International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Hybrid CNN–LSTM Model for Emotion Prediction from Visual Data in Children

[1]Vikas Jangra*, [2]Sumeet Gill, [3]Binny Sharma, [4]Archna Kirar

Submitted: 17/10/2024 **Revised:** 20/11/2024 **Accepted:** 01/12/2024

Abstract: Emotion recognition is one of the crucial components of human-computer interaction that allows systems to respond intelligently to user emotions. In this paper, a hybrid framework for emotion recognition based on facial and gesture-based expression models is proposed. This framework utilizes a deep convolutional neural network for facial expression analysis and an LSTM for gesture recognition, thus providing state-of-the-art performance. A multimodal fusion technique combines the features of both modalities to boost the accuracy in emotion classification. Extensive evaluation was performed on benchmark datasets with an impressive accuracy of 87.5% to predict correct emotions across six basic emotions, a neutral state, valence, and nine complex emotions. Emotion recognition in children aged between 4 and 14 years is challenging since their emotional expressions are subtle and are in a developmental stage, with very few published works in this area. The novelty of this paper is a multimodal emotion recognition model for children, incorporating facial expressions and gestures, novel architecture, and multimodal input fusion, offering a richer and more context-aware emotion recognition framework, especially beneficial in dynamic child interaction environments. This work advances emotion recognition systems for younger populations with applications in education, child development, and healthcare.

Keywords and phrases: *Multimodal Emotion Recognition, Facial Expression Analysis, Gesture Recognition, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM)*

1 Introduction

Human-Computer Interaction (HCI) [15] is very important in this age of interactive and intelligent computing. Emotion and gesture recognition are significant innovations in this area. The last 40 years have seen dramatic changes in computing, and the recent innovations have allowed using 3D gestures on televisions, smartphones, and PCs to make HCI more interesting. The gesture recognition market has huge growth potential. Its annual growth rate is expected to be 22.2% by 2025.

Emotions are conscious responses characterized by intense mental activity and a strong sense of pleasure or displeasure. Humans convey their emotions through various channels, including facial expressions, speech, and body movements. Identifying emotions based on facial expressions is known as Facial Emotion Recognition (FER) [13], which focuses on six primary emotions: disgust, anger, fear, surprise, sadness, and happiness. Gestures are a form of nonverbal communication used to convey specific messages. In gesture recognition, a gesture refers to any physical body movement that can be interpreted by a motion sensor. The ability of computers to understand and respond to gestures enables them to execute commands based on these movements.

The applications of gesture recognition systems are diverse, ranging from lie detection to distance learning, photojournalism, tutoring systems, and biometrics [12]. In the last ten years, automatic facial expression recognition has received significant attention from the artificial intelligence and computer vision research communities, and substantial progress has been achieved in the area of emotion recognition using facial expressions [4]. Increasingly, research has focused on movement and gestures as integral parts of nonverbal communication in both human-human interaction (HHI) and human-computer interaction (HCI) [9].

Emotion recognition is a fast-growing area of research, wherein the majority of the studies are focused on emotion recognition in adults and elderly people by facial expressions and gestures. However, limited work has been done in the understanding of emotions in children, which is critical in many areas of education [14], health care [1], and child development [20][19]. It is inherently much more

[2]*Professor, Department of Mathematics, M. D. University, Rohtak 124001, Haryana, India*
[1,3,4]*Research Scholar, Department of Mathematics, M. D. University, Rohtak 124001, Haryana, India*
vikasjangra96.rs.maths@mdurohtak.ac.in[1,*],
drsumeetgill@mdurohtak.ac.in[2],
binny.rs.maths@mdurohtak.ac.in[3],
archnakirar219@gmail.com[4]

difficult to recognize children's emotions because the patterns of expression of emotions among children are not as developed or as stable as those of adults. In our proposed model, we address this gap by focusing on emotion recognition in children aged 4 to 14 years. The model uses the facial expressions as well as gestures of the upper body as inputs and relies on a hybrid approach of deep learning, which encompasses CNNs [2] for feature extraction and LSTM [22] for sequential pattern recognition [16]. This has been done to have a more robust and reliable approach to the recognition of emotions in children, thus setting the stage for child-friendly applications in a more successful manner.

A novel contribution of this paper is as follows:

- The model proposed here offers a multimodal framework for emotion recognition that integrates facial expressions and upper body movements to enable a better and more robust analysis. The majority of previous models consider only facial expressions, and only one considers speech and gestures. Our model is different in that it integrates two visual modalities.
- Unlike existing research on children's emotion recognition based on CNN-based models, our suggested approach presents a hybrid CNN–LSTM model processing facial expressions and upper body gestures. Facial expressions are derived based on spatial representation through CNN and are modelled as temporal sequences for gestures utilizing LSTM. These two modalities are subsequently combined via feature-level concatenation, which allows an integrated representation capturing static and dynamic emotional cues. This multimodal combination represents a major improvement in the capture of the subtle and progressive emotional expressions of children.

