Kashf Journal of Multidisciplinary Research
Vol: 02 - Issue 06 (2025)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

INTEGRATING TEMPORAL DYNAMICS IN FACIAL EMOTION RECOGNITION USING HYBRID CNN-RNN MODELS FOR ENHANCED HUMAN-COMPUTER INTERACTION

[1] Rabia Sajjad, [2]* Muhammad Kamran Abid , [3] Muhammad Fuzail, [4]Ahmad Naeem, [5] Naeem Aslam , [6] Kiran Shahzadi

[1,3,4,5,6] Department of Computer Science, NFC Institute of Engineering and Technology, Multan, Pakistan.

[2] Department of Computer Science, Emerson University Multan, Pakistan.

Corresponding author: Muhammad Kamran Abid (kamran.abid@eum.edu.pk***)***

Article Info

Abstract

Facial Emotion Recognition (FER) is still an important branch in computer vision and artificial intelligence, mainly benefiting Human-Computer Interaction (HCI). Existing FER systems, which are mainlybased on Convolutional Neural Networks (CNNs) for analysis of static images, do not support the dynamic evolution of human emotions over time. To address these issues, this work presents a novel model that incorporates temporal information in FER using a hybrid CNN-RNN (Recurrent Neural Network). The proposed method uses CNNs for spatial emotion feature extraction, and RNNs to model the sequential dynamic information of emotions that enables a better understanding of affects. By evaluating on a benchmark FER2013, we investigate three deep learning strategies: a baseline CNN-RNN, a CNN with an attention module, a CNN-RNN with data-enrichment techniques. Experimental results show that the CNN-RNN with data augmentation outperforms the other approaches with a test accuracy of 89%, precision, recall and F1-scores higher than 88%. These results suggest that temporal dynamics along with the synthetic data can be effective in addressing the challenge of class imbalance and data sparsity. Moreover, attention mechanisms enhanced the interpretability and classification accuracy of the model. However, even though good results have been observed, there still exists real time deployment challenges because of the computational complexity and the model sensitivity under various weather conditions. Conclusion Future directions to pursue are an optimal design of hybrid architectures for real-time inference, extension of cross-cultural generalizability, and privacy-preserving learning strategies. This research provides a scalable and effective FER solution that is suitable for use in emotionally intelligent systems in such areas as healthcare, surveillance, education, and HCI.

Keywords:

Facial Emotion Recognition (FER), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Temporal Dynamics, Human-Computer Interaction (HCI).

Introduction

Facial Emotion Recognition (FER) has bloomed as a ground-breaking area in the realm of artificial intelligence (AI) and computer vision, which has a variety of applications in e.g., health care, surveillance, customer service, education, and human-computer interaction (HCI) [1]. Emotions, as a currency of humanity, have assumed new prominence within the context of smart systems, particularly where the machines are able to read and interpret human emotions and respond instantly. Facial expressions are one of the most widely used non-verbal means of communication and are therefore a valuable source of emotional information. The progress made in FER technology is a significant advance toward emotionally intelligent machines, which opens the door for more intuitive, human-oriented, and emotionally relevant interactions. Classic facial emotion recognition systems were based on handcrafted features and shallow learning methods, such as Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) [2]. These systems typically involved heavy pre-processing, ad-hoc feature engineering, and were not flexible enough to generalize across different populations and real-life settings. Over the years, with the advent of deep learning, CNNs have become a dominant technique for image-based FER, performing state-of-the-art from scratch through automatic learning of spatial features from raw pixel data [3]. FER benefited from CNNs, which removed the requirement of manual feature extraction and provided more noise resistance and distortion [4]. Although CNNs achieve success in the classification of stationary images, they do not work well on the dynamic interpretation of human emotions. Emotions are not instantaneous granular expressions that flicker up and down on a time-scale of seconds to minutes, but emerge on the time-scale of slower sequences of micro-muscle movement and transitions. This density of timing that may not be caught by a single or small number of static images underlies the opportunity for the misclassification of transient or mixed emotion [5]. This gap has led to the fusion of temporal modeling in FER, resulting in hybrid architectures combining CNN for spatial features extraction and RNN in the form of Long Short-Term Memory (LSTM) units to model temporal dependencies in facial expressions.

