Cureus Journal of Computer Science

Open Access Original Article

Cross-Model Implementation of Emotion Recognition Systems on Indian Datasets

Vinay Sharma [1] , Naveen Aggarwal [1] , Nirmal Kaur [1]

1. Department of Computer Science and Engineering, University Institute of Engineering and Technology, Panjab University, Chandigarh, IND

Corresponding author: Vinay Sharma, vinaysharmapuchd@gmail.com

Received 06/30/2025
Review began 07/29/2025
Review ended 09/03/2025
Published 09/10/2025

DOI:
https://doi.org/10.7759/s44389-025-08677-x

Abstract

The proliferation of digital communication has increased the significance of precise emotion recognition across various applications, including human-computer interaction, mental health monitoring, emotional chatbots, and security systems. However, only a handful of emotion recognition models have been implemented on Indian emotion datasets. However, the majority of existing models have not been tested on Indian datasets, which pose distinct challenges due to cultural differences and variations in emotional expressions. This paper implements deep learning-based emotion detection models - ResNet-50 and Mel-Frequency Cepstral Coefficients-Convolutional Neural Network (MFCC-CNN) - and trains them on popular Indian datasets: InFER, IIITM, ISED, MENDELY, and PUMAVE-D. These datasets provide a wide range of expressions and cultural nuances, posing significant challenges for model performance. Experimental results reveal that the ResNet-50 model attains an accuracy of 99.89% - ISED dataset (video modality), while the MFCC-CNN model achieves 93% accuracy on the PUMAVE-D dataset (audio modality). The multimodal approach achieves 93% accuracy on PUMAVE-D (audio-video), although performance varies across datasets, with MENDELY (87.05%) and PUMAVE-D (video) (84.41%) showing lower accuracy due to increased variations in expressions and data characteristics. This study assesses the performance of ResNet-50 and MFCC-CNN models on image, audio, video, and multimodal datasets, highlighting their adaptability across different datasets, providing insights specific to each dataset, and employing score-based fusion to improve overall performance.

Categories: Affective Computing, Computational Science and Engineering, Deep Learning
Keywords: emotion recognition, indian datasets, multimodal analysis, machine learning models, deep learning models

Introduction

Emotions are feelings through which we judge the world around us and other people, based on our beliefs. This helps us to respond and adapt to different situations [1]. Applications of emotion recognition extend to various domains, including marketing and security, mental health assessment, and human-computer interaction. Machines can interpret human emotions through facial expressions, speech, or physiological signals. Emotion recognition systems enhance user experiences and facilitate more intuitive interactions. For instance, in healthcare, these systems can assist in monitoring patient moods, while in marketing, they can help gauge consumer reactions to products. With advancements in technology, there is an increasing expectation for interactive machines to accurately recognise and exhibit emotions like human emotional expression [2]. Despite global advancements, the performance of emotion recognition models can significantly vary across various demographic and cultural groups due to inherent differences in expressions and gestures. It emphasises the need to develop systems that can accurately react and interpret human emotions, allowing for smooth and effective interaction between people and computers [3,4].

Emotion recognition is classified as unimodal and multimodal. Unimodal systems focus on a single modality, such as facial expressions, speech, to separately recognise emotions [5-7]. However, a single modality might not accurately describe emotions for certain complex emotional states [8]. A multimodal emotion recognition system integrates the data from various modalities, such as speech patterns and facial expressions, for more accurate results [9,10].

India, with its rich diversity of languages, ethnicities, and cultural norms, presents a unique and complex landscape for emotion recognition. Indian datasets offer diverse facial structures, expressions, and emotional cues, providing a robust test bed for developing and refining emotion recognition technologies. Motivation to focus on Indian datasets in this study is twofold: i) these datasets capture the intricate nuances of emotional expressions specific to Indian culture, which are often underrepresented in global emotion recognition research, ii) growing demand for localised AI solutions in India, driven by its expanding technology sector.

The primary contributions of this paper are outlined as follows:

How to cite this article
Sharma V, Aggarwal N, Kaur N (September 10, 2025) Cross-Model Implementation of Emotion Recognition Systems on Indian Datasets . Cureus J Comput Sci 2 : es44389-025-08677-x. DOI https://doi.org/10.7759/s44389-025-08677-x

• Explored Indian emotion datasets containing unimodal (Audio-A, Video-V) and multimodal (A-V) clips.

• Implemented different deep learning models (ResNet-50 and Mel-Frequency Cepstral Coefficients-Convolutional Neural Network [MFCC-CNN]) on unimodal (IIITM, InFer, IIMI, Mendely, IHED, ISAFE, ISED) and multimodal datasets (PUMAVE-D, IndEmoVis).

• Analysed the performance of these models on Indian emotion datasets.

A novel score-level fusion strategy is introduced, where the most dominant modality is selected to achieve more reliable multimodal emotion recognition.

Materials And Methods

Related work

Existing datasets are classified on the basis of the data modality they provide, i.e., either Audio-A, Video-V, or A-V. Certain datasets contain single-modality data, images, or audio, while others offer multimodal data combining both audio and visual modalities. We have categorized the Indian datasets based on their modality-unimodal and multimodal. The unimodal category comprises audio and video clips separately, and the multimodal category includes both audio and video clips together.

