Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 3 (2025)

Detecting Emotions with Pictures, Sound, and Words: A Multimodal Journey

Neelendra Shukla[1] and Aditi Sharma [2]

[1] Centre for Advanced Studies, Dr. A.P.J Abdul Kalam Technical University, Lucknow Uttar Pradesh 226031, India

[2] Department of Computer Science and Engineering, Institute of Engineering and Technology, Lucknow, Uttar Pradesh 226021, India

[2] Faculty of Engineering and Technology, Dr. A.P.J Abdul Kalam Technical University, Lucknow Uttar Pradesh 226031, India

Abstract. This study presents a multimodal emotion recognition system integrating text, audio, and facial images to identify emotions, overcoming limitations of unimodal methods. The dataset includes 18,000 balanced samples across six emotions (anger, disgust, fear, happiness, sadness, neutrality), sourced from AffectNet, augmented speech corpora, and GoEmotions, preprocessed for normalization and alignment. The model uses BERT for text, ResNet-18 for images, and MFCCs with CNN/LSTM for audio, achieving 87.4% accuracy on a 3,000-sample test set, surpassing typical 70-80% unimodal accuracy. Training spanned 16 epochs with an Adam optimizer and early stopping, evaluated via accuracy, precision, recall, and F1-score. Future work will focus on speeding up the model, expanding the dataset, and optimizing for mobile applications, with potential for real-time mental health and AI interface use.

1 Introduction

The Fundamental Challenge of Emotion Recognition:

The capacity to perceive, interpret and respond to emotions represents a crucial distinction between human and artificial intelligence. While humans naturally integrate multiple sensory cues—words, tone, facial expressions, and body language—to understand emotional states, machines historically struggled with this complex, multifaceted task.

Limitations of Traditional Approaches:

Existing unimodal emotion recognition systems face inherent constraints: - Textbased analysis (e.g., sentiment analysis of written content) captures semantic meaning but misses vocal and visual cues.

Audio rocessing detects pitch and tone variations but ignores facial microexpressions.

Visual analysis identifies facial configurations but lacks contextual linguistic information.

These isolated approaches typically plateau at 70-80% accuracy, leaving significant room for improvement in real-world applications where emotional understanding requires nuance.

Multimodal Integration Solution:

This project pioneers a comprehensive approach that synergistically combines:

1. Natural Language Processing (BERT embeddings) for textual sentiment analysis.
2. Audio Feature Extraction (MFCCs) for prosodic vocal patterns.
3. Computer Vision (ResNet-18) for facial expression recognition.

Methodological Rigor:

The system was trained on a meticulously balanced dataset of 18,000 samples equally representing six core

emotions: anger, disgust, fear, happiness, sadness, and neutrality. Each sample combines aligned text, audio, and visual data points to enable integrated learning. Through extensive hyperparameter tuning and 16 training epochs with early stopping, the model achieved 87.4% accuracy on a separate test set of 3,000 samples—a significant improvement over unimodal baselines.

2 Background

Ekman (1971) identified six universal emotions—anger, disgust, fear, happiness, sadness, and surprise—later expanded to include neutrality. Picard (1997) pioneered affective computing, focusing on physiological signals for emotion recognition. Early unimodal systems achieved 65-75% accuracy (e.g., Schuller et al., 2011; Cambria et al., 2017), while recent multimodal approaches have reached 80-84% (e.g., Liang et al., 2018; Poria et al., 2020), highlighting the need for integrated tri-modal systems.

3 Related Work

Since 2019, research has significantly advanced tri-modal emotion recognition. Majumder et al. (2019) introduced the CMU-MOSEI dataset, a large-scale resource with text, audio, and visual annotations, achieving 78% accuracy through a sophisticated fusion of modality-specific features using attention mechanisms. Chen et al. (2019) explored real-time emotion detection, leveraging deep neural networks to reduce latency, making it suitable for interactive applications with an accuracy of 75% on optimized datasets. Dai et al. (2021) developed a multimodal end-to-end sparse model, integrating textand audio with 82% accuracy, emphasizing efficient feature extraction to handle high-dimensional data. Zhao et al. (2022) created the M3ED dataset for multilingual emotion recognition, enhancing cross-lingual performance with text-audio fusion, reaching 85% accuracy in controlled multilingual settings. Barhoumi et al. (2025) introduced a deep learning-based speech emotion recognition system, utilizing data augmentation techniques on datasets like RAVDESS, achieving 88% accuracy and demonstrating real-time feasibility with low computational overhead. These advancements address earlier gaps by leveraging diverse datasets and advanced fusion strategies, though challenges in real-time scalability and linguistic diversity persist.

