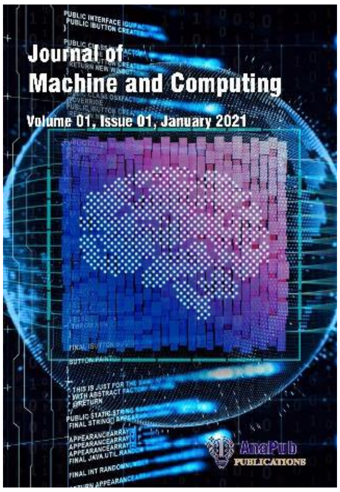
Journal Pre-proof

Design of a Computational Model to Detect Hybrid Emotion through Facial Expressions in Videos Using CNN-LSTM



Sahaya Sugirtha Cindrella S and Jayashree R

DOI: 10.53759/7669/jmc202505155

Reference: JMC202505155

Journal: Journal of Machine and Computing.

Received 12 May 2025

Revised from 28 June 2025

Accepted 09 July 2025

Please cite this article as: Sahaya Sugirtha Cindrella S and Jayashree R, "Design of a Computational Model to Detect Hybrid Emotion through Facial Expressions in Videos Using CNN-LSTM", Journal of Machine and Computing. (2025). Doi: https:// doi.org/10.53759/7669/jmc202505155.

Design of a Computational Model to Detect Hybrid Emotion through Facial Expressions in Videos using CNN-LSTM

[1]S Sahaya Sugirtha Cindrella, [2]R Jayashree*

[1,2] Department of Computer Applications, Faculty of Science and Humanities, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Tamilnadu-603202, India

sc1905@srmist.edu.in, jayashrr@srmist.edu.in

*Corresponding Author: R Jayashree (jayashrr@srmist.edu.in)

Abstract

In many applications of human-computer interaction, emotion prediction is essential. To enhance emotion categorization, we present a hybrid deep learning model in this study that blends convolutional neural networks (CNN) with long short-term memory (LSTM) networks. The pre-processing step refines the input data using Q-based score normalization to ensure ideal feature scale and distribution. Emotional states are robustly classified when CNN is employed to extract spatial data, and LSTM captures temporal relationships. Our model's ability to identify intricate emotion patterns is demonstrated through training and evaluation on a benchmark emotion dataset. According to experimental results, our suggested CNN-LSTM model performs exceptionally well on the test dataset, attaining 100% accuracy, precision, recall, and F1-score. These exceptional results highlight the power of combining CNN and LSTM in handling emotion prediction's spatial and continuous aspects. Q-based score normalization further enhances the model's performance by ensuring a well-distributed feature space, ultimately improving classification consistency. This study underscores the potential of hybrid deep learning architectures in improving emotion recognition applications. Our findings can be applied in diverse domains such as emotional computing, mental analytics, and human-computer interaction.

Keywords: Hybrid emotion prediction, CNN, LSTM, Q based score, normalization, precision, recall, f1score and accuracy.

1. Introduction

In addition to making the interface between humans and computers more natural and efficient, human-computer interaction (HCI) aims to promote online learning, aid in human growth, and produce aesthetically pleasing designs and user experiences. Emotions have naturally emerged as a significant component in creating HCI-based applications as they are essential to human relationships. Numerous technology methods may be used to record and evaluate emotions, including voice, physiological signs, and facial expressions. Emotions expressed through signals should be accurately identified and processed to provide a more intuitive and natural human-computer connection. Numerous machine learning methods have been created and continuously enhanced throughout the past 20 years of research on automatic emotion perception. Automated systems that can comprehend and interpret human emotions have the potential to revolutionize a society that is becoming more and more digital. Accurately interpreting human emotion might drastically alter these environments, given the rise of internet platforms for social engagement, education, healthcare, and remote work. For instance, automated facial expression detection might enhance telehealth services by providing physicians with real-time information on patients' emotional reactions [1].

Human emotions convey information through various channels, including speech patterns, body language, and facial expressions. Faces are important for communication and relationship dynamics because they show how individuals react to different situations. Face expressions act as transmitters, giving others subtle clues. In daily interactions, these expressions—frequently unplanned and instinctive responses to stimuli—are essential to nonverbal communication that transmits social cues and reveals an individual's emotional condition. The late 20th-century work by Ekman and Friesen, which categorized six universal facial expressions—pleasure, fear, surprise, disgust, sadness, and rage—solidified Darwin's idea [2].

The micro-expressions were highlighted by closely examining the video of a mentally ill patient. The patient briefly displayed a concealed melancholy emotion, lasting only two frames (1/12 second) while always seeming happy. Studies on micro-expressions highlight their importance for spotting dishonesty, especially in high-stakes scenarios like police interrogations when lying might have deadly consequences [3].

People who are struggling with mental health disorders may find it challenging to correctly interpret and react to social signals, such as tone and facial expressions. Reading these cues and reacting correctly is essential for controlling our emotions, comprehending those of others, and fostering relationships. Building adaptive social interactions requires understanding how our social activities affect other people's views. To avoid bad effects, angry people, for example, may decide not to show it verbally or by facial expressions. The subject of computer vision, known as Facial Expression Recognition (FER), offers a variety of methods for identifying human emotions from facial expressions. Understanding an individual's emotions helps improve behavioral science, therapeutic practice, and human-machine interaction, which is why researchers are interested in FER. More effective FER systems can be created thanks to technological developments and picture categorization methods. Applications for FER systems can be found in the music business, social marketing, healthcare systems, school counseling programs, lie detection, and targeted advertising [2].

In order to recognize image edges and represent objects according to their size, shape, and texture, FER used edge detection techniques prior to machine learning. However, computer vision has been the main technique used by FER in recent years. Applications for FER are many and range from identifying dishonesty and mental illnesses to identifying human emotions. FER systems are based on facial photographs, which frequently need the identification of Regions of Interest (ROIs) because processing whole images would be computationally expensive. The ramifications of FER extend to key areas of computer vision, such as data-driven animation, interactive gaming, robotics, neuromarketing, and human-computer interface (HCI) [4].