Figure *1* illustrates the taxonomy of these Indian datasets, while Table *1* provides an overview of the available datasets based on emotion categories, modality, sample count, and the inclusion of Indian emotions.

A critical observation is that most existing research on emotion recognition relies mostly on global datasets, with limited research specifically focusing on Indian datasets. Studying Indian datasets is essential due to the cultural and contextual differences in emotional expression, which can significantly impact the accuracy and generalizability of the models. Cultural nuances, language diversity, and distinct social norms in India influence how emotions are expressed and perceived. Therefore, research centred on Indian datasets is crucial to developing more efficient and contextually relevant emotion detection systems tailored to the Indian subcontinent.

Unimodal Emotion Datasets

Rizvi et al. [11] published an image dataset, InFER (Indian Facial Emotion Recognition). It offers a vast amount of labelled facial images captured under controlled conditions. It consists of 4,200 video clips and 10,200 images, which are of nearly 600 human subjects and feature 6,000 spontaneous expression images sourced from the internet through crowdsourcing. A wide range of variations in terms of age, gender, ethnicity, and expressions makes this dataset unique. The expressions are divided into seven basic emotions: neutral, fear, anger, disgust, happiness, sadness, and surprise. The IIITM face dataset is an essential resource for evaluating the efficiency of facial attribute detection models in the context of the Indian subcontinent, introduced by Arya et al. [12]. There are 107 human subjects, having 1,928 images with varied facial characteristics such as pose, gender, emotion, facial hair, glasses, hair, and clothing. The IIMI [13] emotional face dataset contains an emotional face database of Indian subjects from the northern regions of India. Another dataset of Indian face images with varied expressions (Mendely data) was introduced by Dachawar [14].

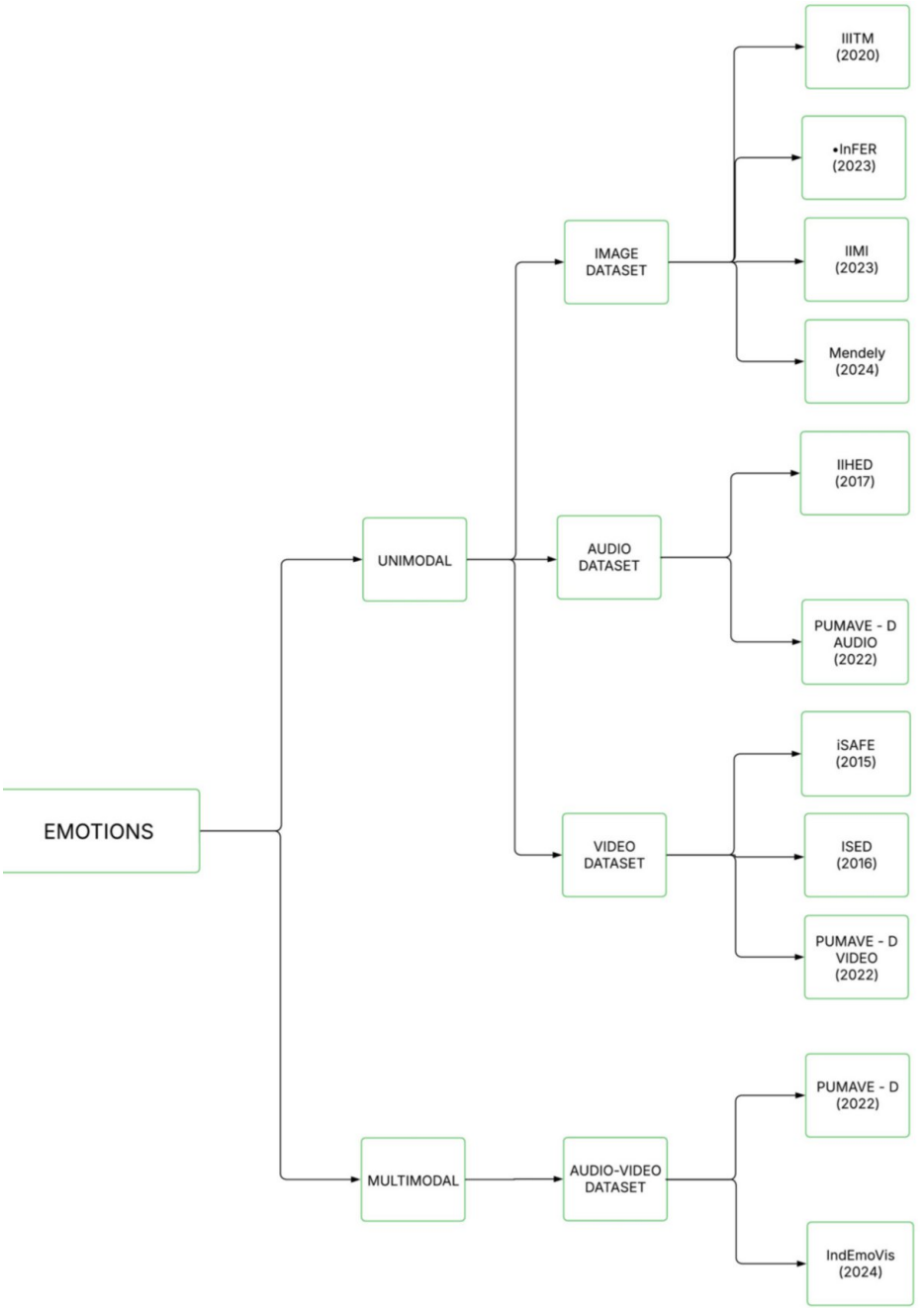

FIGURE 1: Taxonomy of Indian unimodal and multimodal emotion recognition datasets