The rest of the paper is organized as follows: Section 2 briefly reviews the related work for this work. In Section 3, the methodology is discussed in detail, including the dataset and proposed model architecture. In Section 4, the experiment and results are discussed. Finally, the last section includes conclusions and future work.

2 Related Work

2.1 On Adults

Multimodal emotion recognition, which makes use of more than one source of data, such as facial expressions and gestures, has received a lot of attention lately. Recent advancements in multimodal emotion recognition have explored different approaches to enhance the accuracy of emotion detection by integrating facial expressions and body gestures. Gunes et al. (2005, 2006, 2009) [6], [7], [8] used the FABO dataset and showed improved accuracies with classifiers such as AdaBoost, Bayesian Networks, and Random Forest, achieving up to 94.66%. Chen et al. (2012) [5] and Barros et al. (2015) [3] incorporated appearance and motion features, showing notable results with SVM and CNNs, respectively. Keshari et al. (2019) [12] and Nunes et al. (2019) [18] used MultiSVM and CNN classifiers, which have obtained state-of-the-art accuracies on the AED-2 and FABO datasets, respectively. Verma et al. (2021) [20] used Grassmann manifolds for robust feature representation, while Ilyas et al. (2021) [10] utilized the combination of CNNs and LSTMs to model temporal dependencies to obtain 94.41% accuracy. Karatay et al. (2022) [11] and Wei et al. (2024) [21] further improved performance using CNN-Transformer frameworks and TLSTM architectures, achieving accuracies up to 99% and 98.42%, respectively. These studies highlight the effectiveness of combining facial and gesture modalities with advanced machine deep learning techniques for robust emotion recognition in adults.

2.2 On Children's

Child emotion recognition research with audio-visual modalities has seen tremendous growth over the past few years. Lopez-Rincon et al. (2019) [23] proposed a facial expression-based model with the NAO Robot utilizing CNN, obtaining 44.89% accuracy. Filntisis et al. (2019) [24] integrated body posture and facial expressions utilizing HMT networks in the same year, reaching 72% accuracy on the BRED (Baby Robot Emotion Database) dataset. Filntisis et al. (2021) [25] subsequently proposed a deep learning-based approach examining speech and visual behaviours from RGB and optical flow streams. Suhan et al. (2022) [26] reported 90% accuracy for child emotion recognition by utilizing CNN and audio-visual features on the EmoReact dataset. Manish Rathod (2022) [27] utilized explainable AI methods for higher recognition accuracy of 90.98% on a new dataset. Recently, Pandyan et al. (2023) [28] attained 97.8% accuracy for new born emotion detection based on CNN using the City Infants Faces Database. The research points out development in fusing facial expression, body position, speech, and gestures towards enhanced child emotion identification.

3 Methodology

3.1 Data Set

There are many datasets that are used for emotion recognition, which combine facial expressions and gestures to provide diverse modalities for robust analysis. FABO is a widely used resource that contains synchronized facial and body gesture data annotated with six basic emotions. Amrita Emotion Dataset-2 (AED-2) provides multimodal data, including detailed facial and body gestures, for recognizing seven emotional states. This is CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI), multimodal video clips that are annotated by facial expressions, gestures, and speech. Another well-known dataset is EmoReact [17] in particular. The EmoReact dataset was introduced, comprising multimodal data from children aged 4 to 14. This dataset comprises 1,102 videos noted with 17 distinct emotional states, including six basic emotions (happiness, sadness, surprise, fear, disgust, and anger), a neutral state, valence, and nine complex emotions such as curiosity, uncertainty, excitement, attentiveness, exploration, confusion, anxiety, embarrassment, and frustration.

In the proposed model, the EmoReact dataset was employed to recognize emotions in children. Although the dataset contains 17 emotion categories, our work focuses on a subset of six basic emotions: anger, disgust, fear, happiness, sadness, and surprise. To represent gesture-based temporal data, 30 video frames were extracted from each video clip using a frame sampling technique, where videos were processed frame by frame via OpenCV. For facial expression analysis, individual facial images were also extracted from these videos and resized to a uniform resolution of 40×40 pixels. This dual-modality approach captures both temporal dynamics through gesture sequences and spatial features from facial expressions, enabling a more comprehensive and accurate recognition of children's emotions. By combining these modalities, the model effectively leverages complementary emotional cues, improving its ability to detect subtle and dynamic emotional behaviours in children.

