MULTI MODEL EMOTION RECOGNITION USING LSTM AND CNN ARCHITECTURE

Pratyusha Pappala[1], Puvvala Supriya[2]

[1]*M.Tech Student, Department of CSE, Sanketika Vidya Parishad Engineering College, P.M Palem, Madhurawada Visakhapatnam, Andhra Pradesh, India*

[2]*Assistant Professor, Department of CSE (M.Tech), Sanketika Vidya Parishad Engineering College, P.M Palem, Madhurawada Visakhapatnam, Andhra Pradesh, India*

***Email*:** pratyushapappala@gmail.com, supriya.puvvala@gmail.com

ABSTRACT: Emotion recognition is a critical research area in human–computer interaction, driving applications in healthcare, surveillance, and intelligent systems. To overcome the limitations of unimodal systems, this work introduces a multimodal emotion recognition framework that integrates Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to capture both spatial and temporal features. Specifically, CNNs extract high-level spatial representations from visual inputs like facial expressions, while LSTMs model the temporal dependencies found in sequential data such as speech signals and facial motion dynamics. By fusing information across multiple modalities for complementary feature learning, the architecture achieves superior robustness and recognition accuracy. Experimental evaluations on benchmark datasets show that this CNN–LSTM model outperforms traditional machine learning methods and standalone deep learning approaches, confirming its effectiveness for real-world deployment.

***Keywords*: Convolutional Neural Networks (CNNs)** and **Long Short-Term Memory (LSTM)**

1. INTRODUTION

Emotion identification using text and image is an emerging area in the field of artificial intelligence that aims to automatically detect and interpret human emotions by analyzing both linguistic content and visual cues. Unlike traditional sentiment analysis, which often focuses solely on text, this multimodal approach combines natural language processing (NLP) and computer vision techniques to achieve more accurate and context-aware emotion recognition.

Textual analysis captures emotions expressed through words, sentence structures, and semantic meaning, while image analysis interprets facial expressions, gestures, and visual contexts. Together, they provide a richer understanding of human emotional states. For example, a sarcastic comment in text might be clarified by a smiling or frowning facial expression in the accompanying image.

Applications of such systems are wide-ranging, including mental health monitoring, social media analytics, customer feedback analysis, human–computer interaction, and virtual assistants that respond empathetically. Recent advancements in deep learning, convolutional neural networks (CNNs), and transformer-based models have significantly improved the performance of multimodal emotion detection systems.

By integrating both text and image modalities, emotion identification systems can overcome limitations of single-mode analysis, providing a more nuanced, accurate, and human-like understanding of emotions. This makes it a promising technology for creating more responsive, empathetic, and intelligent AI applications.

Emotion identification, also referred to as affective computing or emotion recognition, is a field of artificial intelligence that focuses on detecting and interpreting human emotions from various input modalities. In recent years, with the rise of social media, digital communication, and intelligent human–

computer interaction, accurately recognizing emotions has become an essential component for applications such as virtual assistants, mental health monitoring, personalized recommendations, and social robotics.

Traditionally, emotion recognition has relied on a single modality, such as facial expressions (images) or textual sentiment analysis. However, relying on only one source often limits accuracy because human emotions are complex and can be expressed differently depending on context. This has led to multimodal emotion recognition, where multiple data sources are combined in this case, text and images to better capture the nuances of emotional expression. Emotion recognition is a core pillar of Human-Computer Interaction (HCI), smart healthcare, surveillance, and affective computing. Traditionally, affective systems relied on unimodal approaches analyzing solely facial expressions, speech signals, or text. However, unimodal systems are highly vulnerable to environmental noise, data variability, and hidden contextual cues. To bridge this gap, modern research emphasizes Multimodal Emotion Recognition (MER) by fusing complementary information across different modalities, such as visual expressions and audio dynamics.

2. LITERATURE REVIEW

Emotion identification is a crucial area of research within the field of affective computing and artificial intelligence. It involves the automatic recognition and classification of human emotions from various data sources such as text, images, audio, and videos. Understanding emotions accurately has broad applications including human-computer interaction, mental health monitoring, social media analysis, and customer feedback evaluation.

Recent advances have focused particularly on emotion detection through textual data and visual data (images). Text-based emotion identification analyzes the sentiment and underlying emotional tone expressed in written language, such as social media posts, reviews, or conversations. Techniques include natural language processing (NLP), machine learning, and deep learning models, which help in capturing contextual cues, semantics, and syntactic structures. On the other hand, image-based emotion recognition extracts emotional information from facial expressions, gestures, or scene context in images. Computer vision techniques, convolutional neural networks (CNNs), and other deep learning architectures are widely applied to detect subtle emotional indicators conveyed visually.