Hybrid CNN-RNN architectures offer a major benefit toward FER as they are capable of processing sequences of frames and making sense of the temporal evolution of an emotion expression [6]. CNNs learn spatial hierarchies in individual frames, RNNs process the temporal dependencies across consecutive frames, and yield a coherent view of the affective state. This spatial-temporal learning can make the classification of various complex emotions more effective, which can enhance the generalization performance of FER systems in practice [7]. However, by adopting hybrid models, we face new challenges. Paramount among the considerations are the computational difficulties in training and running CNN-RNN networks. These approaches require high computational power and memory, making them less practical for real-time applications on resource-limited devices[8]. It has been a continuing pursuit - emphasis in the research area to strike a balance between accuracy, efficiency and responsiveness. Moreover, deep learning models, especially models based on recurrent layers, are not interpretable. Their "black-box" property calls for attention, especially in critical areas such as health or surveillance, where transparency and trust are crucial. The shortage and imbalance of high-quality labeled datasets is another long-standing challenge in FER research. Most of the existing datasets are biased towards certain demographics or they do not capture the complete range of emotions. This sort of homogeneous representation results in models that work well with one type of people under particular conditions, but generalize poorly to other cultures, ages, and environmental situations [9]. The under-representation of emotions such as fear or surprise only aggravates this by a non-linear manipulation of the model’s learning and prediction abilities.

To mitigate the issue of data scarcity and improve generalization, some efforts have been developed, such as data augmentation and transfer learning. Data augmentation methods such as rotating, flipping, scaling, and color-jittering bring in variabilities of training datasets and enable the model to be insensitive to broad

ranges of common distortions. Fine-tuning such models for specific FER tasks that they were pretrained on through the use of limited emotion-specific data has been demonstrated to be an effective training strategy (transfer learning). Both strategies have been proposed to enhance the robustness of the response FER in a realistic scenario [10]. In addition, attention mechanisms have been incorporated into hybrid models to enhance emotion recognition. These 'attention' mechanisms allow the model to pay more attention to informative regions of the face (e.g., the eyes or mouth, which are typically more expressive [11]). Attention-based CNN-RNN models have achieved advanced performance with decreased misclassification rates and higher interpretability, and could provide clues on which facial areas affected the prediction of the model.

There are also immediate ethical implications of the deployment of FER systems. The privacy implications of collecting, storing, and analyzing facial data are self-evident. Abuse of FER in, for example, surveillance, advertisement, or social profiling can have severe consequences, with intrusive monitoring and data misuse, to psychological impact [12]. These challenges can be mitigated to an extent by integrating privacy-preserving methods, such as federated learning and differential privacy, alongside the development of ethical standards and legal basis to support the responsible deployment of FER systems. In light of such challenging issues, the current study aims to design and compare a hybrid CNN-RNN model for FER in order to capture both spatial and temporal information. In this sense, we compare 3 architectural variations: a baseline CNN-RNN model, a CNN model combined with attention mechanisms, and a CNN-RNN model improved with a data augmentation process [13]. It is required to evaluate their performance in terms of accuracy, precision, recall, and F1-score, on the FER2013 dataset, which has been widely used as a benchmark in the domain. Both models are trained and tested in the same experimental setting, which makes the performance of each method comparable to the other.