3.2 Proposed Model Architecture

This model is designed to take two types of inputs: images for facial expressions and sequences for gestures, which it then feeds into the network with convolutional layers for images and LSTM layers for sequences to extract relevant features and then classify these input pairs into one of the six emotion classes. The model is optimized using the Adam optimizer and a loss function, categorical cross-entropy, appropriate for multi-class classification problems. Accuracy is applied as the measure of the performance of the model. As CNN operations are parallelizable in GPUs and LSTM sequential operations have greater complexity dominance, the LSTM layer becomes the bottleneck. But the combination of CNN and LSTM maximizes spatial-temporal feature extraction with high accuracy at a reasonable computational cost. Therefore, the model is compatible with real-time applications with just the right balance of accuracy and efficiency on multimodal emotion recognition. Figure 1 illustrates our proposed model with its key components and processing steps in a flowchart.

3.2.1 Components of the Model

The model thus consists of several components, starting with the input of multimodal data that encompasses facial expressions as well as gestures. The features shall be extracted with the help of deep learning techniques via the CNN for extracting spatial features of facial expressions and LSTM networks for the capture of temporal dynamics of gestures. The features so extracted are combined to create an overall representation for the purpose of classifying the emotional state within the final output. It therefore combines the spatial and temporal information from the input data using both CNN and LSTM techniques to robustly recognize the emotions. Components of the Model this section elaborates on each of the model's components.

3.2.1.1 Inputs of Facial Expression

The input is an image of size 48x48 pixels. A CNN extracts spatial features from frames of facial images.

3.2.1.2 Inputs of Gesture

The input is a sequence of 30-time steps, and each time step has 50 features. Sequences of skeletal or body movement data derived from videos are used for the extraction of temporal features with an LSTM network.

3.2.1.3 Facial Feature Extraction

The CNN aspect of the architecture has a hierarchical feature extraction using convolutional layers. Early layers tend to function well for edges and corners; the deeper, more refined network composition progresses further towards the eyes, mouth, and nose. In terms of spatiality, reductions are made using pooling layers that abstract the features into more complex representations. It encodes the entire image in a compact form, giving more importance to facial expressions, and then fuses it with features corresponding to gesture for better accuracy in the recognition of emotion.

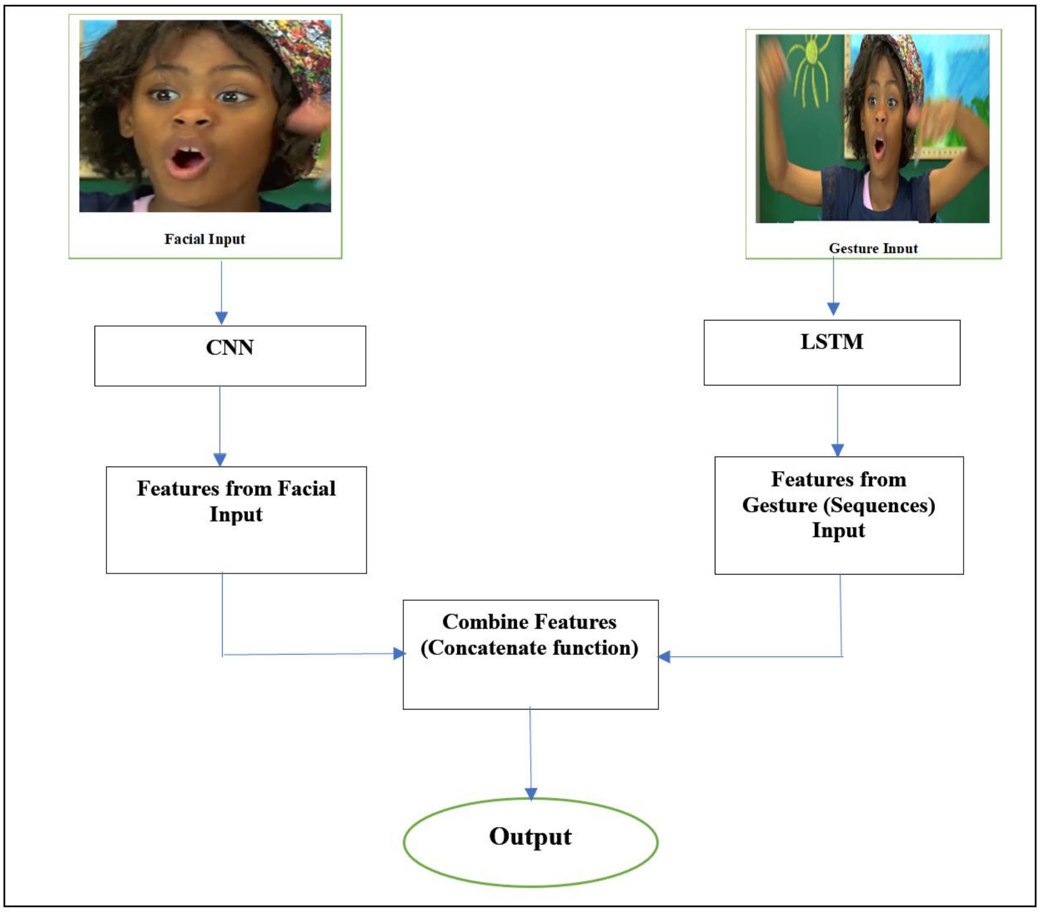

Facial Input
Gesture Input
CNN
LSTM
Features from Facial Input
Features from Gesture (Sequences) Input
Combine Features (Concatenate function)
Output