Modality	Dataset	Unimodal (U)/Multimodal (M)	Number of samples	Emotions
Images	IIITM Face Dataset [12] (2020)	U	1,928	Neutral, Sad, Smiling, Surprised, Surprised with Open Mouth (SO), Bored
	Indian Facial Emotion Recognition [11] (2023) (InFER)	U	10,200 images, 4,200 video clips	Neutral, Fear, Anger, Disgust, Happiness, Sadness, Surprise
	IIMI [13] (2023)	U	1,302	Anger, Disgust, Fear, Happy, Surprise, Neutral, Sad
	Mendely [14] (2024)	U	6,000+	Happiness, Thoughtfulness, Anger, Surprise, Sadness, Fear, Excited
Audio	IIIT-H TEMD [15] (2017)	U	5,317	Anger, Happiness, Sadness, Neutral, Surprise, Fear, Disgust, Sarcastic, Frustrated, Relaxed, Worried, Shy, Excited, Shout
	PUMAVE-D AUDIO (2022)	U	1,188	Anger, Disgust, Fear, Happy, Neutral, Sad
Video	ISED [16] (2016)	U	428	Happiness, Surprise, Sadness, Disgust
	iSAFE [17] (2015)	U	395	Happiness, Sadness, Surprise, Disgust, Fear, Anger, Uncertainty, no emotion
	PUMAVE-D VIDEO (2022)	U	1,188	Anger, Disgust, Fear, Happy, Neutral, Sad
Audio-Video	IndEMOVis [18] (2024)	M	122	Anger, Disgust, Fear, Happy, Neutral, Sad
	PUMAVE –D [19] (2022)	M	1,188	Happiness, Surprise, Neutral, Anger, Sadness, Disgust, Awe, Fear, Sympathy

TABLE 1: Overview of Indian emotion recognition datasets

The IIIT-H TEMD dataset consists of emotional speech, introduced by Gangamohan et al. [15]. It consists of 5,317 annotated utterances, having 2,341 females and 2,976 male speakers. These utterances are annotated with different emotion categories and confidence levels. The emotions are categorized into 10 basic emotions: anger, sad, neutral, happiness, surprise, fear, disgust, sarcasm, frustration, and worry.

Happy et al. [16] developed the Indian Spontaneous Expression Database (ISED) dataset. The dataset contains 428 differentiated video clippings, which are facial expressions of 50 participants. The dataset consists of 213 images depicting 7 different facial expressions - sadness, happiness, surprise, anger, disgust, fear, and neutral - captured from 10 subjects. Additionally, it contains 486 expression sequences from 97 participants, with each sequence illustrating the progression from a neutral state to a peak emotional expression. The iSAFE (Indian Semi-Acted Facial Expression) Dataset for Human Emotions Recognition was introduced by Singh and Benedict [17].

Multimodal Emotion Datasets

Karani and Desai [18] developed a novel multimodal dataset known as the IndEMoVis. It comprises 122 recorded audio-video clips, which were captured in conversations between the individuals. It comprises 61 participants, out of which 25 were females and 36 were males, aged between 18 and 21. These participants were primarily from Gujarat and Maharashtra states in India. A total of nine emotions are there: happiness, surprise, neutral, anger, sadness, disgust, awe, fear, and sympathy. PUMAVE-D was introduced by Singh et al. [19] as a multimodal audio-video dataset developed for emotion recognition within the Indian context. It consists of synchronized audio-visual recordings of subjects expressing seven primary emotions: happiness, sadness, anger, surprise, fear, disgust, and neutral. The dataset ensures high-quality recordings with proper lighting and minimal background noise, making it suitable for both speech emotion recognition and facial expression recognition tasks.

Models and methods

Models play a crucial role by enabling the extraction of complex emotional data from different modalities. Conventional approaches depend on manually designed features and machine learning models for automatic emotion detection. Recent trends have shifted towards sophisticated deep learning models. For image-based emotion recognition, models ResNet-18 and ResNet-34 are used, offering varying depths and complexities. ResNet-50 is preferred due to its 50-layer deep architecture, which maintains a balance

between depth and computational efficiency. In audio-based emotion recognition, MFCC features combined with CNN architectures have shown promising results. In this paper, ReNet-50 [20] and MFCC-CNN models have been implemented for both modalities [21]. ResNet-50, a deep neural network architecture, overcomes the vanishing gradient challenge by integrating residual (skip) connections. It consists of 50 layers, organised into five stages, with each stage containing convolutional and identity blocks. The architecture begins with an input layer for 64 × 64 RGB images, followed by zero-padding, a 7 × 7 convolution, and max-pooling to reduce spatial dimensions. Each stage combines convolutional blocks for down-sampling and identity blocks for deeper feature extraction, with the number of filters increasing in subsequent stages (from 64 to 2048). To finalize the classification process, the model uses global average pooling, which condenses the spatial dimensions to 1 × 1. This is followed by a fully connected layer that employs SoftMax activation to determine the output class. The overall architecture of ResNet-50 used in this study is illustrated in Figure *2*.

