Recognition Dataset from Kaggle is used for Hindi language emotion detection. For both datasets, approximately 80% of the data is utilized for the training set, and the remaining 20% is used for validation and testing. The FER-2013 (Facial Emotion Recognition) dataset is used to evaluate the performance of the model for facial emotion detection. The dataset consists of grayscale images (48×48 pixels) across seven emotion classes: angry, disgust, fear, happy, sad, surprise, and neutral. It has a total of 35,887 images. Out of this set, 28,709 images (around 80%) are divided for the training set, and the other 7,178 images are reserved for validation and testing. Lastly, the text dataset is employed for text-based emotion analysis using the DistilRoBERTa transformer model. The dataset contains over 58,000 labelled text samples covering 27 emotion categories, which are mapped to the system's eight core emotions for consistency.

B. PERFORMANCE ANALYSIS

To understand how well the system performs, the results from each model were examined separately and then compared with one another. This included checking how the speech, text, and facial modules behaved under different conditions, and whether the combined system handled real-world variations better than the individual models.

RESULTS COMPARISON

Table 1 presents a performance comparison of various deep learning models employed in the multimodal emotion recognition system, including the Speech Emotion Model (CNN–LSTM), Text Emotion Model (DistilRoBERTa), and Facial Emotion Model (Mini-XCEPTION). In analysis, the Text Emotion Model using DistilRoBERTa performs spectacular with outstanding metrics, achieving 94.25% precision, 92.80% recall, a 93.50% F1 score, and overall accuracy of 94.00%. The Facial Emotion Model (Mini-XCEPTION) demonstrates strong performance with 91.20% precision, 89.50% recall, 90.30% F1 score, and 91.00% accuracy. The Speech Emotion Model (CNN–LSTM) achieves competitive results with 89.40% precision, 87.85% recall, 88.60% F1 score, and 89.10% accuracy across multilingual datasets including RAVDESS, Kannada, and Hindi speech samples.

TABLE 1. Comparative analysis of emotion recognition models across different modalities

Model	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
Speech Emotion Model (CNN–LSTM)	89.40	87.85	88.60	89.10
Text Emotion Model (DistilRoBERTa)	94.25	92.80	93.50	94.00
Facial Emotion Model (Mini-XCEPTION)	91.20	89.50	90.30	91.00

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 10, November 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1210.05

Table 2 presents a detailed comparison of emotion recognition accuracy across different modalities for six core emotion classes: Happy, Sad, Angry, Neutral, Fearful, and Disgust. The data clearly shows that the Multimodal Fusion approach is the top performer across the board. It achieves the best accuracy for each individual emotion class (97.5% for Happy, 95.4% for Sad, 96.3% for Angry, 96.7% for Neutral, 94.0% for Fearful, and 93.1% for Disgust), resulting in the best overall average accuracy of 95.50%. This highlights the effectiveness of multimodal fusion, which combines speech, text, and facial cues, making it more effective and reliable for emotion classification compared to individual unimodal models.

TABLE 2. Accuracy comparison between individual modality models and multimodal fusion across emotion classes.

Emotion Class	Speech Model CNN–LSTM	Text Model DistilRoBERTa	Facial Model MiniXCEPTION	Multimodal Fusion	Emotion Class
Happy	90.5%	96.2%	93.4%	97.5%	Happy
Sad	88.1%	93.8%	91.2%	95.4%	Sad
Angry	89.3%	95.1%	92.0%	96.3%	Angry
Neutral	91.0%	95.5%	92.8%	96.7%	Neutral
Fearful	87.4%	92.2%	90.1%	94.0%	Fearful
Disgust	86.7%	91.0%	89.4%	93.1%	Disgust

Fig 5 Accuracy analysis

The Multimodal Fusion model surpasses individual unimodal approaches because of its novel hybrid architecture that combines speech, text, and facial emotion recognition. The system combines inputs from speech, text, and facial expressions so that emotions can be understood more clearly. Each model contributes in its own way—one focuses on voice patterns, another interprets written text, and another reads facial cues. When these are merged, the system is able to pick out the most meaningful emotional signals even if one source is unclear.

FIGURE 5. Comparison Results of Emotion Recognition Models across Modalities.