Fig. 1 Flowchart of our Proposed Model

3.2.1.4 Gesture Feature Extraction

Gesture input processes using a Long Short-Term Memory (LSTM) network. LSTM can learn patterns in sequential data, such as time-series gestures, by emphasizing temporal dependencies. Imagine a video of a person waving their arm; now each frame corresponds to the position of this arm at successive time intervals. The LSTM processes that sequence of time and determines how the arm position changes from one frame to another, thus learning the dynamics of the motion. At the end of that sequence, the LSTM captures the temporal dependencies and makes an interpretation of the movement being a wave gesture. Then, this representation is further polished by the dense layer into a relatively compact high-level feature vector that captures the crux of the gesture for further use in classification or recognition tasks. The point is underlined that LSTMs are effective in modeling temporal patterns in gesture recognition. Here is a step-by-step explanation of how gesture detection and feature extraction work: Gesture detection and feature extraction are sequential processes wherein we take every time step that encodes 50 features, such as hand positions, joint angles, or velocities that represent the state of the gesture. An LSTM processes these sequences and captures the temporal dependencies by learning motion patterns like the gradual upward movement in a rising hand gesture. The final output of the LSTM is a vector summarizing the overall motion and can be seen as a unique "fingerprint" of the gesture. A dense layer further refines this representation, creating a higher-level abstraction for robust gesture recognition.

3.2.1.5 Feature Fusion from both Modalities

In our approach, we apply feature-level fusion to improve emotion recognition by combining embeddings from both facial expression and gesture modalities. Outputs from the dense layer of the facial expression model are concatenated with outputs from the dense layer of the LSTM component in the gesture model. This fusion will result in a single feature vector that summarizes the information from both inputs. These features can then be integrated into the model so that better predictions may be made about a given period of time by accounting for both facial and gestural cues of emotions. A dense layer is applied with six output neurons representing each unique class of emotions. For instance, a smiling face with a thumbs-up sign probably means happiness. The SoftMax function is utilized to assign a probability to every one of the six classes. This is interpreted as the probability that each emotion would be expressed. Therefore, this fusion would improve the accuracy and robustness of emotion classification because it would take advantage of the complementary strengths by tapping into facial and gesture-based inputs.

3.2.1.6 Final Output

This last dense layer has six neurons and uses SoftMax activation. There are six basic emotions therefore, each neuron is associated with one of these classes, and SoftMax activation ensures that outputs can be viewed as probabilities.

3.3 Training and Evaluation

For the training of the proposed model, the categorical cross-entropy loss function was utilized, as it is particularly appropriate for multi-class classification problems like multimodal emotion recognition. The model is optimized using the Adam optimizer, which supports efficient convergence with adaptive learning rates. The training was performed for 50 epochs, enabling the model to repeatedly learn from the data and optimize its parameters.

For measuring performance, accuracy was utilized as the main measure for both training and testing. Apart from overall accuracy, the performance of the model was also measured using critical evaluation measures such as precision, recall, and F1-score for every one of the six primary emotions: Anger, Disgust, Fear, Happiness, Sadness, and Surprise. A detailed analysis using confusion matrix for class-wise performance and misclassification trends was also done. Also, loss values were tracked to approximate prediction error and facilitate sufficient convergence of the model. These thorough assessment tactics validate the model's well-balanced performance as well as its capacity for generalization on diverse emotional categories.

4 Results and Discussion

The table 1 shows the per-class performance metrics —precision, recall, and F1-score for the six primary emotion classes: anger, disgust, fear, happiness, sadness, and surprise. The suggested model shows robust and stable performance for most emotion classes. Emotions like anger, disgust, and surprise recorded high F1 scores, signifying the model's capacity to accurately label both common and uncommon emotional displays. Notably, happiness had the best recall, indicating the model's superior sensitivity to positive affect. Whereas Fear has the lowest precision and F1-score. These findings verify that the proposed multimodal CNN–LSTM approach is able to capture emotional differences in children with balanced precision and recall, providing strong generalization across emotion categories.

