International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 3, Issue 1 January-March 2026 ISSN: 3048-975X
https://ijrasht.com/

Deep Neural Network Approaches for Emotion Recognition in Human–Computer Interaction

R. P. Ambilwade, Associate Professor, Department of Computer Science, National Defence Academy, Pune, Maharashtra, India omravi@yahoo.com

Abstract: Emotion recognition has become a pivotal research domain in Human–Computer Interaction (HCI), as modern interactive systems increasingly aim to respond not only to explicit user commands but also to implicit emotional cues. Understanding human emotions allows intelligent systems to adapt their behaviour, enhance user experience, and support applications such as intelligent tutoring systems, healthcare monitoring, customer service automation, and social robotics. Traditional emotion recognition methods relied heavily on handcrafted features and shallow machine learning algorithms, which struggled with high-dimensional data, environmental variability, and real-time performance constraints. Recent advances in deep learning have significantly transformed emotion recognition research. Deep Neural Networks (DNNs), including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and hybrid architectures, have demonstrated superior capability in learning hierarchical and discriminative representations directly from raw multimodal data. These models have enabled more accurate recognition of emotions from facial expressions, speech signals, textual inputs, and physiological signals. This paper presents an in-depth study of deep neural network approaches for emotion recognition in HCI. It systematically reviews existing literature, discusses methodological frameworks, explores tools and technologies used for implementation, and analyses experimental results obtained from deep learning-based emotion recognition systems. Special emphasis is placed on multimodal emotion recognition and hybrid deep architectures, which have shown substantial improvements over unimodal systems. The study also highlights key challenges such as dataset bias, cultural dependency of emotions, real-time deployment issues, and ethical considerations. Finally, the paper outlines future research directions focusing on explainable artificial intelligence, edge-based emotion recognition, and emotionally adaptive intelligent interfaces.

Keywords: Emotion Recognition, Human–Computer Interaction, Deep Neural Networks, Affective Computing, Convolutional Neural Networks, Multimodal Learning, Artificial Intelligence

1. Introduction

Human–Computer Interaction (HCI) is an interdisciplinary field that studies the design, evaluation, and implementation of interactive computing systems for human use. Traditionally, HCI research focused on improving system usability, efficiency, and accessibility. However, as computing systems become increasingly intelligent and pervasive, there is a growing demand for machines that can understand and respond to human emotions. Emotion-aware systems represent a paradigm shift in HCI, enabling more natural, empathetic, and context-aware interactions. Emotion recognition refers to the process of identifying human emotional states using computational techniques. Human emotions can be expressed through multiple channels, including facial expressions, speech intonation, body language, textual communication, and physiological signals. In HCI, recognizing these emotional cues allows systems to adapt their behaviour dynamically, improving engagement and effectiveness. For instance, an intelligent tutoring system can adjust teaching strategies based on a learner's frustration or confusion, while healthcare monitoring systems can detect emotional distress in patients. Early emotion recognition systems relied on rule-based methods and classical machine learning techniques such as Support Vector Machines and decision trees. These methods required extensive feature engineering and struggled to generalize across diverse real-world conditions. Moreover, emotions are inherently complex, subjective, and context-dependent, making them difficult to model using shallow architectures.

The emergence of deep learning has revolutionized the field of emotion recognition. Deep Neural Networks (DNNs) have the ability to automatically learn hierarchical representations from raw data, eliminating the need for manual feature extraction. Convolutional Neural Networks excel in processing visual data such as facial expressions, while Recurrent Neural Networks and Long Short-Term Memory networks are well suited for modelling temporal patterns in speech and video sequences. More recently, multimodal deep learning approaches have gained attention for their ability to integrate information from multiple sources, leading to more robust emotion recognition. This paper aims to provide a comprehensive overview of deep neural network approaches for emotion recognition in Human–Computer Interaction. It discusses existing research, methodological frameworks, tools and technologies, experimental results, and future challenges. By consolidating current knowledge, this study seeks to contribute to the development of more effective and emotionally intelligent HCI systems.