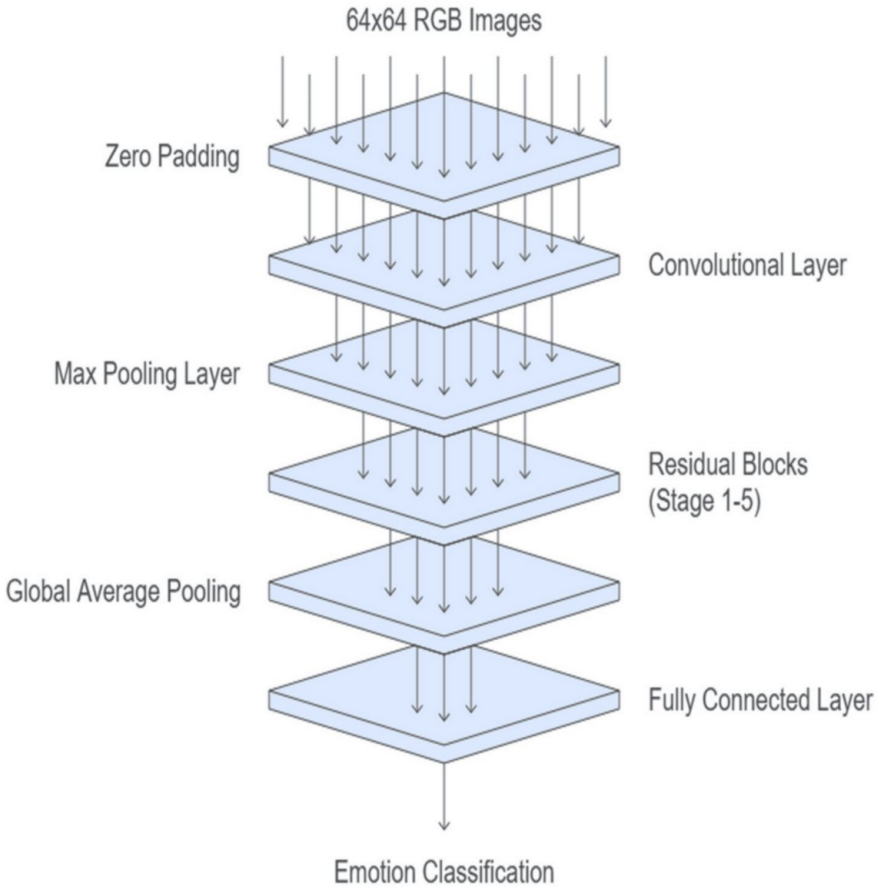

FIGURE 2: Architecture of the ResNet-50 model for image- and video-based emotion recognition

MFCC-CNN is another deep learning model used for recognizing emotions from speech. It is used to analyse and classify emotions from audio modality. MFCCs are widely used features in speech processing because they capture important frequency details that reflect how humans hear sounds. These features are then passed to a CNN that learns audio patterns to capture different emotions. For the video modality, frames were extracted from each video clip and resized to a resolution of 64 × 64 pixels in RGB format. These frames were then normalized before being passed into the ResNet-50 model. In the case of the audio modality, the original audio files were first converted to mono WAV format with a sampling rate of 16 kHz. MFCCs were computed using a 25 ms window with a 10 ms overlap, extracting 13 coefficients per frame. To maintain consistent input dimensions for the CNN, all MFCC sequences were either padded or trimmed to a fixed length. The MFCC-CNN model is built using three 2D convolutional layers, each activated by the ReLU function and followed by 2D max-pooling. The resulting feature maps are then flattened and passed through two fully connected layers with 128 and 64 neurons, respectively. A SoftMax layer is used at the end for classification. To reduce the risk of overfitting, a dropout rate of 0.5 is applied after each dense layer. Figure *3* illustrates the complete architecture of the MFCC-CNN model.