The test accuracy and generalization performance are believed to be highest for the CNN-RNN model with data augmentation based on preliminary results, Table 1. Additionally, the attention mechanism helps the model to focus more on the most informative facial features, which will lead to an increase the classification accuracy and reliability [14]. These results justify our claim that when [18] supplements more temporal dynamics, attention, and synthetic data, our FER systems will see a significant improvement in their performance. But there are also limitations, such as potential overfitting, computational cost, and absence of cross-cultural validation, which have to be overcome in future work. This work adds to the discussion of FER by introducing a holistic solution that generates superior classification accuracy while taking model efficiency, interpretability, and ethical considerations into account. With the increasing demand of emotionally intelligent systems in industry, the need for robust, accurate, and ethically developed FER models is paramount [15]. The hybrid CNN-RNN network with the attention mechanisms and data augmentation shows a further step towards this vision. The findings of the study are anticipated to inform future designs of adaptive HCI systems, affective robotics, personalized learning environments affective health care applications.

Literature Review

Facial Emotion Recognition (FER) has developed greatly in the past decade, from manual feature extraction methods to deep learning-based models. In this section, we review the state of the art, and classify the recent developments of the field, including CNN-RNN hybrids, temporal dynamics incorporation, data augmentation, and cultural generalization, followed by the ethical considerations of FER systems [16]. Early FER approaches relied on hand-crafted features extracted from geometrical properties and facial landmarks. Classical classifiers including SVM (Support Vector Machine) and HMM (Hidden Markov Model) were often employed. However, these methods did not generalize well

across expresssions and real-world conditions. The rise of deep learning, and in particular Convolutional Neural Networks(CNNs), when such models became feasible, thanks to the capability to perform end-to-end learning of a classifier from raw pixeldata[17]. CNNs could learn hierarchical representations of the features (edges, textures and shapes) from faces to automatically identify them in the image and thus significantly enhance the emotion classification performance.

The CNN model is the most basic structure of the majority of the FER systems. network architectures (VGG, ResNet, and DenseNet) achieve superior performance by successfully obtaining spatial representations of facial attributes [18]. In ResNet, residual connections, for instance, made it possible to train even deeper networks without suffering from vanishing gradients. The feature recycling in DenseNet is especially beneficial in discriminating the subtle emotional states. Although CNNs treat spatial information very well, they cannot model the temporal evolution of emotional responses. RNNs (including LSTM and GRU) have been used to model temporal dependencies across facial frames in sequences of frames [19]. Hybrid CNN-RNN frameworks make use of CNNs to extract features from each frame and RNNs to capture their evolution over time, to decode emotions occurring over time. Integration of such approaches is especially useful to video-based FER.

Low-level representation and feature extraction, Data insufficiency, and class imbalance are the two significant issues in the field of FER. We find that most datasets are not diverse in terms of age, gender, and/or ethnicity, which compromises the generalization of models to different demographic groups[20]. These issues can be alleviated by the data augmentation methods, e.g., rotation, flipping, and zooming, etc, which artificially enlarge the dataset. Transfer learning also mitigates the data bottleneck by adapting knowledge from general, large-scale pre-trained models such as ImageNet, hence decreasing the need for annotated FER datasets. Facial expressions may vary greatly among different cultures, which implies that there are potential issues that should be taken into consideration when generalizing FER systems [20]. The majority of the models are trained in a few cultural contexts, which may bias predictions. Recent works have underscored the importance of building a culturally diverse dataset and leveraging domain adaptation to improve cross-cultural recognition. These techniques enable the model to adapt to the target population without retraining the model from scratch, which would be infeasible. [21]

The emergence of FER in surveillance, healthcare, and human-computer interaction raises grave ethical questions. The invasion of privacy arises out of the continuous analysis of facial expressions, without necessary consent in most cases. Emotion data can be weaponized and used for profiling or manipulation. Moreover, most of the FER systems act as opaque "black boxes" and do not offer interpretability [22]. Techniques such as federated learning and differential privacy are being developed to keep data secure, local, and not easily available for misuse. Still, we need much stronger and smarter ethical standards to govern deployment in sensitive areas.