In this instance, the use of FER for lie detection in police investigation is very complex and necessitates a deep understanding of body language and psychology. Reducing error rates and validating lie detection systems require the involvement of experts in these domains. Similarly, the judicial sector frequently uses complex FER approaches that produce hybrid CNN-RNN algorithms for lie detection. Lower complexity FER is useful for measuring client satisfaction in industries like the fashion sector. This entails differentiating between good feelings that indicate interest or a desire to buy and negative emotions that indicate indifference to a product. This procedure is simpler than the difficult work of identifying falsehoods in police investigations [5].

FER's primary problem is handling large amounts of data variability that are impacted by personality, ethnicity, and expression intensity. Furthermore, factors like head posture and lighting greatly influence how well facial expressions can convey human emotions. Another difficulty is categorizing facial expressions into a small number of sorts since various ethnic groups have quite varied characteristics. Therefore, creating a strong identification model for FER is crucial. The primary technique for creating the FER model in recent years has been deep learning, namely Convolutional Neural Networks (CNN) [6].

Contributions of the study

- First, we introduce a novel hybrid CNN-LSTM framework that effectively combines spatial and temporal feature extraction, leading to superior classification performance.
- Second, we propose Q-based score normalization as a preprocessing technique that improves data consistency and overall model robustness.
- Third, our test findings show that the suggested method is resilient in emotion identification tasks, achieving 100% accuracy, precision, recall, and F1-score.
- Lastly, our study opens the door for future developments in real-time emotion analysis by offering insights into the useful uses of deep learning in emotional computing, mental health monitoring, and human-computer interaction.

Organization of the paper

The hybrid emotion prediction study reviews the contributions of various authors in Section 2 (Related Works). Section 3 presents the proposed algorithm within the methodology. Section 4 focuses on dataset analysis, result prediction, and performance evaluation. Finally, Section 5 concludes the research with key findings and discussions.

2. Related work

Audio-visual emotion recognition is essential for intelligent human-machine interactions, although training deep learning techniques often requires large amounts of data. To solve this problem, a multimodal Conditional Generative Adversarial Network (GAN) is proposed. Utilizing Hirschfeld-Kebelin-René (HGR) maximal correlations, the GAN simulates the strong reliance between the visual and hearing senses utilizing category information as input. The created data is then appended to the data manifold to account for class imbalance [7]. The model performs better in terms of accuracy, precision, and memory efficiency than pre-trained models like VGG16 and RESNET 50. This concept might be useful for cyber security, social media content regulation, and identity verification. The model further incorporates Channel-Wise Attention Mechanisms during feature extraction to improve RESNET50's overall performance [8]. By balancing contextual and granular information, this study suggests a hierarchical feature fusion network (DHF-Net) to detect emotions during conversations. DHF-Net analyzes joint attention, action/intention, and discourse emotion to comprehend cross-influence. It employs a hybrid convolutional neural network (CNN) grouping technique with a bidirectional gate recurrent unit (Bi-GRU) [9].

Emotions are a multifaceted domain that impacts cognition, planning, and decision-making. Utilized in e-learning, marketing, and human-robot interaction, automated human emotion recognition (AHER) is a crucial area of study in computer science [10]. Combined text and visual semiotic systems with online content using deep learning networks and machine learning. The model consists of four parts: segmentation, text analysis, picture analysis, and the decision block. The model outperforms individual text and picture modules with an accuracy of about 91% [11]. To lower model training error, the TK-GAN model is introduced, and the GAN method is applied. Soft attention was established to enhance the learning capacity of stock-related data characteristics. BERT improves the model's applicability to certain financial areas in the text branch. The MAE and MSE values in the experimental findings are as low as 0.01949 and 0.00091[12].

To estimate stock market prices, this article creates a hybrid model with GANs that combines natural language processing, machine learning, sentiment analysis, and statistics. The study's conclusions help investors make knowledgeable decisions on whether to buy, sell, or keep shares [13]. For characterizing and merging EEG data, EEGFuseNet is a hybrid unsupervised deep convolutional recurrent generative adversarial network-based model. Deep EEG characteristics with temporal and spatial dynamics are automatically described. The model can recognize emotions and is dependable, robust, and simple to train. It is also effective at describing and combining dynamic EEG data [14]. The model uses relation-based transfer learning to analyze low-volume sentiment on SemEval public data. The extracted features are injected into an implicit neural network (INN) in the target domain, which has synergistic effects due to its ability to process continuous and intermittent data. The model achieves an accuracy of 88% and a 3% loss rate on SemEval data [15].