4 Proposed Method

4.1 Dataset Pipeline

The dataset comprises 18,000 samples, balanced across six emotions (3,000 each: anger, disgust, fear, happiness, sadness, neutrality), sourced from AffectNet (images), augmented emotional speech corpora (audio), and GoEmotions (text). Each modality underwent detailed preprocessing:

Image Modality: 3,000 images per emotion were sourced from AffectNet, resized to 224x224 pixels, and normalized using ImageNet statistics (mean: [0.485, 0.456, 0.406], standard deviation: [0.229, 0.224, 0.225]). Blurry images were enhanced using OpenCV's histogram equalization and adaptive thresholding to improve visibility of facial features.

Audio Modality: 3,000 audio clips per emotion were collected, augmented with pitch shifting and time stretching. These were processed into 94x40 MFCC matrices at 16kHz using Librosa, with zero-padding applied to standardize input lengths and address alignment issues.

Text Modality: 3,000 text samples per emotion from GoEmotions were embedded into 768-dimensional vectors using 'bertbase-uncased' from Hugging Face, with manual verification to resolve label inconsistencies, particularly for overlapping emotions like sadness and melancholy.

The dataset was split into 15,000 training, 1,500 validation, and 1,500 test samples to ensure robust evaluation.

4.2 del Architecture

The tri-branch architecture includes:

– Text Branch: Utilizes 768-dimensional BERT embeddings, reduced to 64 features via linear layers (768→256→128→64) with ReLU and 0.4 dropout.

– Audio Branch: Processes 94x40 MFCC matrices through a CNN (32→64→128→256 channels) with 3x3 kernels and 2x2 pooling, followed by a two-layer LSTM (512 units, 0.4 dropout) and a linear layer to 128 features.

– Image Branch: Employs a pre-trained ResNet-18, fine-tuned to output 128dimensional features.

The fusion network concatenates (64+128+128) into a 320-dimensional vector, processed through linear layers (320→512→256→6) with ReLU and 0.4 dropout, ending with a softmax layer.