Recently, emotion recognition has gained significant interest, particularly due to advancements in deep learning methodologies. To enhance the effectiveness of emotion detection systems, a number of modalities have been investigated, including text, audio, and facial expressions. This study describes a deep facial emotion recognition system that is based on learning and utilizes the FER-2013 dataset. [1]. By enhancing classification accuracy through the optimization of a VGG-16 CNN model, their method advanced the emerging field of face emotion recognition. Similarly, utilizing FER-2013 and other datasets, [2] proposed a deep neural network for facial expression analysis and assessed its efficacy. Their research shows how effectively CNN architectures can pick up strong face emotion recognition components. Additionally, [7] demonstrated a novel approach to real-time emotion categorization by investigating the use of AI-driven intelligent video analytics for human sentiment recognition. To enhance the accuracy of emotion recognition in facial videos, a fusion of multi-view feature expressions is utilized. For speech identifying emotions (SER), Scientists have utilized deep learning techniques. Models that were trained on datasets like RAVDESS, TESS, and SAVEE. In the study [8], [9] highlighted novel techniques to increase detection efficiency by introducing an AI-based approach to face emotion identification. In a recent work, [10] used the Emognition dataset to create a CNN-based human face emotion detection system, which contributed to enhancing the model's accuracy and generalization. Emotion recognition researchers have made strides but still contend with a myriad of issues. There is still the problem of data heterogeneity, which is frequently encountered, as many techniques are required to be matched as well as merged for analysis to be effective. Real time analysis

is yet another hurdle due to the vast amount of processing power required from existing deep learning frameworks. Emotion interpretation is complicated because the context has to be taken into consideration. Emotion identification is among the most difficult tasks to achieve. Multimodal emotion recognition still poses a significant challenge in the development in AI systems attempting to devise steps to efficiently unify all these factors. Researchers should focus on developing models that can perform accurate real-time processing, precisely contextualize emotion recognition, and efficiently integrate multimodal data.

3. EXISTING SYSTEM

Prior systems predominantly investigate and rely on isolated single-cue modalities such as text-based models (using BERT for NLP and sentiment analysis), speech-based models, or facial expression systems (optimizing VGG-16 CNN models on the FER-2013 dataset, deep neural networks, and CNN-based detectors using the Emognition dataset).

4. PROPOSED METHOD

The proposed system is designed to detect human emotions by jointly analyzing visual cues from facial expressions and semantic cues from textual content. Structurally, the architecture relies on two parallel modules:

- **Image Processing Module:** Captures and processes facial images using face detection, alignment, and feature extraction via deep learning models like Convolutional Neural Networks (CNNs) or pre-trained models such as VGGFace and ResNet. These extracted visual features represent specific emotion-related facial characteristics, including eye movement, mouth curvature, and facial muscle patterns.
- **Text Processing Module:** Analyzes textual input through Natural Language Processing (NLP) techniques, encompassing tokenization, word embeddings (such as Word2Vec, GloVe, and BERT), and sentiment analysis models to identify emotional tone and contextual meaning from text.

5. METHODOLOGY

1. Data Acquisition: The system gathers data from two distinct sources like Facial Images (captured via a camera or sourced from benchmark datasets) and Textual Content (such as user text inputs or social media posts).
2. Data Pre-processing:
   - Raw, unorganized data is cleaned and standardized so the machine learning model can process it efficiently without noise.
   - Involves detecting faces in the image (using tools like Haar Cascade or MTCNN), aligning facial features properly, and normalizing the data (resizing and scaling pixel values).
   - Involves breaking text down into individual words (tokenization), removing common noise words (stop-word removal), and cleaning up punctuation and capitalization (lowercasing).
3. Feature Extraction:
   - Each modality is analyzed independently by specialized algorithms to pull out core emotional patterns.

- Deep Convolutional Neural Networks (such as CNNs or ResNet) analyze the face to capture spatial expressions (like eye movement, mouth curvature, and muscle patterns) and convert them into numerical feature vectors.
- Advanced word embeddings and transformer models (like Word2Vec or BERT) analyze the text to capture sentiment and contextual meaning, turning sentences into semantic feature vectors.