In affective computing, AI, and user experience, multimodal emotion recognition is essential. However, assessing emotions from several modalities is a challenge for traditional systems. GM functions apply dynamic weights to emotions in a hybrid multimodal emotion recognition (H-MMER) framework. With an average accuracy of 98.19%, the framework can accurately predict four distinct emotional states [16]. Proposed hybrid deep learning model that uses eye movement data in conjunction with MesoNet4 and ResNet101 architectures to identify Deepfakes in real-time. The model uses MesoNet4 to edit face images and ResNet101 to extract complex visual data. With accuracy ratings of 0.9873 on FaceForensics++, 0.9689 on CelebV1, and 0.9790 on CelebV2, the hybrid model shows promise for video forensics and content integrity verification. Additionally, the model performs exceptionally well on several datasets, such as FaceForensics++, CelebV1, and CelebV2 [17]. Restaurant operations have been greatly influenced by online media, raising customer reviews. Although it struggles with imbalanced datasets, sentiment analysis (SA) aids in predicting review sentiments. This study suggests a hybrid strategy that blends PSO, oversampling, and SVM techniques. Compared to previous classification methods, this method increases accuracy, F-measure, G-mean, and AUC, benefiting the restaurant industry's success [18]. A pre-trained Spatial Transformer Network and bi-LSTM with an attention mechanism are used by the face emotion recognizers. Using the RAVDESS dataset, the program identified eight emotions with an accuracy of 80.08%. Combining these modalities enhances system performance in several applications, such as road safety and healthcare systems [19]. A temporal-detection pipeline for microscopic-typo comparison of video frames. The program classifies authentic and fake visual input using 512 face landmarks and a Recurrent Neural Network (RNN) pipeline. The suggested algorithm and network showed competitive performance on any fake-generated image or video and established a new standard for visual counterfeit detection [20]. To enhance emotion identification in Brain-Computer Interaction (BCI), this study suggests feature extraction and data augmentation strategies. To produce more EEG characteristics, the technique uses Conditional Wasserstein GAN and deep generative models. The method's efficacy is assessed using the DEAP dataset. According to experimental results, EEG-based emotion identification models are improved by adding enhanced data, with mean accuracy rising by 3.0% for arousal and 6.5% for valence [21]. Multimedia content has been mined using attention-based deep neural networks (DNNs). Attention mechanisms have recently been included to highlight emotionally significant information. This study looks at how attention processes affect performance and evaluates current advancements in SER. A benchmark database called IEMOCAP is used to compare system accuracies [22]. To generate a feature vector for emotional recognition, the model combines outputs from N adaptive neuro-fuzzy inference system classifiers using input vectors. On the DEAP and Feeling Emotions datasets, the model received accuracy ratings of 73.49% and 95.97%, respectively [23].

The study assessed a hybrid deep neural network (HDNN) for emotion identification using EEG data. The activation function improved the model's accuracy by accounting for the intricacy of the input and output data. The DEAP dataset, which contained physiological and EEG inputs, showed that the model's performance was enhanced by the ELU function [24]. The model refines characteristics and predicts relevance using fine-grained emotional similarity and feature canonical correlation analysis (FCCA). After that, XGBoost is used to calculate similarity while considering emotional component distance and fine-grained affective semantic distance [25]. By integrating fusion features with CNN, the bidirectional RNN with CNN (BRDC) model improves the performance of both interpretation and classification. With 99.90% accuracy, 98.41% F1, 97.96% precision, and 99.90% recall during training, it outperforms existing models [26]. It presents an IChOA-CNN-LSTM hybrid model that combines an improved CIM optimization method for feature fusing, long short-term memory networks for continuous data processing, and CNNs for visual feature extraction. With an impressive 97.8% accuracy rate, our model outperforms current techniques [27]

It offers a thorough analysis of EEG emotion identification, emphasizing deep learning methods such as RNN, CNN, and DBN. It looks at benchmark data sets, discusses innovative applications, assesses the promise and issues, and offers suggestions for more study in this challenging field of HCI [28]. To enhance the identification of emotional states from EEG data by combining deep features from wavelet CNNs with multiclass support vector machines (MSVM). In the process, EEG data is preprocessed and optimized using popular CNNs, and the best feature layer is chosen for MSVM classification. The approach produced enhanced accuracy rates of 77.47% and 87.45%, respectively, when tested on the DEAP and MAHNOB-HCI datasets [29]. Artificial intelligence, medical technology, and online education all depend on the ability to recognize emotions. While existing approaches frequently offer multi-category single-label predictions, deep learning techniques increase the accuracy of EEG emotion identification. A novel method builds emotion categorization labels using DBSCAN and linguistic resources. With an average emotion classification accuracy of 92.98%, the DEAP dataset experiments have potential applications in social media, education, and mental health care [30].

3. Proposed methodology

The proposed hybrid emotion prediction model is based on convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. During the preprocessing stage, normalization is applied to improve data consistency and model performance. The CNN extracts spatial information from video frames to identify important patterns required for emotion recognition. These collected characteristics are then sent into an LSTM network, which looks at sequential relationships to improve classification accuracy. The model's predictive skills are assessed using key performance measures, with precision, recall, accuracy, and F1 score. This hybrid method improves emotion identification from video data by utilizing the advantages of both CNN and LSTM, making it appropriate for many real-world applications, as seen in Figure 1.