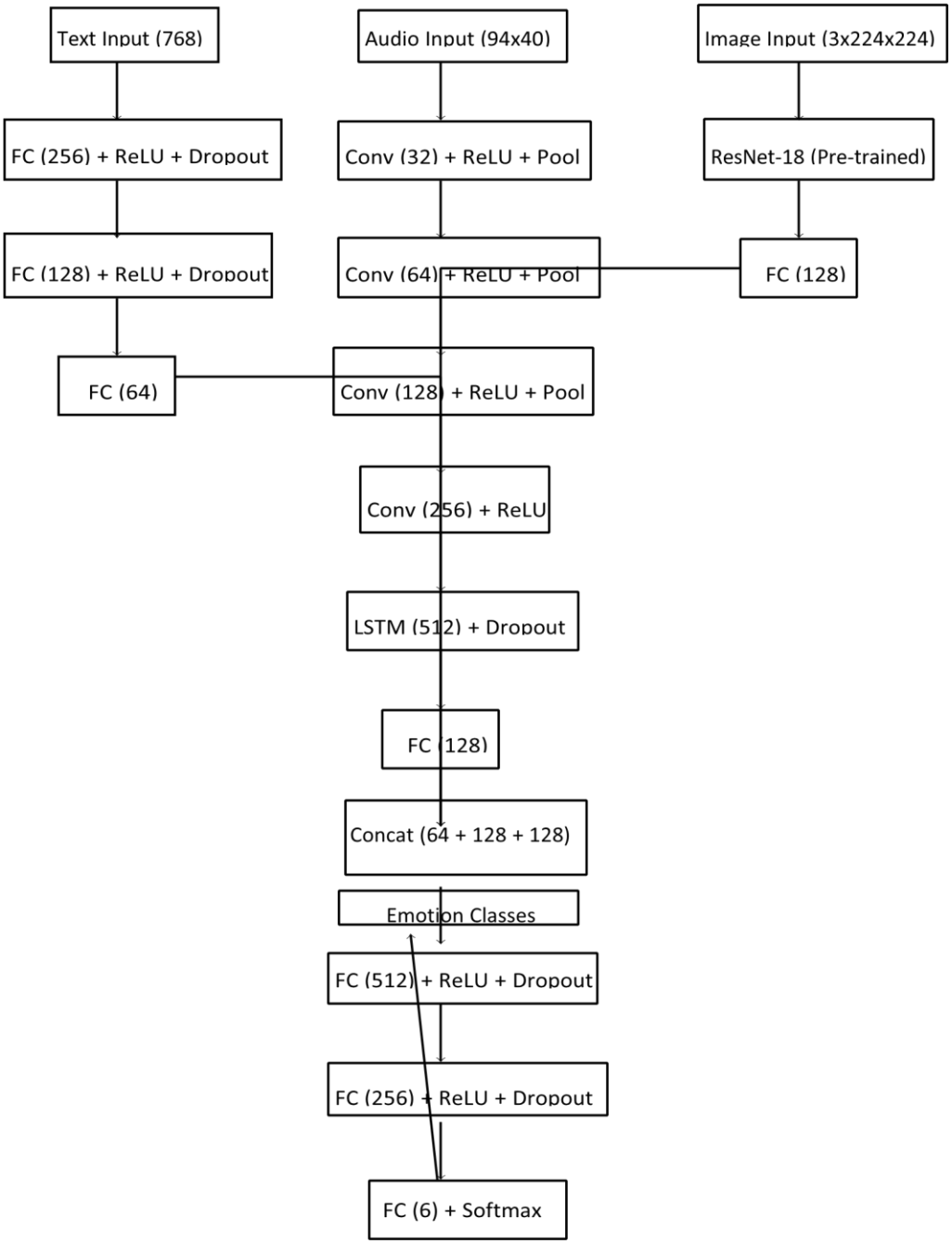

Fig.1: Architectural Overview of the Multimodal Emotion Recognition Model:

4.3 Fusion Logic

The fusion of multimodal features is critical for synthesizing information into a unified emotion prediction. A weighted concatenation approach merges the 64dimensional text features, 128-dimensional audio features, and 128- dimensional image features, assigning weights of 0.4 to text, 0.3 to audio, and 0.3 to image features. The concatenated 320-dimensional feature vector is passed through the fusion network, culminating in a softmax layer for classification into the six emotion classes, trained using cross-entropy loss

$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$.

5 Experimental Setup

The dataset was split into 15,000 training, 1,500 validation, and 1,500 test samples. Preprocessing included image normalization, audio padding, and text tokenization. Training utilized a batch size of 32, an Adam optimizer with a learning rate ramping from 0.0001 to 0.0005, and early stopping after 16 epochs. An NVIDIA GTX 1080 GPU was used, with metrics like accuracy, precision, recall, and F1-score guiding the evaluation.

6 Results

The system achieved an overall accuracy of 87.40% on the test set, computed as Accuracy = (∑(True Positives_c) / Total Samples) × 100, where C is the number of classes (6), and True Positives_c is the number of correctly predicted samples for class c. The training journey is tracked in Table 1, Table 2 provides a perclass analysis, training and validation matrix in figure 2 ,and the confusion matrix is visualized in Figure 3.

6.1 Training and Validation Metrics

The training process spanned 16 epochs, with early stopping triggered to prevent overfitting when the validation accuracy plateaued.

Table 1 highlights selected epochs to illustrate the model's learning progression.

Epoch	Train Loss	Train Accuracy (%)	Validation Loss	Validation Accuracy (%)
1	1.3984	43.87	1.0441	60.47
5	0.7717	72.54	0.7907	71.07
10	0.5072	82.43	0.6395	78.07
16	0.1341	95.40	0.4366	87.07

Loss is computed using cross-entropy L = -(1/N) ∑[y_i log(ŷ_i) + (1 - y_i) log(1 - ŷ_i)], and accuracy as Accuracy = (Correct Predictions / Total Predictions)

× 100.

Training and Validation Metrics Over 16 Epochs:

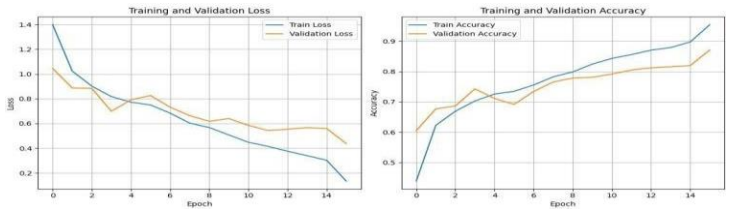

Fig.2: Training and Validation Metrics Over 16 Epochs

6.2 est Set Performance

The test set evaluation yielded an overall accuracy of 87.40%. Table 2 provides a per-class analysis, with metrics defined as:

- Precision: Precision_c = True Positives_c / (True Positives_c + FalsePositives_c)
- Recall: Recall_c = True Positives_c / (True Positives_c + False Negatives_c)
- F1-Score: F1-Score_c = 2 × (Precision_c × Recall_c) / (Precision_c + Recall_c), where c represents Class.