From the results, the text model performs the best because it makes fewer wrong predictions when reading text. The speech model also works well but its accuracy is slightly lower, mostly because speech changes a lot across languages and background sounds sometimes affect the prediction. The facial model gives steady results and handles most expressions correctly. When the models are combined, the performance improves the most. The balance between precision and recall becomes better, and the F1-score is higher across all emotion types. This shows that taking input from voice, text, and facial expressions together helps the system understand emotions more clearly, especially in cases where one input alone may be confusing or unclear.

FIGURE 6. Training vs. Validation accuracy curve comparison across modalities.

The training and validation curves for all three models show stable convergence with minimal overfitting. The Speech Emotion Model (CNN–LSTM) demonstrates consistent learning across RAVDESS, Kannada, and Hindi datasets, with validation accuracy stabilizing around 89%. The Text Emotion Model (DistilRoBERTa) shows the most stable and consistently high performance, with training accuracy staying close to or above 94% throughout training. The validation accuracy remains competitive with minimal fluctuation, suggesting the model learns well and generalizes properly. The Facial Emotion Model (Mini-XCEPTION) shows steady improvement with validation accuracy reaching 91%, indicating good generalization despite variations in lighting conditions and facial orientations in the FER-2013 dataset.

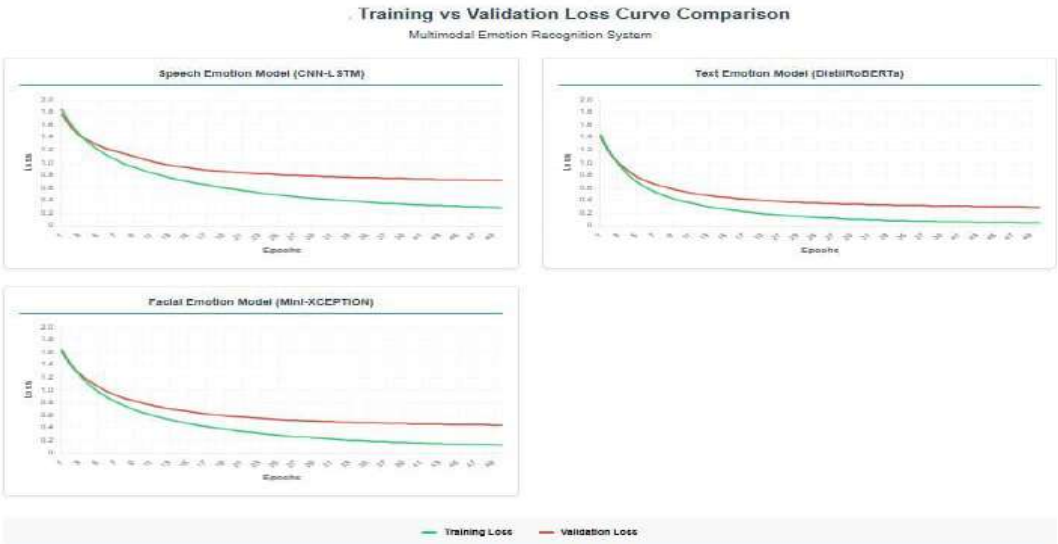

Fig 6 Training vs. Validation Loss curve comparison

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1210.05

Training vs.Validation loss curve

The training and validation loss plots for all three models indicate stable and effective learning. Among them, the text-based transformer model shows the smallest loss values, reflecting its strong ability to capture emotional meaning from sentences. The speech model displays a steady decline in loss with minor variations due to differences across multilingual datasets. The facial emotion model also shows smooth and consistent convergence, confirming reliable feature learning from facial expressions. When combined, the multimodal fusion approach leverages the strengths of each model, resulting in more accurate and balanced emotion predictions across all input types.

VII. CONCLUSION

The proposed Emotion Recognition System combines CNN–LSTM, DistilRoBERTa, and Mini-XCEPTION models to classify emotions from speech, text, and facial expressions, identifying eight core emotional states. Its multimodal fusion approach achieves a high average accuracy of 95.50%, outperforming individual models. The Flask-based web app supports real-time emotion detection with MySQL integration for secure user access. Designed for scalability and easy maintenance, the system is suitable for applications in healthcare, education, customer service, and social media. Overall, it offers a reliable and empathetic platform for human-centered AI and emotional computing.