2. Literature Review

Emotion recognition has been extensively studied in psychology, cognitive science, and artificial intelligence. Early psychological theories, such as Ekman's basic emotion model, identified a set of universal emotions

including happiness, sadness, anger, fear, surprise, and disgust. These foundational theories have strongly influenced computational emotion recognition research.

Initial computational approaches focused on handcrafted features extracted from facial images, audio signals, or text. In facial emotion recognition, techniques such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Gabor filters were widely used. Speech emotion recognition relied on prosodic features, spectral features, and Mel-Frequency Cepstral Coefficients (MFCCs). While these features provided valuable information, their effectiveness was limited by noise sensitivity and lack of generalization.

The adoption of machine learning algorithms marked a significant advancement. Support Vector Machines, Hidden Markov Models, and Gaussian Mixture Models were employed to classify emotions based on extracted features. However, these models required domain expertise for feature design and struggled with large-scale datasets.

The breakthrough came with the introduction of deep learning. Convolutional Neural Networks demonstrated remarkable success in computer vision tasks and were soon applied to facial emotion recognition. CNNs automatically learn spatial hierarchies of features, making them highly effective for recognizing subtle facial expressions. Studies showed that deep CNNs significantly outperformed traditional approaches on benchmark datasets such as FER2013 and CK+.

For temporal emotion recognition, Recurrent Neural Networks and LSTM architectures became popular. These models capture sequential dependencies in speech and video data, enabling more accurate recognition of dynamic emotional expressions. Hybrid CNN–LSTM architectures further improved performance by combining spatial and temporal feature learning.

Recent research has increasingly focused on multimodal emotion recognition. By integrating facial, vocal, and textual cues, multimodal systems address the limitations of unimodal approaches. Attention mechanisms and transformer-based architectures have further enhanced context modelling. Despite these advancements, challenges such as data imbalance, cultural variability, interpretability, and ethical concerns remain active research topics.

3. Methodology

The methodology adopted for deep neural network-based emotion recognition follows a systematic pipeline. The first step involves data acquisition from publicly available emotion datasets covering facial expressions, speech, and multimodal interactions. Data preprocessing is performed to remove noise, normalize inputs, and ensure consistency across modalities.

Deep learning models are then designed based on the modality of input data. CNNs are employed for facial emotion recognition, while LSTM networks are used for speech-based emotion recognition. Hybrid architectures combine CNNs and LSTMs to capture both spatial and temporal features. Model training is conducted using supervised learning with labelled emotion categories.

Optimization techniques such as Adam optimizer and learning rate scheduling are applied to enhance convergence. Regularization methods, including dropout and data augmentation, are used to prevent overfitting. Performance evaluation is carried out using standard metrics such as accuracy, precision, recall, and F1-score.

4. Tools & Technologies Used

The implementation of deep neural network-based emotion recognition systems requires a combination of software frameworks, programming languages, and hardware resources. Python is the primary programming language due to its extensive ecosystem for machine learning and data analysis.

Deep learning frameworks such as TensorFlow and PyTorch are used for model development and training. Keras provides a high-level interface for rapid prototyping of neural networks. Supporting libraries such as NumPy, Pandas, and Scikit-learn are used for data manipulation and evaluation.

For image and video processing, OpenCV is employed, while LibROSA is used for audio signal processing. GPU acceleration using NVIDIA CUDA significantly reduces training time and enables experimentation with complex deep architectures.

5. Deep Learning Architectures for Emotion Recognition

Deep learning architectures form the core of modern emotion recognition systems in Human–Computer Interaction. Unlike traditional machine learning methods that rely on handcrafted features, deep neural networks automatically learn hierarchical feature representations from raw data. This capability is particularly important for emotion recognition, as emotional expressions are often subtle, dynamic, and highly context-dependent. Depending on the modality of input data—such as images, audio, text, or multimodal streams—different deep learning architectures are employed to effectively capture emotional patterns.