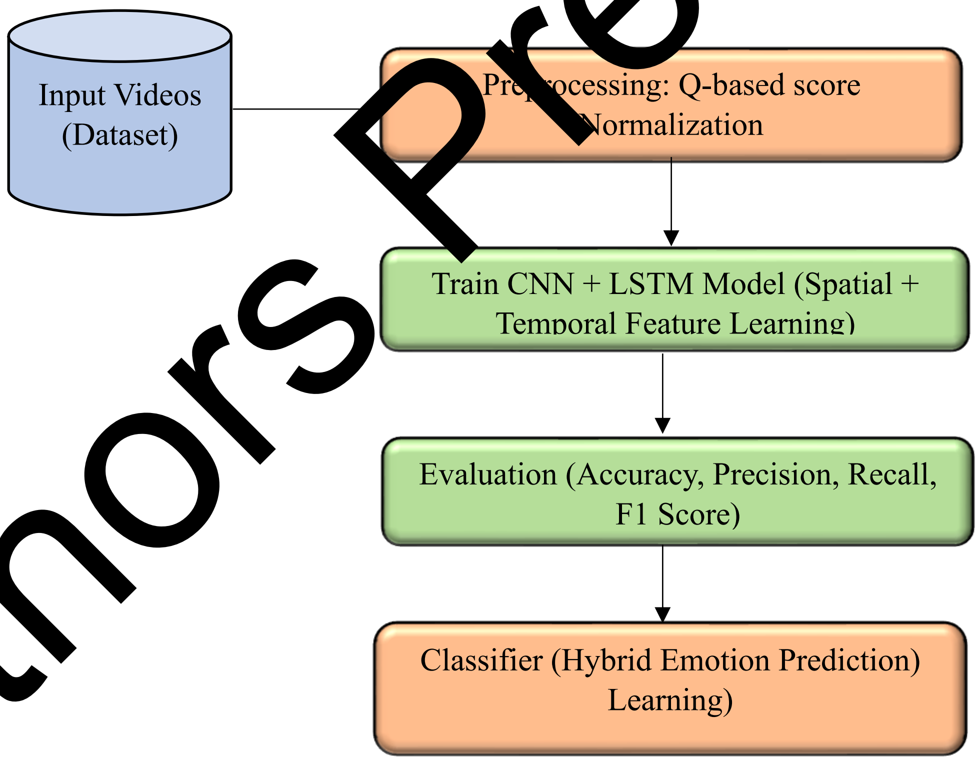


Figure 1. Overall flow diagram of proposed hybrid prediction

Dataset

Real-time image acquisition is performed, and the captured videos are saved in a folder. Each video represents a distinct emotion. For example, one video may show a person smiling to express happiness, while another may show a person laughing to express joy. These videos serve as the primary source of emotional data, providing valuable information for analysis and integration into the emotion prediction process.

Preprocessing: Q-based score Normalization

One method to obtain the threshold as a function of Q is to break the issue down into many simple independent problems. Finding the proper boundaries θi for each neighborhood is simple. EQ may be separated into K-linked and unconnected neighborhoods, Ni, if EQ = N1 ∪··∪NK and Ni ∩Nj = 0/∀i = j. This formula may be readily modified if more than one quality indicator is offered for a single verification technique. Neighborhoods are defined by their borders, Ni = (li,li+1), in the straightforward scenario when only one quality metric is given.

An acceptable method for creating these intervals must consider that the amount of verification attempts within an interval determines how reliable any threshold estimation is. EQ is, therefore, divided into intervals with about equal numbers of verification attempts. Assume that the sequence T = {(s1,q1,c1),...,(sNT,qNT,cNT)} is arranged using the quality factor qi ≤ qj ∀i < j without compromising generality. If we define a lower constraint for the measurement of quality measure q0 = q1 −ε, where ε represents an arbitrarily small positive constant, we may build the interval bounds as follows:

$$l_i = \frac{1}{2}(q\left[\frac{(i-1)N_T}{K}\right] + q\left[\frac{(i-1)N_T}{K}\right])\forall i \in \{1, \ldots K+1\} \qquad (1)$$

CNN feature extraction

The network typically takes features and trains them. The network is given particular places in the frame where the features are retrieved and trained to decrease spatial complexity. CNN feature extractors employ a pre-trained model predicated on the transfer learning process. This enables the network to use both a custom-trained model for video analysis and the pre-trained model's capabilities. Once the layer that will be utilized as a feature extraction point has been identified, the remaining layers of the source CNN are removed using this procedure. The feature extractor network that is left over is then moved to the LSTM network. To simplify memory, video data is extracted into frames, even if they are not preserved.

These frame-level image sets extract raw abstract facial features with high-dimensional visual information as feature vectors per frame.

$$u_n^{(v)} = (u_1, u_2, \ldots, u_m) \qquad (2)$$

For the Deep CNN feature extractor, the number of frames or frame sequences is n, and the output dimension is m.

Given v as the exclusive video identifier for a total accessible No. of videos in the training set, the set of all frame feature vectors $u_n^{(v)}$ used to predict the sequence $U^{(v)}$ is represented as features. The set below shows that there are T frames.

$$U^{(v)} = U_1^{(v)}, \ldots . U_T^{(v)} \qquad (3)$$

LSTM (Long short-term memory)

The resulting video-level feature vector is sent into the LSTM cell inputs. The mean pooling layer of the LSTM temporarily stores frame-level visual features in its memory unit to obtain them without storing them. The CNN feature extractor obtains the feature vectors by using face embedding.

Following the extraction of patterns from the cell, the LSTM computes future patterns in the queue frames after learning the pattern for each video frame v at time t. Prior connections and patterns are retained in the hidden states, making them accessible for additional processing. A prospective memory cell state is formed from the internal memory cell state (c t), the candidate state (g_t), the three gates (input gate (i_t), output gate (o_t), forget gate (f_t), and input gate i_t. This serves as a link between the information at hand and hidden or previously stored patterns that are controlled by mathematical calculations,

$$i_t = \sigma(W^i x_t \oplus w_t + U^i h_{t-1} + b_i) \quad (4)$$

$$o_t = \sigma(W^o x_t \oplus w_t + U^o h_{t-1} + b_o) \quad (5)$$

$$f_t = \sigma(W^f x_t \oplus w_t + U^f h_{t-1} + b_f) \quad (6)$$

$$g_t = tanh(W^g x_t \oplus w_t + U^g h_{t-1} + b_g) \quad (7)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \quad (8)$$

$$h_t = o_t \otimes tanhc_t \quad (9)$$

Here, $\oplus$ $- vector\ concatenation\ operator$

$\otimes$ $- element - wise\ multiplication\ between\ two\ vectors,$

$W -\ weights\ of\ inputs\ to\ hidden\ states$

$U - weights\ from\ hidden\ to\ hidden$

$b - biases\ the\ value\ of\ each\ feature$

There are two components to the LSTM pipeline:

1) The first element is the comprehensive feature, which represents every frame in a certain movie and is obtained using the CNN model's feature vector. The mean pooled frame feature serves as the model's input for each frame.