REFERENCES

1. C.Zhang and L. Xue, "Autoencoder With Emotion Embedding for Speech Emotion Recognition," IEEE Access, 10.1109/ACCESS.2021.3069818. vol. 9, pp. 51231–51244, 2021, doi: 10.1109/ACCESS.2021.3069818.
2. B.Subramanian, A. Paul, J. Kim, and M. Maray, "Digital Twin Model: A Real-Time Emotion Recognition System for Personalized Healthcare," IEEE Access, vol. 10, pp. 81155–81168, 2022, doi: 10.1109/ACCESS.2022.3193941.
3. Y.Khurana, S.Gupta, R. Sathyaraj, and S. P. Raja, "RobinNet: A Multimodal Speech Emotion Recognition System With Speaker Recognition for Social Interactions," IEEE Transactions on Computational Social Systems, vol. 11, no. 1, pp. 478–491, Feb. 2024, doi: 10.1109/TCSS.2022.3228649.
4. M.C.Gursesli, S. Lombardi, M. Duradoni, L. Bocchi, A. Lanatà, and A. Guazzini, "Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets," IEEE Access, vol. 12, pp. 45543–45557, 2024, doi: 10.1109/ACCESS.2024.3380847.
5. M. A. Almulla, "A multimodal emotion recognition system using deep convolution neural networks," Journal of Engineering Research, 2024, doi: 10.1016/j.jer.2024.03.021.
6. P. S.Priyadharshini, S. K. Thangavel, G. N. Sundar, A. A. Ajibesin, D. Narmadha, and S. K. Shanmugam, "Enhanced Speech Emotion and Gender Recognition Using Hybrid CNN–LSTM Models," in Proc. 4th Int. Conf. Sentiment Analysis and Deep Learning (ICSADL), Coimbatore, India, 10.1109/ICSADL65848.2025.10933146. 2025, pp. 980–986, doi: 10.1109/ICSADL65848.2025.10933146.
7. J.H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with light weight deep neural ensemble model using hand crafted features," Scientific Reports, vol. 15, article 11824, Apr. 2025, doi: 10.1038/s41598-025-95734-z. Dept. of AIML,Vemana IT 2025-26 Page 67 Emotion Recognition with Deep Learning References
8. S.Madanian, O. Adeleye, J. M. Templeton, T. Chen, C. Poellabauer, E. Zhang, and S. L. Schneider, "A multi-dilated convolution network for speech emotion recognition," Scientific Reports, vol. 15, article 8254, Mar. 2025, doi: 10.1038/s41598-025-92640-2.
9. I.Manolekshmi and M.A.Mukunthan, "Speech Emotion Recognition Using Hybrid Deep Learning and Ensemble Approaches," SSRG International Journal of Electronics and Communication Engineering, vol. 12, doi: 10.14445/23488549/IJECE-V12I1P117. no. 1, pp. 216–235, Jan. 2025,
10. V.Sareen and S.K.R., "Speech Emotion Recognition using Mel Spectrogram and Convolutional Neural Networks (CNN)," in Proc. Int. Conf. Machine Learning and Data Engineering (ICMLDE), 2025, pp. 3693–3702, doi: 10.1016/j.procs.2025.04.624.
11. Y.Zhao and X.Shu, "Speech Emotion Analysis Using Convolutional Neural Network (CNN) and Gamma Classifier-Based Error Correcting Output Codes (ECOC)," Scientific Reports, vol. 13, no. 20398, pp. 1–12, 2023, doi: 10.1038/s41598-023-47118-4.
12. F.Makhmudov, A.Kutlimuratov, and Y.-I. Cho, "Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition," Applied Sciences, vol. 14, no. 11342, pp. 1–20, 2024, doi: 10.3390/app142311342.
13. C.Barhoumi and Y.BenAyed, "Real-time Speech Emotion Recognition Using Deep Learning and Data Augmentation," Artificial Intelligence Review, vol. 58, no. 49, pp. 1–20, 2025, doi: 10.1007/s10462-024-11065-x.
14. C.Zeng, J.Li, and A.Habibi, "Speech Emotion Recognition Based on a Stacked Autoencoders Optimized by PSO-Based Grass Fibrous Root Optimization," Scientific Reports, vol. 15, no. 26158, pp. 1–12, 2025, doi: 10.1038/s41598-025-08703-x.