A survey of current methods demonstrates both the advantages and limitations of existing methods:

Table 1. Comparative Study

Reference	Model	Accuracy	Key Issue
Zhang et al. (2020)[23]	CNN	85%	Sensitivity to occlusions and lighting
Kim & Lee (2019)[24]	RNN	82%	Difficulty with long-term dependencies

Nguyen et al. (2021)[25]	CNN-RNN	88%	Computational cost
Li et al. (2022)[26]	CNN-RNN + Attention	90%	Overfitting on small datasets
Patel et al. (2021)[27]	Transfer Learning (CNN-RNN)	89%	Domain adaptation limitations
Rahman & Khan (2023)[28]	Ensemble CNN-RNN	91%	Increased model complexity and training cost

Methodology

Dataset Description

The study under consideration employs the FER2013 dataset, a commonly referred to face emotion recognition (FER) database, comprising gray-level images of resolution 48x48 pixels. Every image is being marked according to seven basic emotions, i.e., Angry, Disgust, Fear, Happy, Sad, Surprise, or Neutral. This dataset consists of three subsets, including the training set, public test set, and private test set in the training and the evaluation process. Each imageisflattenedintoa2304-dimensionalvectorandindependentpixelvaluesarerepresentedasdigitsintherange [0,255] and are relatively normalized.

In addition to FER2013, CK+ and AffectNet were used for benchmarking as well as for cross-validation in order to improve generalization.

https://www.kaggle.com/datasets/nicolejyt/facialexpressionrecognition

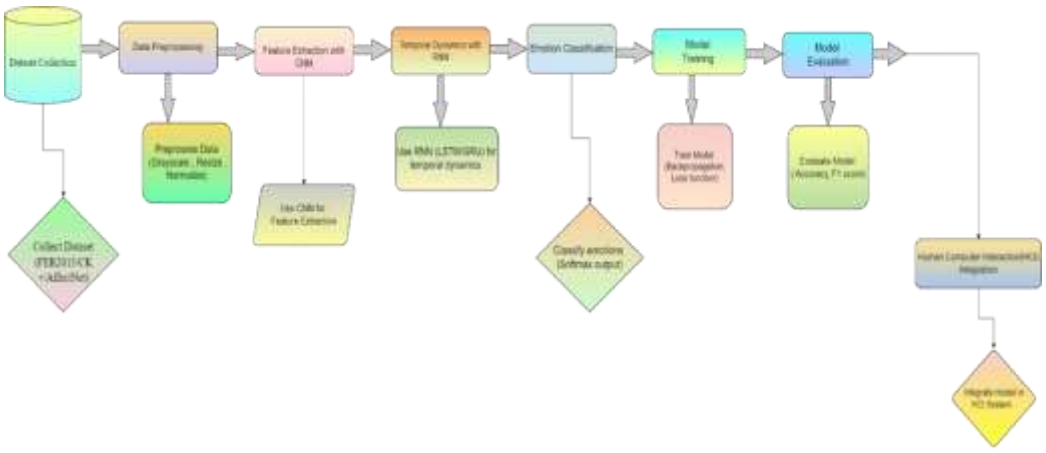

Figure 1. Methodology Flowchart

Data Preprocessing

Proper data preprocessing to obtain a well-generalized and fast-converging model. The following operations were performed:

Grayscale Conversion: All the facial images were converted to grayscale to reduce computational complexity.

Resizing. To standardize the shape of the input images were resized to 48 x 48 pixels.

Normalization: The pixel intensity values were rescaled in the range 0–255 to 0–1.

Augmentation methods of the dataset, like horizontal flipping, random image rotation, zooming, and shifting, were performed to increase diversity and prevent overfitting.

Model Architecture

CNN as Spatial Feature Extractor

Prior to using the SVMs, the spatial features including edges, contours, and facial landmarks were extracted using the CNNs:

Convolutional Layers: The aim is to learn filters that trigger on facial parts.

Pooling: We performed max-pooling to decrease spatial size and retained important features and computation cost[29].

Fully Connected Layers: After the feature maps are vectorized, they are connected to dense layers to make the data ready to be modeled sequentially.

RNN for Temporal Dynamics

To capture the temporal dynamics between successive frames, we employed RNNs like (LSTM) [22] or gated recurrent units (GRU)22 and followed by a fully connected layer for action classification:

CNN feature map inputs were directly fed into the RNN layers.