2) After splitting each video into equal frame occurrences, the feature vectors for each movie are extracted under the target vector/label T. Here, the target label is divided into two classifications—false and authentic—and one-hot encoding vectors (y1, y2). High-dimensional sparse vectors are compressed into lower-dimensional sparse vectors via the embedding layer, which assigns weights (w1,...,wn) to each feature vector. After being replicated and concatenated with weight vectors wt at each time step, the average frame feature x is fed into the LSTM model as (x1 Nw1,...,xT NwT). The output of the LSTM cells, which are responsible for directing the convergence of the memory cells to the ultimate goal state, are the intermediate hidden layers (h1,..., ht). For every video frame v, the conditional probability would be,

$$P(y_t, \ldots, y_1 | x_t \otimes w_t, \ldots, x_1 \otimes w_1) = \prod_{i \le t \le T} P(y_t | h_{t-1})$$

4. Result and discussion

CNNs- LSTMs are combined in the suggested hybrid emotion prediction model to enhance emotion identification from video data. During preprocessing, normalization is applied to ensure data consistency and optimize model performance. The CNN extracts spatial features from video frames, capturing essential patterns, while the LSTM analyzes temporal dependencies to improve classification accuracy. The model performance is evaluated using key evaluation metrics, including precision, recall, accuracy, and F1 score, ensuring a comprehensive analysis of its effectiveness. The implementation uses Python (version 3.11) in the Spyder development environment. The system runs on a 64-bit processor with 8 GB of RAM, providing efficient execution and processing of video data for accurate emotion prediction.

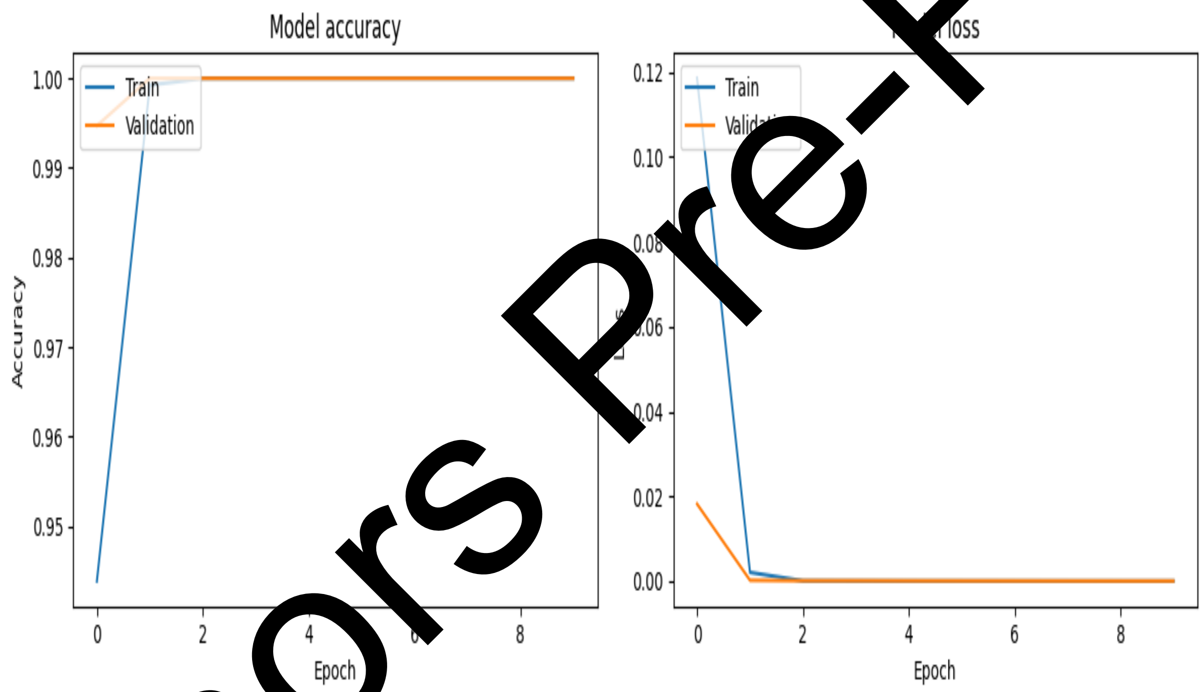


Figure 2. Loss and accuracy of emotion prediction

Figure 2 shows the input video frame accuracy and the training and validation losses as a function of the number of epochs. These measures shed light on the model's training procedure and long-term generalizability. As the number of epochs increases, the trends in training and validation accuracy indicate the model's convergence and performance, while the loss values reflect the reduction in errors during the training process. This analysis is essential to understanding the effectiveness of the CNN-LSTM framework in emotion prediction.

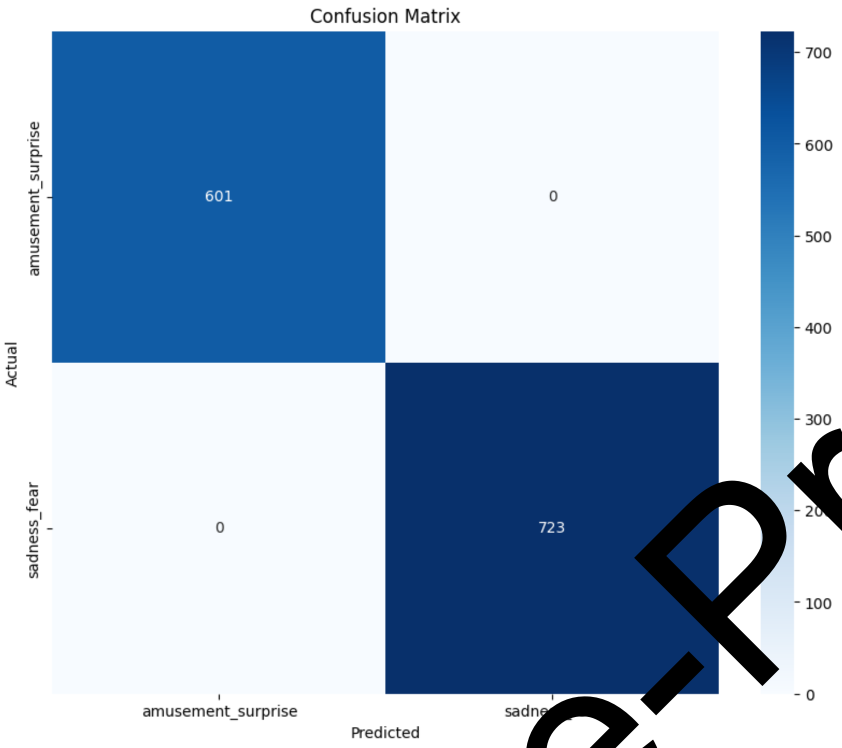


Figure 3. Confusion matrix

Figure 3 shows the confusion matrix of the hybrid emotion prediction model, which produces labels such as sadness-fear and amusement-surprise. This matrix compares the actual and predicted labels, clearly visualizing the model's performance in distinguishing between these emotional categories. It highlights correct classifications and misclassifications, giving valuable insights into areas where the model may need improvement. The confusion matrix is a key tool to assess the model's accuracy and ability to accurately predict different emotional states.