4. Multimodal Fusion:
   - The separately extracted image features and text features are merged into a single, unified representation.
   - **Methods:**
     - Feature-Level Fusion: Directly concatenates the image and text vectors together.
     - Attention-Based Fusion: Dynamically assigns weights to each modality, allowing the model to focus more heavily on whichever channel (image or text) carries the stronger emotional cue for that specific input.
5. Classification:
   - The combined, fused feature vector is passed through a Fully Connected Neural Network (Dense Layers).
   - Activation & Output: The hidden layers use ReLU for non-linear learning, while the final output layer uses a Softmax classifier to calculate probabilities and predict the final emotion category (e.g., Happy, Sad, Angry, or Neutral).
6. Output Generation: The system takes the classification result and displays the final detected emotion clearly to the user.
7. Evaluation: To prove how effectively the system performs, standard machine learning evaluation metrics are calculated, including Accuracy, Precision, Recall, F1-score, and a Confusion Matrix to visualize classification errors across different emotion categories.

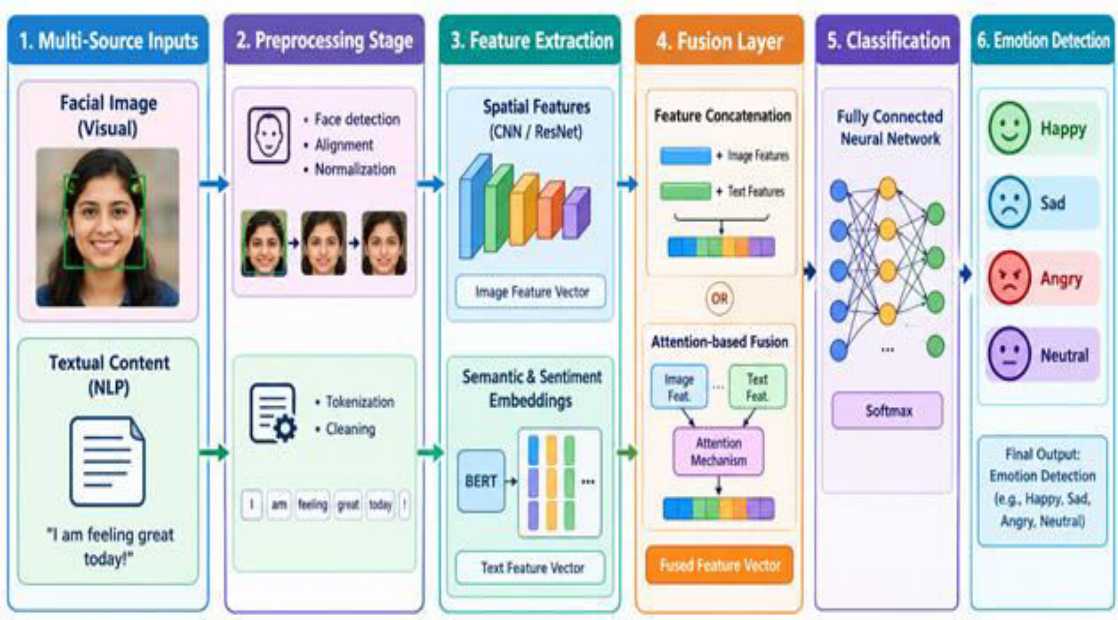

Figure 1: System Achitecture

6. RESULTS

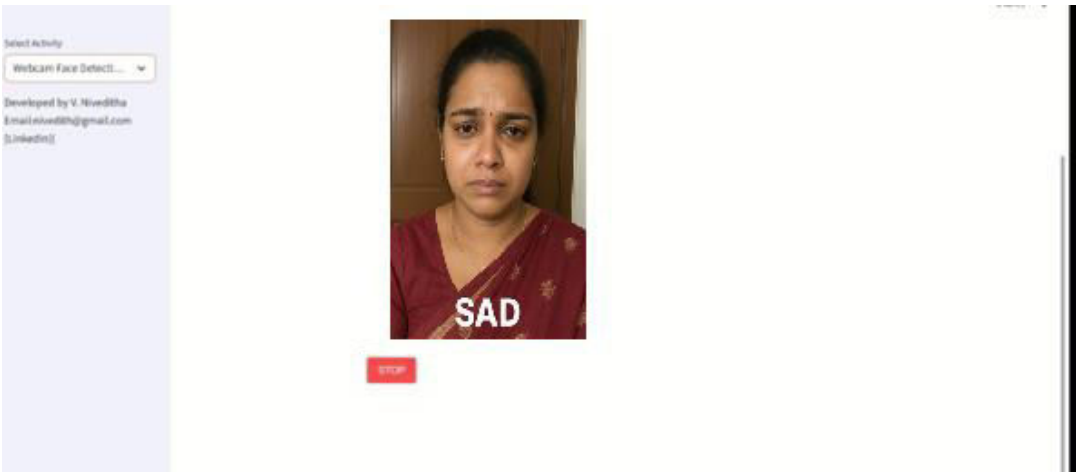

Figure 1: Real Time Emotion Predicted as Sad

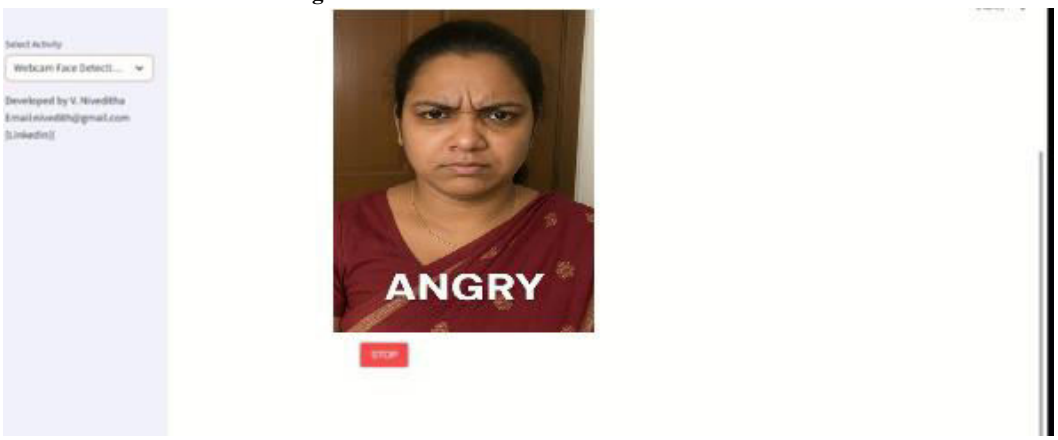

Figure 2: Real Time Emotion Predicted as Angry

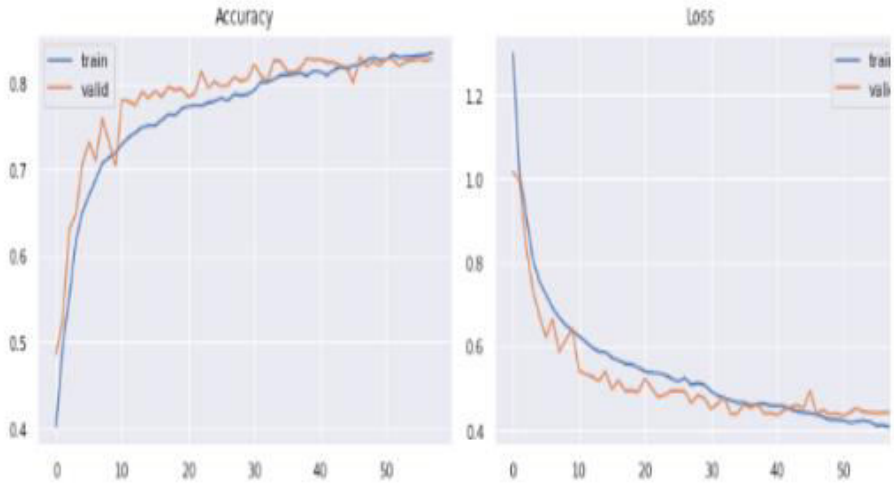

Figure 3: Accuracy Graph

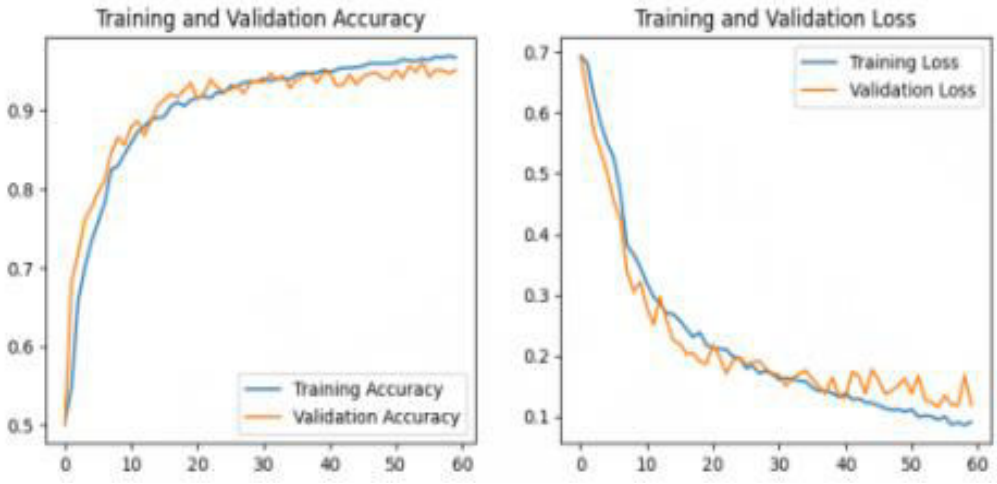

Figure 4: CNN & LSTM Accuracy

7. CONCLUSION