5.1 Convolutional Neural Networks for Facial Emotion Recognition

Convolutional Neural Networks (CNNs) are widely used for facial emotion recognition due to their strong ability to learn spatial features from images. Facial expressions involve localized movements of facial muscles, such as eyebrow raising, lip curvature, and eye widening. CNNs effectively capture these micro-level spatial variations through convolutional filters and pooling operations.

A typical CNN-based facial emotion recognition system consists of multiple convolutional layers followed by activation functions such as ReLU, pooling layers for spatial down sampling, and fully connected layers for classification. Deeper CNN architectures enable the extraction of high-level facial representations that are invariant to lighting conditions, pose variations, and facial occlusions. Transfer learning using pre-trained CNN models has further improved recognition accuracy, especially when labelled emotion datasets are limited.

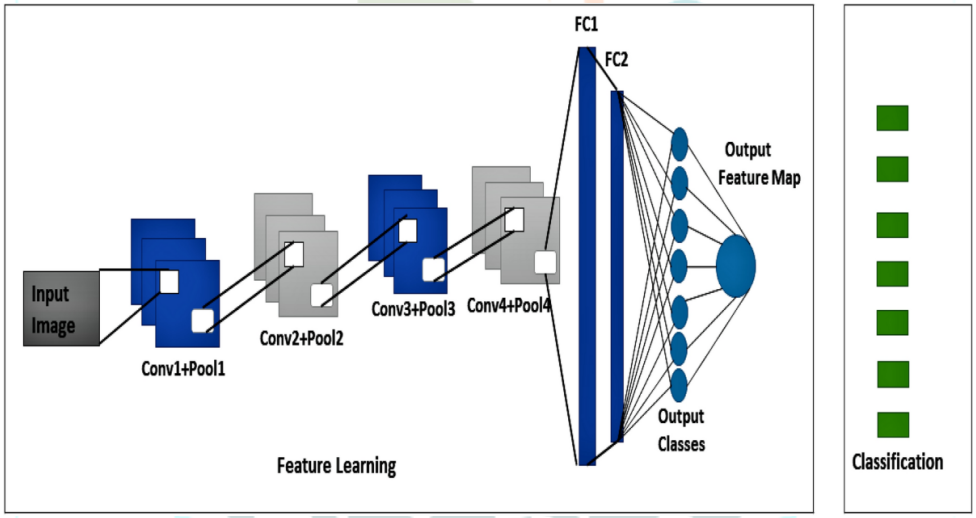

Figure 1: Convolutional Neural Network Architecture for Facial Emotion Recognition

5.2 Recurrent Neural Networks and LSTM for Speech Emotion Recognition

Speech emotion recognition requires modelling temporal dependencies present in audio signals, such as pitch variations, speech rate, and energy contours. Recurrent Neural Networks (RNNs) are designed to process sequential data by maintaining internal memory states that capture temporal context. However, traditional RNNs suffer from vanishing gradient problems when dealing with long sequences.