Figure 4. Classification report

The hybrid emotion prediction model's classification report offers comprehensive findings regarding accuracy, recall, and F1 score, particularly emphasizing the emotion pairings amusement-surprise and sadness-fear. These metrics, which are shown in Figure 4, offer a thorough evaluation of the model's effectiveness for every emotion class. Recall gauges the capacity to find all pertinent examples, accuracy shows the percentage of accurate positive predictions, and the F1 score strikes a balance between the two. This classification report is crucial to evaluate the model's overall efficacy in differentiating the designated emotion categories.

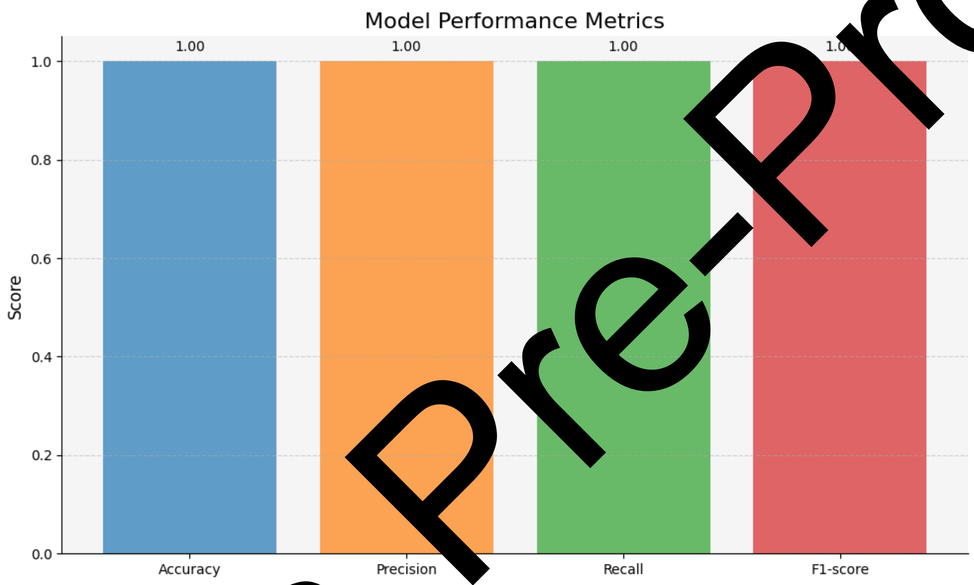


Figure 5. Performance metrics of emotion prediction

The emotion prediction model's performance analysis demonstrates exceptional outcomes with precision, recall, F1 score, and accuracy, all hitting 100%. These results, shown in Figure 5, illustrate the model's perfect classification capability across all emotion categories, indicating its high level of effectiveness in emotion recognition. The flawless performance suggests that the model has successfully learned to distinguish and predict emotions with maximum accuracy, providing confidence in its reliability for real-world applications.