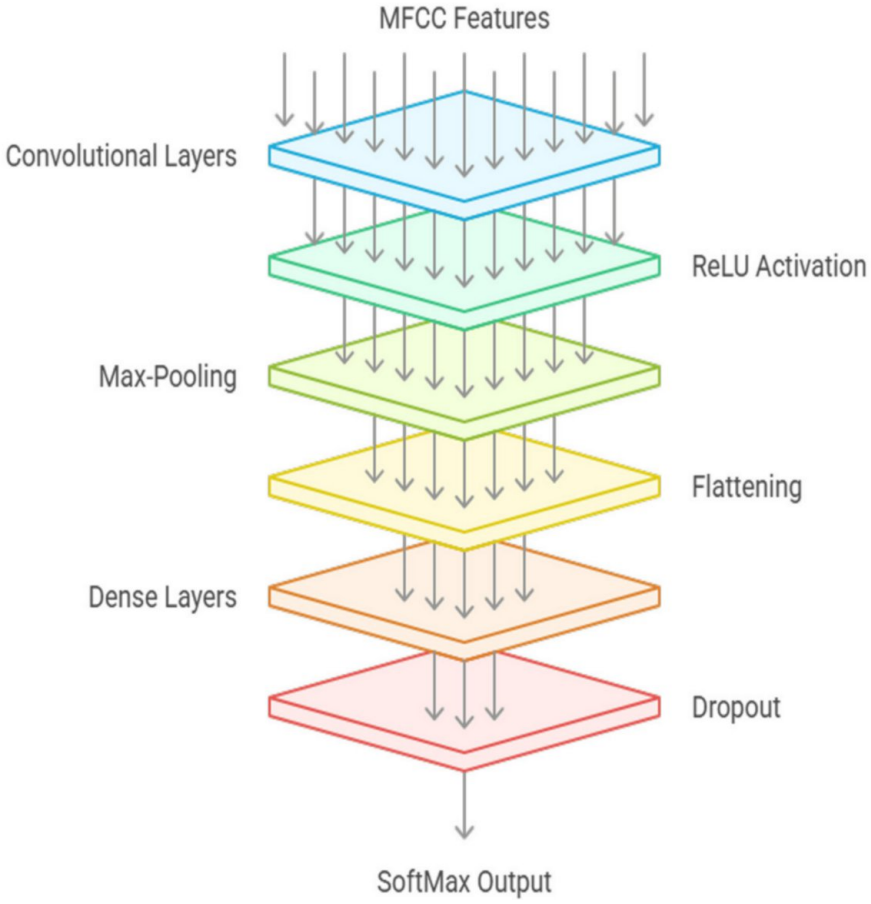

FIGURE 3: Architecture of the MFCC-CNN model for audio-based emotion recognition

MFCC-CNN, Mel-Frequency Cepstral Coefficients-Convolutional Neural Network; ReLU, Rectified Linear Unit

MFCC-CNN inputs extracted MFCC features of the audio data and then applied multiple convolutional layers to detect patterns, followed by pooling layers. Afterwards, the flattened data is passed through fully connected layers that classify emotions. The final output layer employs the SoftMax activation function to categorize six different emotional states. Score-based modality selection method is used for multimodal emotion recognition. Instead of directly fusing audio and video features, the models are evaluated separately to determine the stronger modality. ResNet-50 is trained on video data (PUMAVE-D-Video) and MFCC-CNN on audio data (PUMAVE-D-Audio), with accuracy tracked at each training epoch. Final classification is based on the modality with the greatest accuracy, ensuring that the dominant model contributes to the final decision. This approach prevents performance degradation that can occur when combining weaker modalities and allows for a stronger and more effective emotion recognition system.

Evaluated three fusion strategies: early fusion (by concatenating features), late fusion (by averaging prediction scores), and score-based selection. Among these, score-based selection achieved the highest accuracy of 93% on the PUMAVE-D dataset. In comparison, early fusion and late fusion resulted in accuracies of 89.2% and 90.4%, respectively. Due to its better performance and straightforward implementation, score-based fusion was chosen for the final model.

Both models were trained using the categorical cross-entropy loss function, with optimization carried out using the Adam optimizer. The training process was conducted over 50 epochs, incorporating early stopping based on validation loss to prevent overfitting. An 80/20 stratified train-test split was employed to maintain a balanced distribution of emotion classes across both sets. Model performance was assessed using standard evaluation metrics, including accuracy, precision, recall, and F1-score. To ensure reliability and minimize variability in the results, all experiments were repeated three times, and the final metrics were reported as the average across these runs.

Results And Discussion

Experimental results

The performance of deep learning-based models (ResNet-50, MFCC-CNN) is analyzed on several Indian emotion datasets (InFER, IIITM Face Dataset, IIMI, Mendeley, IIIT-H TEMD, PUMAVE-D AUDIO, ISED, iSAFE, PUMAVE-D-VIDEO, IndEMoVis, PUMAVE-D-AUDIO-VIDEO). While cross-validation is commonly used, it was not adopted in this study to maintain consistency across all datasets and to minimize training time, considering the large number of dataset-model combinations involved. Instead, an 80-20 stratified split was used to ensure that emotion classes were evenly represented in both the training and test sets. Each dataset is split into training and testing (80-20). To maintain uniformity, we selected the same parameters, such as 50 epochs, a dropout rate of 0.5, and the Adam optimizer configured with a learning rate of 0.001. The models are trained and tested with GPU support with Intel Xeon W7-2475X (37.5 MB cache, 20 cores, 40 threads, 2.6 GHz to 4.8 GHz) and a Graphics Card of NVIDIA RTX A5500 (CUDA enabled).

Evaluation metrics

Proposed models for emotion recognition have been assessed using evaluation metrics such as accuracy, F1-score, precision, and recall, which are defined in Table *3*, and Table *2* summarizes these evaluation metrics.

Metrics	Equation
Accuracy	(TP + TN)/(TP + TN + FP + FN)
Precision (P)	(TP)/(TP + FP)
Recall (R)	TP/(TP + FN)
F1-score	2 (P × R)/(P + R)

TABLE 2: Performance metrics

TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative

Performance analysis

Performance of models (ResNet-50 and MFCC-CNN) on the existing Indian datasets is shown in Table *3*. The pre-trained ResNet-50 model is known for its depth and the residual connections, which help in capturing emotional expressions from images/video. For audio datasets, the MFCC-CNN model is used to process sequential data.