Table 1: Evaluation Metrics for Emotion Classification across Six Emotions

Emotions	Precision	Recall	F1-Score
Anger	0.90	0.86	0.88
Disgust	0.89	0.85	0.87
Fear	0.76	0.80	0.78
Happiness	0.82	0.98	0.89
Sadness	0.87	0.81	0.84
Surprise	0.81	0.95	0.88
Total Accuracy	---	---	87.5%
Average	0.842	0.875	0.873

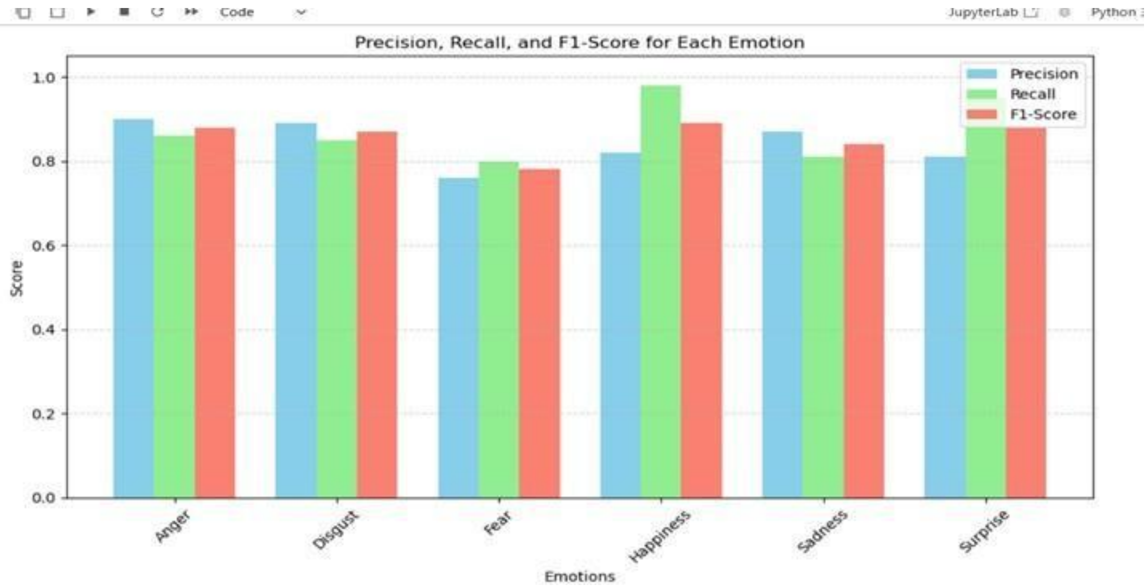
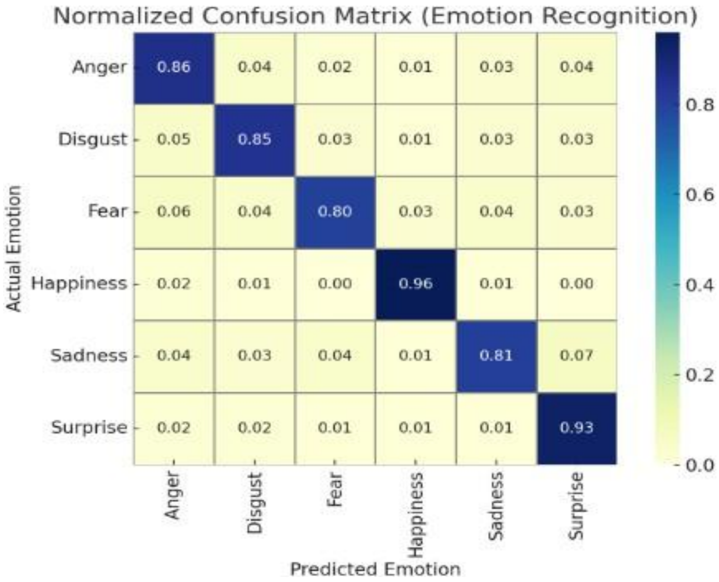


Fig. 2 Visualization of Classification Performance Metrics for Each Emotion Class



	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	0.86	0.04	0.02	0.01	0.03	0.04
Disgust	0.05	0.85	0.03	0.01	0.03	0.03
Fear	0.06	0.04	0.80	0.03	0.04	0.03
Happiness	0.02	0.01	0.00	0.96	0.01	0.00
Sadness	0.04	0.03	0.04	0.01	0.81	0.07
Surprise	0.02	0.02	0.01	0.01	0.01	0.93

Fig. 3 Normalized Confusion Matrix for Multimodal Emotion Recognition

The performance of the proposed emotion recognition model was tested based on a normalized confusion matrix, as presented in Figure 3. The matrix displays the classification performance on six basic emotions. The model shows good classification accuracy for happiness and surprise. Anger and disgust are also identified well, showing the capability of the model in classifying negative emotional expressions.

Nonetheless, the emotion Fear has a somewhat lower true positive score of 0.80, with significant confusions with anger (6%) and disgust (4%). Likewise, sadness has a true positive score of 0.81, with some classifications into surprise (7%) and fear (4%). The confusions can be traced to similarity in facial expressions or upper body movements among these emotion classes. In general, the confusion matrix verifies that the model generalizes well for the majority of emotion classes, particularly performing best in identifying separate emotions such as happiness and surprise.