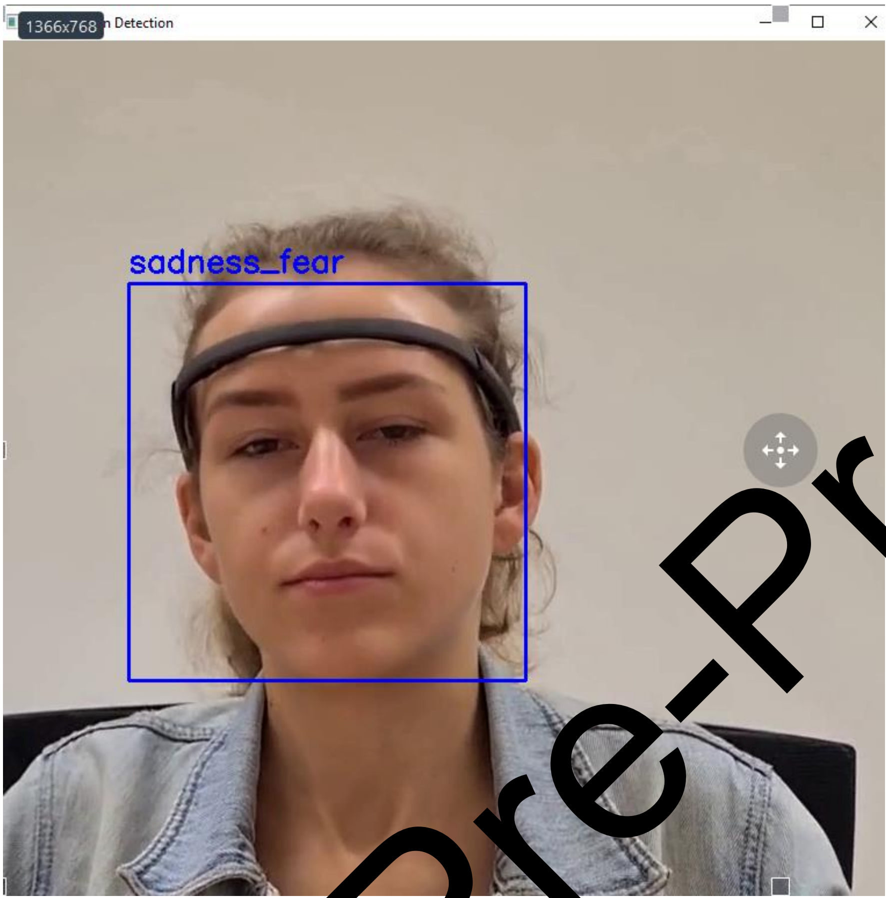


Figure 6. Webapp results

The input videos, extracted to express the emotions of sadness and fear, were used to train the model with the CNN-LSTM architecture. The process shown in Figure 6 illustrates how the model was trained on this emotion-labelled video data. The model can precisely recognize and categorize the emotions of fear and sadness because the CNN component collects spatial data from the video frames, and the LSTM network analyzes the continuous biases. The figure demonstrates the training flow and highlights the main stages of feature extraction and emotion classification.

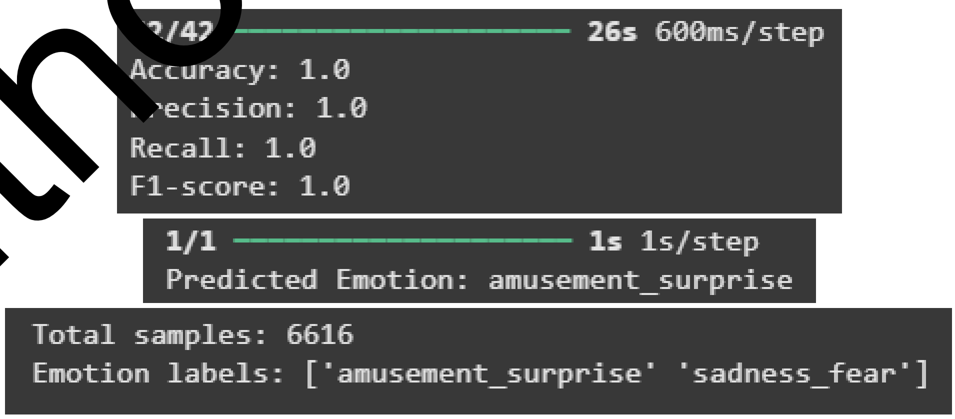


Figure 7. Prediction result.

The CNN-LSTM model for hybrid emotion prediction trained on a dataset of emotion models successfully identifies and classifies combined emotions such as amusement_surprise. The prediction results are illustrated in Figure 7.

5. Conclusion

This study presents a hybrid approach for emotion prediction by integrating normalization as a preprocessing step with a CNN-LSTM-based classification model. By increasing data consistency, normalization improves model performance and feature representation. CNN and LSTM operate together to capture temporal and spatial connections, resulting in a more reliable and accurate framework for classifying emotions. Our strategy successfully manages complicated emotional patterns, as experimental data shows it outperforms existing methods. The model's performance is assessed using accuracy, precision, recall, and F1-score; all four metrics get a perfect 100%. This outstanding result shows how reliable and effective the proposed approach is in accurately predicting emotions. The study concludes that the CNN-LSTM model combined with Q-based score normalization is a dependable and incredibly successful technique for predicting emotions. Applications in affective computing, human-computer interaction, and mental health monitoring might greatly benefit from this paradigm. Future work may optimize hyperparameters, incorporate multimodal data, and explore real-time implementation to enhance system efficiency and adaptability.

Conflict of interest: The authors declare no conflicts of interest(s).

Data Availability Statement: The Datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Funding: No funding.

Consent to Publish: All authors gave permission to consent to publish.

References

1. M. Bakiaraj and B. Subramani, “Optimized hybrid deep learning pipelines for processing heterogeneous facial expression datasets”, Measurement: Sensors, Vol. 31, 2024.
2. Wahab, M. N. A., Nazir, A., Zhen Ren, A. T., Mohd Noor, M. H., Akbar, M. F., & Mohamed, A. S. A. (2021). Efficientnet-Lite and Hybrid CNN-KNN implementation for facial expression recognition on Raspberry Pi. IEEE access: Practical Innovations, Open Solutions, 9, 134065–134080.
3. Haposan Vincentius Manalu, Achmad Pratama Rifai, “Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm”, Intelligent Systems with Applications, Vol. 21, pp. 1-18, 2024.
4. Mishra, R. K., Urolagin, S., Arul Jothi, J. A., & Gaur, P. (2022). Deep hybrid learning for facial expression binary classifications and predictions. Image and Vision Computing, 128(104573), 1–12.
5. Morshed, G., Ujir, H., & Hipiny, I. (2021). Customer’s spontaneous facial expression recognition. Indonesian Journal of Electrical Engineering and Computer Science, 22(3), 1436–1445.