Model	Dataset	Modality	Accuracy	F1-score	Precision	Recall
ResNet-50	IIITM (2020)	Image	99.07%	0.990	0.991	0.990
	MENDELY (2024)	Image	87.05%	0.866	0.867	0.864
	ISED (2016)	Video	99.89%	0.96	0.96	0.96
	PUMAVE-D–VIDEO (2022)	Video	84.41%	0.78	0.78	0.78
MFCC-CNN	PUMAVE-D–AUDIO (2022)	Audio	93.00%	0.95	0.95	0.95
ResNet-50	PUMAVE-D–AUDIO-VIDEO (2022)	Audio-video	93.00%	0.95	0.95	0.95

TABLE 3: Performance of ResNet-50 and MFCC-CNN models on emotion datasets

MFCC-CNN, Mel-Frequency Cepstral Coefficients-Convolutional Neural Network

As analysed from Table *3*, it is found that the ResNet-50 model provides an accuracy of 99.07% on the image dataset - IIIT and 99.89% on the video dataset - ISED.

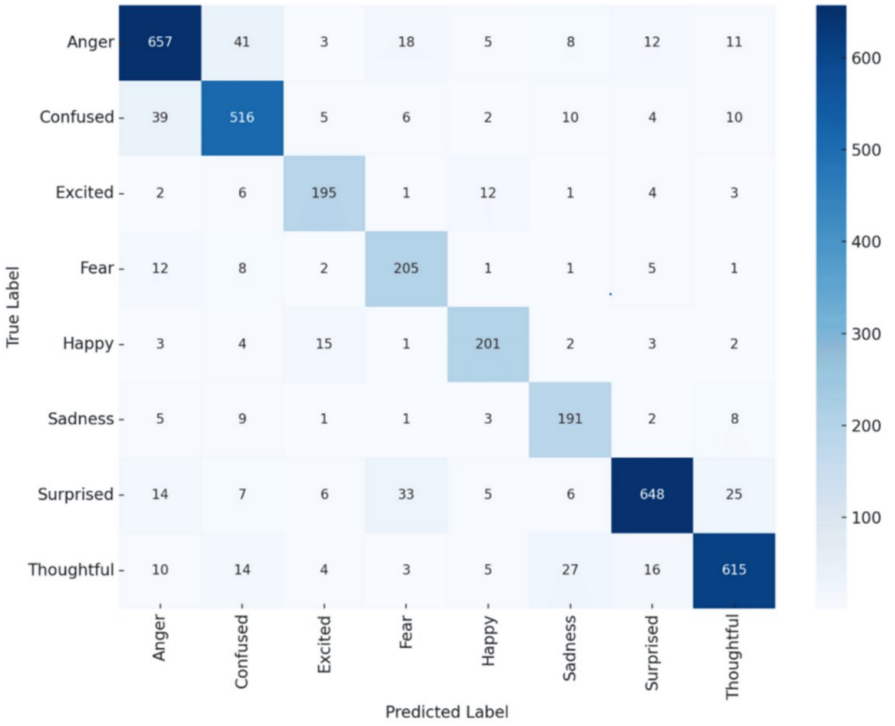

FIGURE 4: Confusion matrix of ResNet-50 on Mendeley dataset

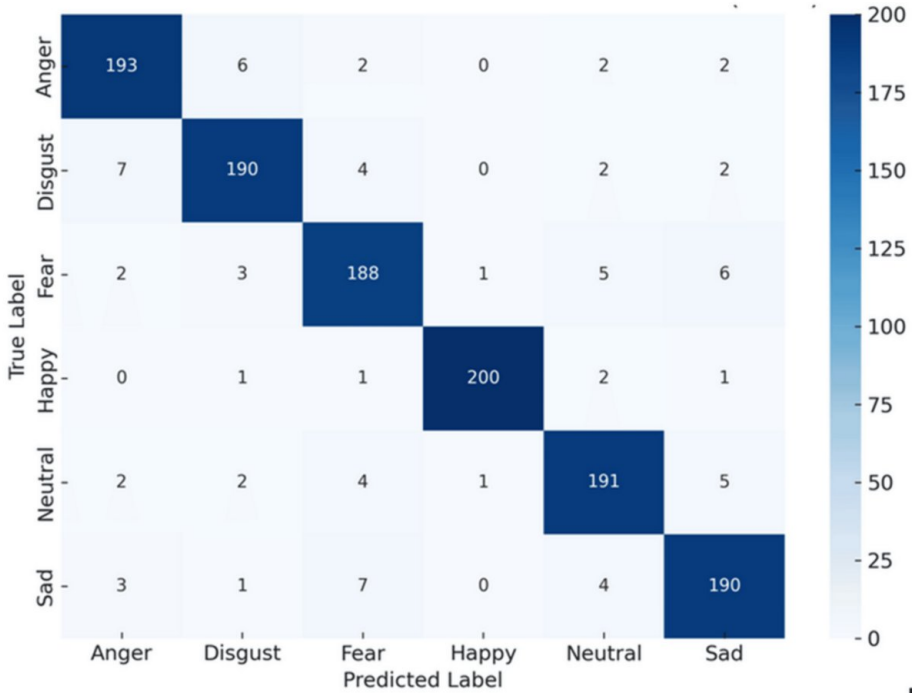

FIGURE 5: Confusion matrix of MFCC-CNN on PUMAVE-D Audio dataset

MFCC-CNN, Mel-Frequency Cepstral Coefficients-Convolutional Neural Network

In addition to overall performance metrics, confusion matrices (Figures *4* and 5) and per-class analyses offer deeper insights into model behavior, especially on challenging datasets such as Mendeley and PUMAVE-D. These analyses reveal class-specific strengths and weaknesses: emotions like happiness and neutral are recognized with higher reliability, whereas emotions such as fear and disgust are frequently