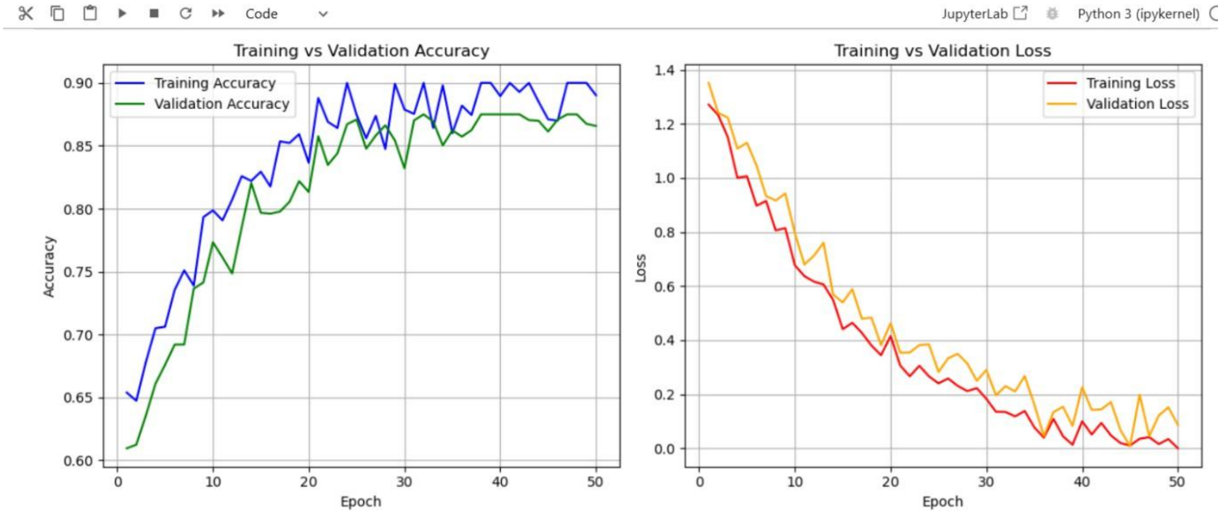


Fig. 4: Training vs. Validation Accuracy and Loss Graphs over Epochs

In this experiment, 75% of the data was used for training and 25% for validation. The training process of the suggested emotion recognition model was tracked for 50 epochs, and the accuracy and loss trend are shown in Figure. As evident from the Training vs. Validation Accuracy plot, training accuracy and validation accuracy both grew consistently throughout early epochs, with the model learning a training accuracy of about 90% and a validation accuracy of 87.5% for the last epochs. This steady growth implies that the model learns discriminative features from the dataset successfully without serious overfitting. This conclusion is further reaffirmed by the training vs. validation loss plot.

Both training and validation loss decreased sharply in the early epochs and continued to decline gradually, stabilizing around 0.1 for training and 0.15–0.2 for validation. The small and consistent gap between training and validation curves suggests strong generalization performance of the model. Minor oscillations in the validation loss after epoch 35 can be attributed to differences in sample complexity, but no meaningful divergence occurs. Overall, these curves demonstrate that the model converges well, with neither overfitting nor underfitting, and generalizes well to unseen data during validation.

5 Comparison with Background Models

The table 2 provides an overview of recent research on emotion recognition in kids based on different datasets, age ranges, and modalities. The majority of the current methods, including those by Lopez-Rincon et al. (2019), Suhan et al. (2022), and Manish Rathod et al. (2022), make use of CNN alone and mostly take facial expressions as the input modality. Though some of the models, including Suhan et al., have included speech and gesture, they all employ single-stage CNNs that do not involve temporal modelling. Unlike these works, our suggested model employs a hybrid deep learning strategy involving CNN and Long Short-Term Memory (LSTM) networks for the concurrent spatial and temporal feature extraction. Additionally, our model uniquely combines upper body gestures and facial expressions, offering a multimodal insight towards emotion identification. With the EmoReact dataset of children between 4 to 14 years old, the suggested model had competitive accuracy at 87.5%. Our method introduces new architecture and multimodal input fusion, providing richer and more context-dependent emotion recognition infrastructure, particularly useful in dynamic child interaction environments.