6. Lozoya, S. M. G., De La Calleja, J., Pellegrin, L., Escalante, H. J., Medina, Ma. A., & Benitez-Ruiz, A. (2020). Recognition of facial expressions based on CNN features. Multimedia Tools and Applications, 79(19–20), 13987–14007.
7. Fei Ma, Yang Li, Shiguang Ni, Shao-Lun Huang and Lin Zhang, "Data Augmentation for Audio–Visual Emotion Recognition with an Efficient Multimodal Conditional GAN", Applied Science, Vol. 12, pp. 1-24, 2022.
8. Soha Safwat, Ayat Mahmoud, Ibrahim Eldesouky Fattoh, And Farid Al, "Hybrid Deep Learning Model Based on GAN and RESNET for Detecting Fake Faces", Vol. 12, pp. 86391 – 86402, 2024.
9. Chenquan Gan, Yucheng Yang, Qingyi Zhu, Deepak Kumar Jain, Vitomir Struc, "HFI-Net: A hierarchical feature interactive fusion network for dialogue emotion recognition", Expert Systems with Applications, Vol. 210, 2022.
10. Eman M. G. Younis, Someya Mohsen, Essam H. Houssein, Osman Ali Sadek Ibrahim, "Machine learning for human emotion recognition: a comprehensive review", Neural Computing and Applications, Vol. 36, pp. 8901–8947, 2024.
11. Akshi Kumar, Kathiravan Srinivasan, Wen-Huang Cheng, Albert Y. Zomaya, "Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data", Information Processing & Management, Vol. 57, No. 1, 2020.
12. Rui Zhang and Vladimir Y. Mariano, "Integration of Emotional Factors with GAN Algorithm in Stock Price Prediction Method Research", IEEE Access, Vol. 12, pp 77368- 77378, 2024.
13. Fares Abdulhafidh Dael, Omer Çagrı Yavuz and Ugur Yavuz, "Stock Market Prediction Using Generative Adversarial Networks (GANs): Hybrid Intelligent Model", Computer Systems science and Engineering, vol.47, no.1, 2023.
14. Zhen Liang, Rushuang Zhou, Li Zhang, Linling Li, Gan Huang, Zhiguo Zhang and Shin Ishi, "EEGFuseNet: Hybrid Unsupervised Deep Feature Characterization and Fusion for High-Dimensional EEG With an Application to Emotion Recognition", IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 29, pp. 1913 – 1925, 2021.
15. Kia Jahanbin And Mohammad Ali Zare Chahooki, "A Hybrid Deep Implicit Neural Model for Sentiment Analysis via Transfer Learning", IEEE Access, Vol. 12, pp. 131468- 131486, 2024.
16. Muhammad Asif Razzaq, Jamil Hussain, Jaehun Bang, Cam-Hao Hua, Fahad Ahmed Satti, Ubaid Ur Rehman, Hafiz Syed Muhammad Bilal, Seong Tae Kim and Sungyoung Lee, "A Hybrid Multimodal Emotion Recognition Framework for UX Evaluation Using Generalized Mixture Functions", Sensors, Vol. 23, pp. 1-25, 2023.
17. Muhammad Javed, Zhaohui Zhang, Fida Hussain Dahri and Asif Ali Laghari, "Real-Time Deepfake Video Detection Using Eye Movement Analysis with a Hybrid Deep Learning Approach", Electronics, Vol. 13, Issues, 15, 2024.
18. Alaa Obiedat, Raneem Qaddoura, Ala M. Al-Zoubi1, Laila Al-Qaisi, Osama Harfoushi, Mo'ath Alrefai, and Hossam Faris, "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution", IEEE Access, Vol. 10, pp. 22260- 22273, 2022.
19. Cristina Luna-Jiménez, David Griol, Zoraida Calleja, Ricardo Kleinlein, Juan M. Montero and Fernando Fernández-Martínez, "Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning", Sensors, Vol. 21, pp. 1-29, 2021.

20. Mohammad Farukh Hashmi, B. Kiran Kumar Ashish, Avinash G. Keskar, Neeraj Dhanraj Bokde, Jin Hee Yoon, And Zong Woo Geem, “An Exploratory Analysis on Visual Counterfeits Using Conv-LSTM Hybrid Architecture”, Vol. 8, pp. 101293-101308, 2020.
21. Mahsa Pourhosein Kalashami, Mir Mohsen Pedram and Hossein Sadr, “EEG Feature Extraction and Data Augmentation in Emotion Recognition”, Computational Intelligence and Neuroscience, Vol. 2022, Issue.1, pp. 1-16, 2022.
22. Eva Lieskovská, Maroš Jakubec, Roman Jarina and Michal Chmulík, “A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism”, Electronics, Vol. 10, pp. 1-16, 2021.
23. F. Kebire Bardak, M. Nuri Seyman and Feyzullah Temurtas¸ “Adaptive neuro-fuzzy based hybrid classification model for emotion recognition from EEG signals”, Neural Computing and Applications, Vol. 36, pp. 9189–9202, 2024.
24. Jehosheba Margaret Matthew, Masoodhu Banu Noordheen and Mohammad Mustafa, “Enhancement of Hybrid Deep Neural Network Using Activation Function for EEG Based Emotion Recognition”, Traitement du Signal, Vol. 41, No. 4, 2024.
25. Zhibin Su, Yiming Feng, Jinyu Liu, Jing Peng, Wei Jiang and Jingyu Liu, “An Audiovisual Correlation Matching Method Based on Fine-Grained Emotion and Feature Fusion”, Sensors, Vol. 24, Issue 17, 2024.
26. Shofiqul Islam, Nahidul Islam, Noramiza Hashim, Mamunur Rashid, Bifta Sama Bari and Fahmid Al Farid, “New Hybrid Deep Learning Approach Using BiGRU-BiLSTM and Multilayered Dilated CNN to Detect Arrhythmia”, IEEE Access, Vol. 10, pp. 58081 – 58082, 2022.
27. R. Geethanjali and A. Valarmathi, “A novel hybrid deep learning IChOA-CNN-LSTM model for modality-enriched and multilingual emotion recognition in social media”, Scientific Reports, Vol. 14, 2024.
28. Xiaohu Wang, Yongmei Ren2, Bo Luo, Wei He, Jun Hong and Yinzhen Huang, “Deep learning-based EEG emotion recognition: Current trends and future perspectives”, Frontiers in Psychology, Vol. 14, No. 1126994, 2023.
29. Sara Bagherzadeh, Keivan Maghooli, Ahmad Shalbaf and Arash Maghsoudi, “A Hybrid EEG-based Emotion Recognition Approach Using Wavelet Convolutional Neural Networks and Support Vector Machine”, Basic clinical neuro science, Vol. 14, Issue 1, 2023.
30. Liumei Zhang, Bowen Xia, Yichuan Wang, Wei Zhang and Yu Han, “A Fine-Grained Approach for EEG-Based Emotion Recognition Using Clustering and Hybrid Deep Neural Networks”, Electronics, Vol. 12, No. 23, 2023.