misclassified, largely due to limited training data and greater variability in expression. These results show that the model successfully detects emotional expressions in high-quality, structured images, where facial features are well-defined and consistently represented. However, performance dropped to 87.05% and 84.41% for diverse datasets such as MENDELY and PUMAVE-D - VIDEO dataset. The decline results can be attributed to variations in lighting conditions, occlusions, background clutter, and motion artifacts present in video sequences. The differences between static images and dynamic video frames also play a crucial role, as temporal variations in facial expressions can introduce inconsistencies in feature extraction. These findings highlight the importance of robust feature extraction techniques and domain adaptation methods to enhance the generalizability of image- and video-based emotion recognition models.

The audio model MFCC-CNN achieves an accuracy of 93% and an F1-score of 0.95 on PUMAVE-D-AUDIO. The use of MFCC effectively captures essential acoustic features that distinguish different emotional states. Audio signals convey critical emotional cues through variations in pitch, tone, intensity, and speech modulation, making MFCC an essential component in speech-based emotion classification. While the model performed well, challenges such as background noise, speaker variability, and differences in pronunciation could affect classification accuracy.

The integration of audio and video modalities has been evaluated using ResNet-50 on PUMAVE-D - AUDIO-VIDEO (2022), where the model achieved 93% accuracy and an F1-score of 0.95. Multimodal models leverage complementary information from both visual and auditory inputs, enhancing the robustness of emotion recognition systems. The results indicate that combining both modalities leads to improved classification performance compared to unimodal systems. However, the similarity in accuracy between the audio-only and multimodal models suggests that the fusion strategy used may not be fully optimized to exploit the benefits of multimodal learning. Advanced fusion techniques, including attention mechanisms and late fusion strategies, can be explored to enhance multimodal emotion recognition.

In a nutshell, deep learning models excel on well-structured datasets but face challenges on diverse data sources. Image and video-based models perform well with clear visual data but require robust feature extraction techniques for unstructured content. MFCC-CNN provides good results for audio-based recognition, while multimodal fusion holds promise but demands optimised integration strategies for significant performance gains.

One of the major challenges faced was an imbalance in class. By this imbalance, the precision and recall were affected, which ultimately led to lower accuracy. These results show the importance of using data augmentation effectively for enhancing the model's resilience and balancing the datasets. These findings suggest that the performance of the models can be significantly enhanced on diverse datasets designed for Indian datasets. Another limitation arises from class imbalance and variability introduced by background noise and uncontrolled recording conditions. These challenges may be mitigated through techniques such as data augmentation, resampling, and noise-robust training, which we propose to explore in future work. For enhanced feature extraction, transformer models could be explored in future research, which can help in improving the accuracy and efficiency. Making more diverse and larger datasets, which have more emotional variations and improved labeling, will create more comprehensive models. Making more advanced multimodal architectures that effectively combine audio and video could enhance the applicability and accuracy of emotion recognition.

Conclusions

Performance of emotion recognition models was evaluated across various Indian datasets, leveraging pre-trained ResNet-50 models for image and video data, and MFCC-CNN models for audio data. The key findings show that ResNet-50 achieved high accuracy on the IIITM and ISED datasets, whereas MFCC-CNN performed best on the PUMAVE-D audio dataset. Additionally, combining modalities through score-based fusion further enhanced overall performance, highlighting the benefits of selecting models based on data type. The results revealed that ResNet-50 performed well on image-based datasets (IIITM, ISED), achieving accuracy (99.07% and 99.89%, respectively). However, some datasets presented challenges due to class imbalance and noise, which impacted the models' precision and recall. MFCC-CNN models demonstrated strong performance on audio-based emotion recognition, with 93% accuracy on the audio dataset, while multimodal fusion of audio and video data improved overall performance, highlighting the benefits of both modalities. Our results show that datasets with higher expressive and structural variability, such as MENDELY and PUMAVE-D, tend to result in lower model performance. This highlights the importance of developing robust, culturally aware emotion recognition systems and provides a foundation for future research in multimodal and cross-cultural approaches. Challenges encountered include class imbalance, background noise, and modality-specific limitations, which affected precision and recall for certain datasets. Despite these successes, issues like class imbalance and real-world data variability, inconsistent lighting, and background noise affected the model's robustness. along with the implementation of score-based multimodal fusion, providing dataset-specific insights and practical guidance for emotion recognition tasks in the Indian context. Future research could focus on exploring more advanced fusion methods, creating larger and more diverse Indian datasets, and integrating

transformer-based architectures to improve the accuracy and robustness of multimodal emotion recognition.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Vinay Sharma, Naveen Aggarwal, Nirmal Kaur