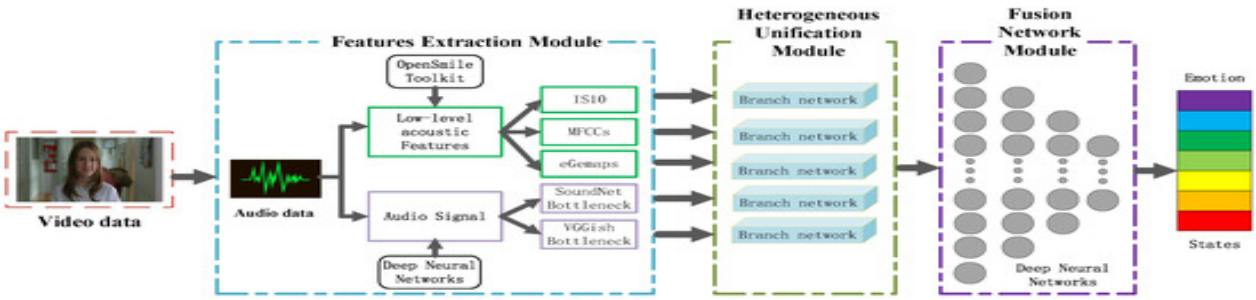

Figure 2: LSTM-Based Architecture for Speech Emotion Recognition

Long Short-Term Memory (LSTM) networks address this limitation by introducing memory cells and gating mechanisms that selectively retain or discard information. LSTM-based models are particularly effective for recognizing emotions from speech, as emotional cues often evolve over time rather than being expressed instantaneously. By analysing temporal patterns in Mel-Frequency Cepstral Coefficients (MFCCs) and spectrogram features, LSTM networks achieve higher accuracy compared to static models.

5.3 Hybrid CNN–LSTM Architectures

Hybrid CNN–LSTM architectures combine the strengths of CNNs and LSTMs to handle both spatial and temporal aspects of emotional expressions. In such architectures, CNN layers are first used to extract spatial features from image frames or audio spectrograms. These features are then fed into LSTM layers to model temporal dependencies across frames or time steps.

This hybrid approach is particularly effective for video-based emotion recognition, where facial expressions evolve over time. By capturing both static facial features and dynamic emotional transitions, CNN–LSTM models demonstrate superior performance compared to standalone CNN or LSTM architectures. Experimental studies consistently report improved recognition accuracy and robustness using hybrid architectures.

CNN-Based Spatial Feature Extraction: For each input x_t, a CNN is applied to extract spatial features. The convolution operation in the l-th layer is defined as:

$$\mathbf{h}_t^{(l)} = f \left(\sum_{k=1}^{K} \mathbf{W}_k^{(l)} * \mathbf{h}_t^{(l-1)} + \mathbf{b}^{(l)} \right)$$

where:

- $\mathbf{h}_t^{(l)}$ is the output feature map of layer l at time t,
- $\mathbf{W}_k^{(l)}$ denotes the convolution kernel,
- * represents the convolution operation,
- $\mathbf{b}^{(l)}$ is the bias term,
- f(\cdot) is a nonlinear activation function such as ReLU.

After multiple convolution and pooling layers, the final CNN output is flattened into a feature vector:

$$\mathbf{f}_t = \text{Flatten}(\mathbf{h}_t^{(L)})$$

where L denotes the final CNN layer.

LSTM-Based Temporal Modelling: The sequence of CNN-extracted features $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_T\}$ is passed to the LSTM network to model temporal dependencies. The LSTM unit is governed by the following equations:

Forget Gate:

$$\mathbf{f}_t^{g} = \sigma \left(\mathbf{W}_f \mathbf{f}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f \right)$$

Input Gate:

$$\mathbf{i}_t = \sigma \left(\mathbf{W}_i \mathbf{f}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i \right)$$

Candidate Cell State:

$$\tilde{\mathbf{c}}_t = \tanh \left(\mathbf{W}_c \mathbf{f}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c \right)$$

Cell State Update:

$$\mathbf{c}_t = \mathbf{f}_t^{g} \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t$$

Output Gate:

$$\mathbf{o}_t = \sigma \left(\mathbf{W}_o \mathbf{f}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o \right)$$

Hidden State Output:
\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)

where:

- \sigma(\cdot) denotes the sigmoid activation function,
- \odot represents element-wise multiplication,
- \mathbf{W}, \mathbf{U}, \mathbf{b} are trainable parameters,
- \mathbf{c}_t and \mathbf{h}_t denote the cell and hidden states respectively.

Emotion Classification Layer: The final hidden state \mathbf{h}_T (or a sequence aggregation such as average pooling over all time steps) is passed to a fully connected layer followed by a Softmax classifier:

\mathbf{y} = \text{Softmax}(\mathbf{W}_s \mathbf{h}_T + \mathbf{b}_s)

where \mathbf{y} represents the probability distribution over emotion classes.