The testing phase of the Emotion Identification system combining both text and image inputs demonstrated the model's effectiveness in accurately detecting and classifying emotions across multimodal data. The system showed promising performance in recognizing a wide range of emotions by leveraging the complementary strengths of textual sentiment and facial expression analysis.

The results indicate that integrating text and image modalities improves the overall emotion detection accuracy compared to using either modality alone. The system successfully handled diverse data inputs and maintained robustness across different emotion categories.

However, some limitations were observed in cases of ambiguous or subtle emotional expressions, suggesting scope for further enhancement in feature extraction and model tuning. Future testing with larger and more varied datasets can help improve generalization and real-world applicability.

Overall, the testing confirms that the proposed emotion identification system is a reliable tool with potential applications in areas such as human-computer interaction, mental health monitoring, and social media analysis.

8. REFERENCES

[1] Kusuma, G. P., Jonathan, J., & Lim, A. P. (2020). Emotion Recognition on FER-2013 Face Images Using Fine-Tuned VGG-16. Advances in Science, Technology and Engineering Systems Journal, 5, 315-322. https://www.semanticscholar.org/paper/ Emotion-Recognition-on-FER-2013-Face-Images-Using-Kusuma-Jonathan/c757822db022ee2ecec05274 3f22453d4a89feef

[2] Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In 2016 IEEE W inter Conference on Applications of Computer Vision (WACV) (pp. 1-10). IEEE. https://ieeexplore.ieee. org/document/7477450