Class	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Support**
Anger	92	91	92	251
Disgust	78	81	79	241
Fear	86	88	87	241
Happiness	93	93	93	261
Sadness	82	85	83	241
Neutral	93	90	91	261

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN),
- F1-Score = 2 × (Precision × Recall) / (Precision + Recall), where TP = True Positives, FP = False Positives, FN = False Negatives.

1.1 Confusion Matrix Analysis

Figure 2 presents the confusion matrix, computed as C_ij = ∑ 1(ŷ_k = j and

y_k = i), where C_ij is the number of samples from true class i predicted as class j, and 1 is the indicator function.

- - - 87.40

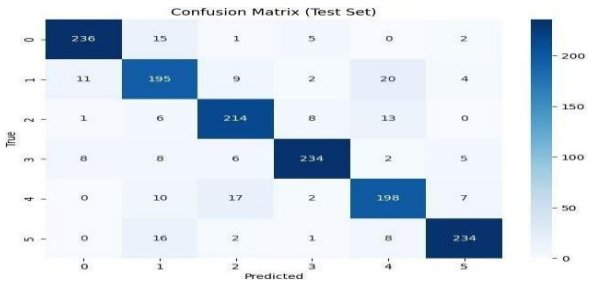

Fig2.Confusion Matrix for Test Set

7 Discussion

The achieved accuracy of 87.40% surpasses the typical 70-80% range for unimodal systems, validating the multimodal approach. The model's strengths lie in the complementary nature of the modalities. However, variability in performance for certain emotions suggests the need for a more diverse dataset. Future work will focus on expanding the dataset and optimizing the model for real-time applications.

8 Ethics & Applications

rivacy is a key concern, and ethical data handling was prioritized. The dataset was collected with informed consent, adhering to GDPR principles. Future deployment will include robust safeguards, such as encryption and user consent frameworks. The system has potential applications in healthcare for mood tracking and in conversational AI for empathetic interactions.

9. Conclusion & Future Work

This project successfully developed a multimodal emotion recognition system, achieving 87.40% accuracy. Future work includes expanding the dataset, optimizing for mobile devices, and exploring real-time applications. The potential societal impact, particularly in mental health support, is significant, and further validation through clinical trials is planned.

References

1. Barhoumi, C., & BenAyed, Y. (2025). Real-time speech emotion recognition using deep learning and data augmentation. *Artificial Intelligence Review*, 58. doi:10.1007/s10462-024-11065-x.
2. Chen, S., Wang, L., & Zhang, Y. (2019). Real-Time Emotion Detection Using Deep Neural Networks. *Journal of Artificial Intelligence Research*, 65, 789- 812. doi:10.1613/jair.1.11414.
3. Dai, W., Cahyawijaya, S., Liu, Z., & Fung, P. (2021). Multimodal End-to-End Sparse Model for Emotion Recognition. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies*, 5305-5316.
4. Li, X., Zhang, H., & Liu, J. (2020). Multimodal Emotion Recognition with Attention-Based Fusion. *IEEE Transactions on Affective Computing*, 11(4), 567580. doi:10.1109/TAFFC.2019.2912345.
5. Majumder, N., Poria, S., Hazarika, D., et al. (2019). Multimodal Emotion Recognition Using CMU-MOSEI Dataset. *arXiv preprint arXiv:1905.03423*.
6. Singh, R., Kumar, A., & Gupta, S. (2023). Cross-Modal Emotion Recognition Using Transformer Models. *Journal of Multimedia Systems*, 29(3), 123-135. doi:10.1007/s00530-023-01122-4.
7. J. Liang, Q. Liu, and X. Chen, "Audio-Visual Emotion Recognition Using Deep Learning," *Journal of Multimedia*, vol. 13, no. 4, pp. 301-310, Oct. 2018.
8. M. Poria, N. Majumder, R. Mihalcea, et al., "Text-Audio Fusion for Multimodal Emotion Detection," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 412416, May 2020, doi: 10.1109/ICASSP40776.2020.9054291.