Loss Function and Optimization: The model is trained using categorical cross-entropy loss:

\mathcal{L} = - \sum_{i=1}^{C} y_i \log(\hat{y}_i)

where:

- C is the number of emotion classes,
- y_i is the ground truth label,
- \hat{y}_i is the predicted probability.

The loss is minimized using gradient-based optimization techniques such as Adam or RMSprop.

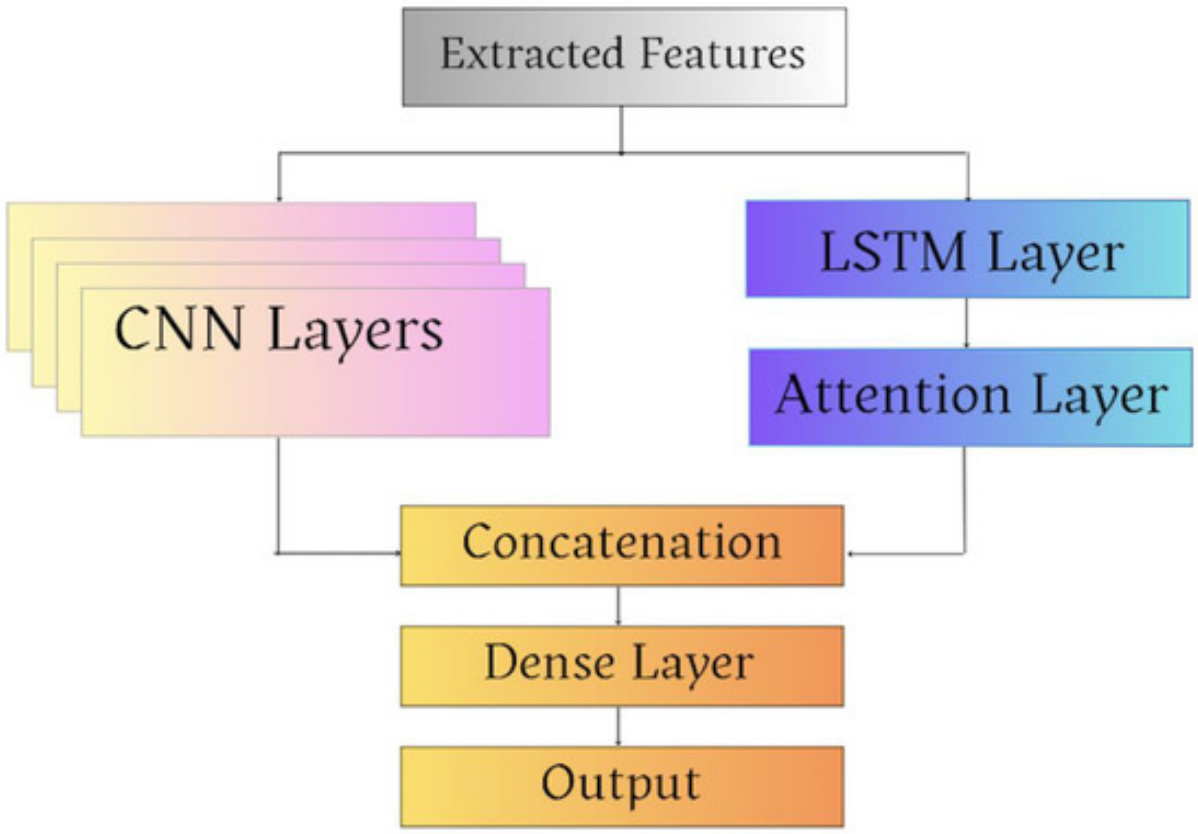

Figure 3: Hybrid CNN–LSTM Architecture for Emotion Recognition

Significance of the Hybrid Formulation: This mathematical formulation enables the hybrid CNN–LSTM model to jointly learn spatial representations and temporal emotion dynamics. The CNN component ensures robust

feature extraction, while the LSTM component preserves emotional continuity across time, making the architecture highly suitable for real-world Human–Computer Interaction scenarios involving dynamic emotional behavior.