Table 2: Comparison between the accuracy results, methods, modalities, age group, and data sets for different systems with our proposed model

Sr. No.	Authors	Method	Data Set	Children's Age Group	Modalities	Accuracy (%)

1	Lopez-Rincon et al. (2019)	CNN	CAFE	2 to 8 years	Facial Expression	44.89
2	**Suhan et al. (2022)**	**CNN**	***EmoReact***	***4 to 14 years***	***Speech, Upper body gesture***	***90***
3	Manish Rathod et. al. (2022)	***CNN***	LIRIS	*7 to 10 years*	*Facial Expression*	*89.31*
4			Manish Rathod's	*7 to 10 years*	*Facial Expression*	*90.98*
5	Pandyan et al. (2023)	***CNN***	City Infants Faces Database	*4 to 6 months*	*Facial Expression*	*97.8*
6	***Our Model***	***CNN and LSTM***	***EmoReact***	***4 to 14 years***	***Upper body gesture and facial expression***	***87.5***

6 Conclusion

In this work, we introduced a hybrid CNN–LSTM deep learning architecture for recognizing emotion in children through the fusion of facial expressions and upper body gestures. In contrast to earlier efforts that adopted mainly CNN-based models with single-mode inputs, our solution captures spatial and temporal information using multimodal fusion, accessing efficiently the intricate and delicate emotional signals that are expressed in children between 4 and 14 years. Based on the EmoReact dataset, the model obtained competitive accuracy of 87.5% with its capabilities across six basic emotions. Performance assessment based on precision, recall, F1-score, and the confusion matrix proved the model's excellent generalization and trustworthy recognition for both positive and negative emotional states. Additionally, training and validation trends authenticated constant learning without overfitting, proving the capability of the model in real-world scenarios. The novelty of our contribution is its dual-modality input handling and temporal modeling, which are especially useful in dynamic child interaction settings like educational systems, child healthcare, and affective computing. In being uniquely focused on child emotion recognition—a somewhat understudied domain—this work adds to human-computer interaction research a child-sensitive and context-aware framework. Future research can investigate incorporating other modalities like speech, full-body gesture or posture, and physiological signals; age or developmental stage-based personalization; and real-time implementation on embedded systems for interactive learning or therapy.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability Statement: The data used in this research are available with the authors. Permission to use the dataset was obtained from its owner via email, and the dataset was provided to the authors through email communication. The dataset is publicly available, and anyone interested in using it can obtain access by requesting permission from the dataset owner.

Acknowledgement: I would like to express my sincere gratitude to B. Nojavanasghari, T. Baltrušaitis, C. Hughes, and L.-P. Morency for providing me with the EmoReact dataset for my research. Their contribution has been invaluable in facilitating this study. Thank you very much for your support and assistance. I also extend my heartfelt gratitude to all my co-authors for their invaluable help, guidance, and support throughout this work. Their contributions have been essential in shaping this study.

References

[1] Alhussein, M. (2016). Automatic facial emotion recognition using the weber local descriptor for the e-Healthcare system. *Cluster Computing*, *19*, 99-108. https://doi.org/10.1007/s10586-016-0535-3

[2] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. (2021). Review of deep learning:

concepts, CNN architectures, challenges, applications, and future directions. *Journal of Big Data*, *8*, 1-74. https://doi.org/10.1186/s40537-021-00444-8

[3] Barros, P., Jirak, D., Weber, C., & Wermter, S. (2015). Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks*, *72*, 140-151. https://doi.org/10.1016/j.neunet.2015.09.009

[4] Bartlett, M. S., Littlewort, G., Lainscsek, C., Fasel, I., & Movellan, J. (2004, October). Machine learning methods for fully automatic recognition of facial expressions and facial actions. In the *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)* (Vol. 1, pp. 592-597). IEEE. **DOI:** 10.1109/ICSMC.2004.1398364

[5] Chen, S., Tian, Y., Liu, Q., & Metaxas, D. N. (2013). Recognizing expressions from face and body gestures by temporal normalized motion and appearance features. *Image and Vision Computing*, *31*(2), 175-185. https://doi.org/10.1016/j.imavis.2012.06.014

[6] Gunes, H., & Piccardi, M. (2005, August). Fusing face and body gestures for machine recognition of emotions. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* (pp. 306-311). IEEE. **DOI:** 10.1109/ROMAN.2005.1513796

[7] Gunes, H., & Piccardi, M. (2007). Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, *30*(4), 1334-1345. https://doi.org/10.1016/j.jnca.2006.09.007

[8] Gunes, H., & Piccardi, M. (2008). Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(1), 64-84. **DOI:** 10.1109/TSMCB.2008.927269

[9] Hudlicka, E. (2003). To feel or not to feel: The role of affect in human–computer interaction. *International journal of human-computer studies*, *59*(1-2), 1-32. https://doi.org/10.1016/S1071-5819(03)00047-8

[10] Ilyas, C. M. A., Nunes, R., Nasrollahi, K., Rehm, M., & Moeslund, T. B. (2021, February). Deep Emotion Recognition through Upper Body Movements and Facial Expression. In *VISIGRAPP (5: VISAPP)* (pp. 669-679). DOI: 10.5220/0010359506690679

[11] Karatay, B., Bestepe, D., Sailunaz, K., Ozyer, T., & Alhajj, R. (2022, March). A multi-modal emotion recognition system based on the CNN-transformer deep learning technique. In *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)* (pp. 145-150). IEEE. **DOI:** 10.1109/CDMA54072.2022.00029