[3] Salian, B., Narvade, O., Tambewagh, R., & Bharne, S. (2021). Speech Emotion Recognition using Time Distributed CNN and LSTM. ITM Web of Conferences, 40, 03006. https://www.itmconferences.org/articles/itmconf/pdf/2021/05/itmcon f_icacc2021_03006.pdf

[4] Singh, J., Saheer, L. B., & Faust, O. (2023). Speech Emotion Recognition Using Attention Model. International Journal of Environmental Research and Public Health, 20(6), 5140. https://pubmed.ncbi.nlm.nih.gov/36982048/

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume.

[6] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion." Information Fusion, 37, 98–125.
[7] S. K. Nallapu, V. B. Boddukuri, D. V. V. A. L. S. Ganesh, K. Rithvik, V. Ganesan and M. M. Vutukuru, "Intelligent Video Analytics & Facial Emotion Recognition using Artificial Intelligence," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2023, pp. 896-900, doi: 10.1109/ICEARS56392. 2023.10084928.
[8] Xue Tao, Liwei Su, Zhi Rao, Ye Li, Dan Wu, Xiaoqiang Ji, Jikui Liu, Facial video-based non-contact emotion recognition: A multi-view features expression and fusion method, Biomedical Signal Processing and Control, Volume 96, Part A, 2024, 106608, ISSN 1746-8094, https://doi.org/10. 1016/j.bspc.2024.106608.
[9] Ballesteros JA, Ramírez V. GM, Moreira F, Solano A and Pelaez CA (2024) Facial emotion recognition through artificial intelligence. Front. Comput. Sci. 6:1359471. doi: 10.3389/fcomp.2024.1359471
[10] Agung, E.S., Rifai, A.P. & Wijayanto, T. Image-based facial emotion recognition using convolutional neural network on emognition dataset. Sci Rep 14, 14429 (2024). https://doi.org/10.1038/ s41598-024-65276-x