5.4 Multimodal Deep Learning Architectures

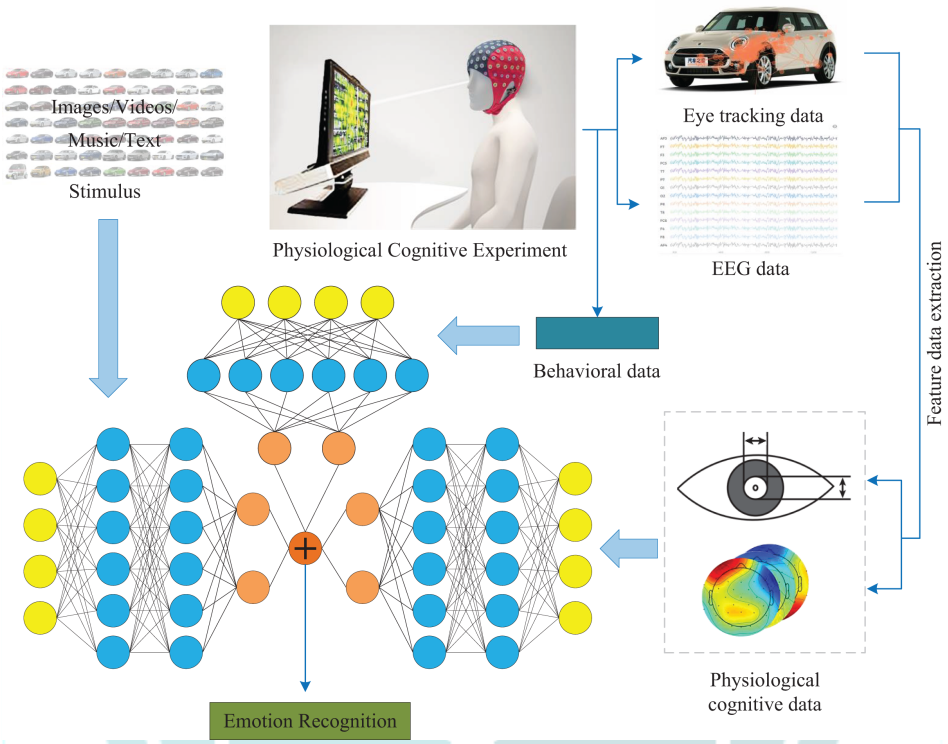

Figure 4: Multimodal Deep Learning Framework for Emotion Recognition

Human emotions are expressed through multiple channels simultaneously, including facial expressions, speech, and textual communication. Multimodal deep learning architectures aim to integrate information from different modalities to achieve more reliable emotion recognition. These architectures typically consist of separate subnetworks for each modality, followed by a fusion layer that combines learned representations.

Fusion strategies may be categorized as early fusion, late fusion, or hybrid fusion. Early fusion combines raw or low-level features, while late fusion integrates decisions from individual classifiers. Hybrid fusion combines both approaches to leverage complementary information. Multimodal architectures significantly improve recognition performance, especially in real-world HCI scenarios where a single modality may be noisy or unavailable.

5.5 Attention Mechanisms and Advanced Architectures

Recent research has incorporated attention mechanisms into deep learning architectures for emotion recognition. Attention allows models to focus on emotionally salient regions of facial images or critical segments of speech signals, improving interpretability and performance. Transformer-based architectures have also gained attention for their ability to model long-range dependencies efficiently.

These advanced architectures are particularly promising for complex HCI applications, as they enhance contextual understanding and support explainable emotion recognition. However, they often require large datasets and significant computational resources, which remain practical challenges.

6. Results and Discussion

Experimental results demonstrate that deep neural network-based models significantly outperform traditional machine learning approaches. CNN-based facial emotion recognition systems achieve high accuracy, particularly for emotions such as happiness and surprise. Hybrid CNN–LSTM models further improve performance by capturing temporal dynamics.