Acquisition, analysis, or interpretation of data: Vinay Sharma, Naveen Aggarwal, Nirmal Kaur

Drafting of the manuscript: Vinay Sharma, Naveen Aggarwal, Nirmal Kaur

Critical review of the manuscript for important intellectual content: Vinay Sharma, Naveen Aggarwal, Nirmal Kaur

Supervision: Naveen Aggarwal, Nirmal Kaur

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** The project was funded by IIT Mandi iHub & HCI Foundation under grant number: IIT MANDI iHub/RD/2023-2025/04. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

This article was previously presented at the ICETTEM Conference at Amity France on May 19, 2025. The authors are grateful to the IIT Mandi iHub & HCI Foundation for funding under the project (Development of Multimodal and Multilingual Human Emotion Detection System) and Design Innovation Centre (DIC), UIET, Panjab University, Chandigarh.

References

1. Hudlicka E: Guidelines for designing computational models of emotions. International Journal of Synthetic Emotions. 2011, 2:26-79. 10.4018/jse.2011010103
2. Byun SW, Kim JH, Lee SP: Multi-modal emotion recognition using speech features and text-embedding. Applied Sciences. 2021, 11:7967. 10.3390/app11177967
3. Ayata D, Yaslan Y, Kamasak ME: Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. Journal of Medical and Biological Engineering. 2020, 40:149-157. 10.1007/s40846-019-00505-7
4. Ahmed N, Al Aghbari Z, Girija S: A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications. 2023, 17:200171. 10.1016/j.iswa.2022.200171
5. Issa D, Demirci MF, Yazici: Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control. 2020, 59:101894. 10.1016/j.bspc.2020.101894
6. Lian Z, Li Y, Tao JH, et al.: Expression analysis based on face regions in real-world conditions. International Journal of Automation and Computing. 2019, 17:96-107. 10.1007/s11633-019-1176-9
7. Nakisa B, Rastgoo MN, Rakotonirainy A, Maire F, Chandran V: Automatic emotion recognition using temporal multimodal deep learning. IEEE Access. 2020, 8:225463-225474. 10.1109/access.2020.3027026
8. He J, Yu X, Sun B, Yu L: Facial expression and action unit recognition augmented by their dependencies on graph convolutional networks. Journal on Multimodal User Interfaces. 2021, 15:429-440. 10.1007/s12193-020-00363-7
9. Zou S, Huang X, Shen X, Liu H: Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation. Knowledge-Based Systems. 2022, 258:109978. 10.1016/j.knosys.2022.109978
10. Wang L, Li R, Wu Y, Jiang Z: A multiturn complementary generative framework for conversational emotion recognition. International Journal of Intelligent Systems. 2022, 37:5643-5671. 10.1002/int.22805
11. Rizvi S, Agrawal P, Challa J, Narang P: InFER: A multi-ethnic Indian facial expression recognition dataset. Proceedings of the 15th International Conference on Agents and Artificial Intelligence. 2023, 3:550-557. 10.5220/0011699400003393
12. Arya KV, Verma S, Gupta RK, Agarwal S, Gupta P: Face: A database for facial attribute detection in constrained and simulated unconstrained environments. CoDS COMAD 2020: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. 2020, 185-189. 10.1145/3371158.3371182
13. Tewari S, Mehta S, Srinivasan N: IIMI Emotional Face Database. (2023). Accessed: July 27, 2025:

https://osf.io/f7zbv/.
14. Dachawar M: Dataset of Indian face images with various expressions. Mendeley Data. 2024, 10.17632/zfcd4bny82.1
15. Gangamohan P, Botsa KK, Rambabu B, Gangashgetty SV: IIIT-H TEMD semi-natural emotional speech database from professional actors and non-actors. Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020, 1538-1545.
16. Happy SL, Patnaik P, Routray A, Guha R: The Indian spontaneous expression database for emotion recognition. IEEE Transactions on Affective Computing. 2017, 8:131-142. 10.1109/taffc.2015.2498174
17. Singh S, Benedict S: Indian semi-acted facial expression (iSAFE) dataset for human emotions recognition. Advances in Signal Processing and Intelligent Recognition Systems. Thampi SM, Hegde RM, Krishnan S (ed): Springer, Singapore; 2019. 1209:150-162. 10.1007/978-981-15-4828-4_13
18. Karani R, Desai S: IndEmoVis: A multimodal repository for in-depth emotion analysis in conversational contexts. Procedia Computer Science. 2024, 233:108-118. 10.1016/j.procs.2024.03.200
19. Singh L, Aggarwal N, Singh S: PUMAVE-D: Panjab University multilingual audio and video facial expression dataset. Multimedia Tools and Applications. 2023, 82:10117-10144. 10.1007/s11042-022-14102-5
20. ResNet-50 Implementation on Google Colab. (2025). Accessed: July 27, 2025: https://colab.research.google.com/github/yashclone999/ResNet_MODEL/blob/master/ResNet50.ipynb.
21. Geetha AV, Mala T, Priyanka D, Uma E: Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. Information Fusion. 2024, 105:102218. 10.1016/j.inffus.2023.102218