[12] Keshari, T., & Palaniswamy, S. (2019, July). Emotion recognition using feature-level fusion of facial expressions and body gestures. In *2019 International conference on Communication and Electronics Systems (ICCES)* (pp. 1184-1189). IEEE. **DOI:** 10.1109/ICCES45898.2019.9002175

[13] Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, *18*(2), 401. **https://doi.org/10.3390/s18020401**

[14] Llurba, C., & Palau, R. (2024). Real-Time Emotion Recognition for Improving the Teaching–Learning Process: A Scoping Review. *Journal of Imaging*, *10*(12), 313. https://doi.org/10.3390/jimaging10120313

[15] Mathew, A. R., Al Hajj, A., & Al Abri, A. (2011, June). Human-computer interaction (HCI): An overview. In *2011 IEEE International Conference on Computer Science and Automation Engineering* (Vol. 1, pp. 99-100). IEEE. **DOI:** 10.1109/CSAE.2011.5953178

[16] Newen, A., Welpinghus, A., & Juckel, G. (2015). Emotion recognition as pattern recognition: the relevance of perception. *Mind & Language*, *30*(2), 187-208. **https://doi.org/10.1111/mila.12077**

[17] Nojavanasghari, B., Baltrušaitis, T., Hughes, C. E., & Morency, L. P. (2016, October). Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th ACM International Conference on Multimodal*

Interaction (pp. 137-144). https://doi.org/10.1145/2993148.2993168

[18] Nunes, A. R. V. (2019). *Deep emotion recognition through upper body movements and facial expression* (Doctoral dissertation, Master's Thesis, Aalborg University). https://projekter.aau.dk/projekter/files/307194482/Thesis_Rita_Nunes.pdf

[19] Rathod, M., Dalvi, C., Kaur, K., Patil, S., Gite, S., Kamat, P., ... & Gabralla, L. A. (2022). Kids' emotion recognition using various deep-learning models with explainable AI. *Sensors*, *22*(20), 8066. **https://doi.org/10.3390/s22208066**

[20] Verma, B., & Choudhary, A. (2021). Affective state recognition from hand gestures and facial expressions using Grassmann manifolds. *Multimedia Tools and Applications*, *80*(9), 14019-14040. https://doi.org/10.1007/s11042-020-10341-6

[21] Wei, J., Hu, G., Yang, X., Luu, A. T., & Dong, Y. (2024). Learning facial expression and body gesture visual information for video emotion recognition. *Expert Systems with Applications*, *237*, 121419. https://doi.org/10.1016/j.eswa.2023.121419

[22] Yang, R., Singh, S. K., Tavakkoli, M., Amiri, N., Yang, Y., Karami, M. A., & Rai, R. (2020). CNN-LSTM deep learning architecture for computer vision-based modal frequency detection. *Mechanical Systems and Signal Processing*, *144*, 106885. https://doi.org/10.1016/j.ymssp.2020.106885

[23] Lopez-Rincon, A. (2019, February). Emotion recognition using facial expressions in children using the NAO Robot. In *2019 international conference on electronics, communications, and computers (CONIELECOMP)* (pp. 146-153). IEEE, **DOI:** 10.1109/CONIELECOMP.2019.8673111

[24] Filntisis, P. P., Efthymiou, N., Koutras, P., Potamianos, G., & Maragos, P. (2019). Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction. IEEE Robotics and Automation letters, 4(4), 4011-4018, **DOI:** 10.1109/LRA.2019.2930434

[25] Filntisis, P. P., Efthymiou, N., Potamianos, G., & Maragos, P. (2021, August). An Audiovisual Child Emotion Recognition System for Child-Robot Interaction Applications. In *2021 29th European Signal Processing Conference (EUSIPCO)* (pp. 791-795). IEEE, **DOI:** 10.23919/EUSIPCO54536.2021.9616106

[26] Suhan, S., Kalaichelvan, K., Samarage, L., Alahakoon, D., Samarasinghe, P., & Nadeeshani, M. (2022, November). Automated Evaluation of Child Emotion Expression and Recognition Abilities. In *2022 International Conference on Information Technology Systems and Innovation (ICITSI)* (pp. 388-393). IEEE, **DOI:** 10.1109/ICITSI56531.2022.9970990

[27] Rathod, M., Dalvi, C., Kaur, K., Patil, S., Gite, S., Kamat, P., ... & Gabralla, L. A. (2022). Kids' emotion recognition using various deep-learning models with explainable AI. *Sensors*, *22*(20), 8066, **https://doi.org/10.3390/s22208066**

[28] Pandyan, U. M., Sindha, M. M. R., Kannapiran, P., Marimuthu, S., & Anbunathan, V. (2023). Application of Machine and Deep Learning Techniques to Facial Emotion Recognition in Infants. In *Emotion Recognition-Recent Advances, New Perspectives, and Applications*, https://www.intechopen.com/chapters/85877