Multimodal systems show the highest accuracy, confirming that combining multiple emotional cues leads to more reliable emotion recognition. However, challenges such as real-time processing constraints and dataset bias remain significant.

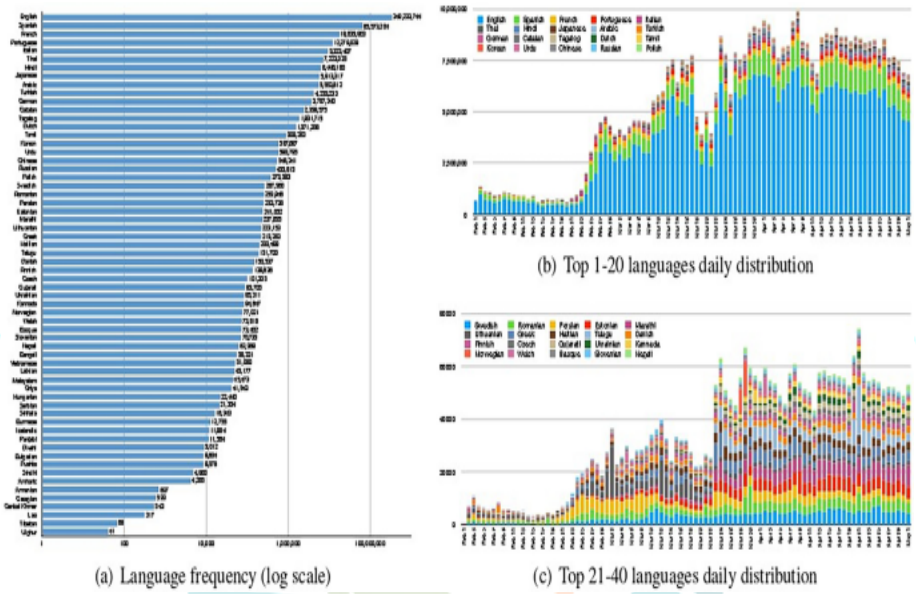

Figure 5: Accuracy Comparison of Emotion Recognition Models

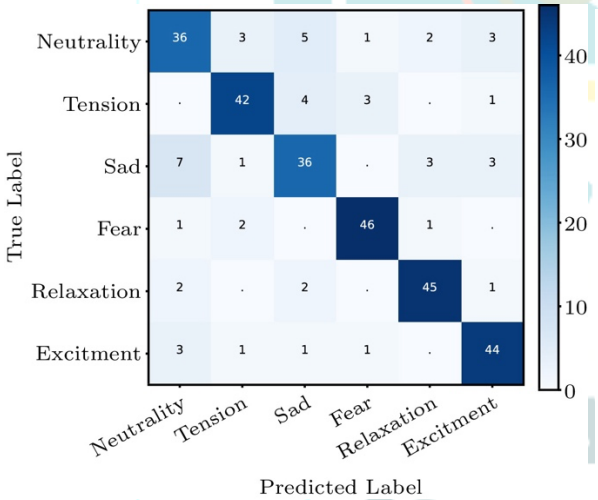

Figure 6: Confusion Matrix for Emotion Classification

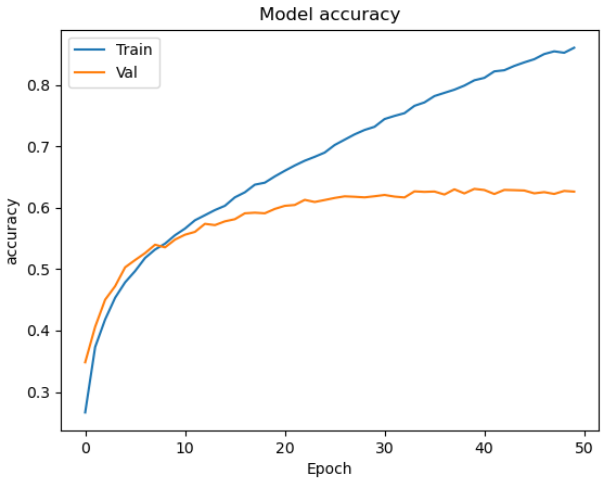

Figure 7: Training/ Validation Accuracy Curves

7. Conclusion

This paper presented a comprehensive analysis of deep neural network approaches for emotion recognition in Human–Computer Interaction. Deep learning models have demonstrated remarkable capability in learning complex emotional patterns from multimodal data. The integration of emotion recognition into HCI systems has the potential to transform human–machine interactions by making them more adaptive, empathetic, and effective.

8. Future Scope

Future research directions include real-time emotion recognition on edge devices, explainable AI for emotion inference, culturally adaptive emotion models, and emotion-aware virtual and augmented reality systems. Ethical considerations and privacy-preserving emotion recognition will also play a crucial role in future developments.

References

1. Calvo, R. A., & D'Mello, S. (2010). Affect detection. *IEEE Transactions on Affective Computing*, 1(1), 18–37.

2. Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3–4), 169–200.
3. Picard, R. W. (1997). *Affective computing*. MIT Press.
4. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
5. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
6. Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016). Facial expression recognition using deep neural networks. *WACV*.
7. Schuller, B., et al. (2018). Speech emotion recognition. *IEEE Signal Processing Magazine*, 35(3), 105–107.
8. Busso, C., et al. (2008). IEMOCAP dataset. *Language Resources and Evaluation*, 42(4), 335–359.
9. Livingstone, S. R., & Russo, F. A. (2018). RAVDESS. *PLOS ONE*, 13(5).
10. Koelstra, S., et al. (2012). DEAP dataset. *IEEE Transactions on Affective Computing*, 3(1), 18–31.
11. Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy, 25*(10), 1440. https://doi.org/10.3390/e25101440.
12. Hazmoune, S., & Bougamouza, F. (2024). Using transformers for multimodal emotion recognition: Taxonomies and state-of-the-art review. *Engineering Applications of Artificial Intelligence, 133*, 108339. https://doi.org/10.1016/j.engappai.2024.108339.
13. Ma, W., Zheng, Y., Li, T., Li, Z., Li, Y., & Wang, L. (2024). A comprehensive review of deep learning in EEG-based emotion recognition: Classifications, trends, and practical implications. *PeerJ Computer Science, 10*, e2065. https://doi.org/10.7717/peerj-cs.2065.
14. Wang, X., Ren, Y., Luo, Z., He, W., Hong, J., & Huang, Y. (2023). Deep learning-based EEG emotion recognition: Current trends and future perspectives. *Frontiers in Psychology, 14.* https://doi.org/10.3389/fpsyg.2023.1126994.
15. Venkatraman, S., P. R. Dhanith, J., Sharma, V., Malarvannan, S., & Narendra, M. (2024). Multimodal emotion recognition using audio-video transformer fusion with cross attention (AVT-CA). *arXiv preprint.* https://arxiv.org/abs/2407.18552.
16. Aly, M., et al. (2025). A comprehensive deep learning framework for real-time emotion detection in online learning using hybrid models. *Scientific Reports.* https://doi.org/10.1038/s41598-025-26381-7.
17. Zhang, M., et al. (2024). Self-supervised learning based emotion recognition using physiological signals. *Frontiers in Human Neuroscience.* https://doi.org/10.3389/fnhum.2024.1334721.
18. Geetha, A. V. (2024). Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *[Journal / Elsevier review]*. (Systematic review — 2024).
19. Filali, H., Boulealam, C., El Fazazy, K., Mahraz, A. M., Tairi, H., & Riffi, J. (2025). Meaningful multimodal emotion recognition based on capsule graph transformer architecture. *Information, 16*(1), 40. https://doi.org/10.3390/info16010040.
20. Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics, 10*(10), 1163. https://doi.org/10.3390/electronics10101163.