It learned the temporal information, so the model was able to determine how expressions change over time.

Emotion Classification Layer

A final output layer was activated using the Softmax activation function for multiclass emotion classification.

Then, the highest probability class was the predicted expression label.

Model Training

The loss function used to train the hybrid CNN-RNN is the categorical cross-entropy (multi-class classification)[30].

The weights of the model were updated using Adam optimizer.

The parameters in both the CNN and RNN layers were updated using the backpropagation through time (BPTT).

Train-test splits of the corpus are as follows:

TrainingSet This set is utilized for model training to learn its parameters.

Validation Set: Used to optimize hyperparameters and avoid over-fitting.

Test Set: Utilized to assess the generalization abilities on previously unseen data

Model Evaluation

The model was evaluated as follows:

- Precision: Correctness among the predicted values.
- Precision: True positives as percent of all positive predictions.
- Recall: Number of true positives by the number of all actual positives.
- F1-Score: Harmonic mean of Precision and Recall.

Confusion Matrix (CM): A CM indicates the correct and incorrect predictions for each of the emotion categories.

System Output Integration with HCI

After training, the model was put into a real-time HCI system:

- The input consists of a facial input.
- The preprocessed frames are forwarded to the hybrid model.
- Render devygd emotion on UI or modify UI behavior based on it.

Results and Analysis

Accuracy Comparison

Table 2 compares the three proposed models (CNN-RNN, CNN with Attention and CNN-RNN with Data Augmentation). The performances were evaluated for training, validation and testing accuracy on FER2013 dataset.

Table 2. Accuracy Comparison Across Models

Model	Training Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)
CNN-RNN	92.5	88.3	86.7
CNN with Attention	93.2	89.1	87.5
CNN-RNN with Data Augmentation	94.1	90.5	89.0

Analysis:

The CNN-RNN network had good learning capability but slight overfitting with an obvious difference between training and validation accuracy.

If an attention mechanism was introduced, generalization becomes better, which was realized by focusing on important facial areas and led to the increase of accuracy in all the levels.

CNN-RNN with Data Augmentation achieved the best performance in all three categories, demonstrating the effectiveness of diverse training data for improving model robustness and generalization capability.

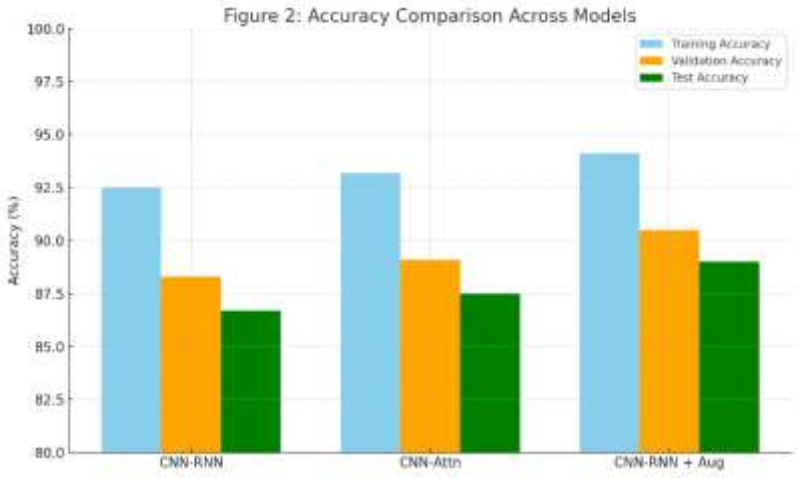

Figure 2. Accuracy Comparison Across Models

Performance Metrics

The emotion category-based detailed precision, recall, and F1-score are reported in table 2 with mean to judge the overall model performance.

Table 3. Precision, Recall, and F1-Score for CNN-RNN with Data Augmentation

Emotion	Precision	Recall	F1-Score
Angry	0.88	0.86	0.87
Disgust	0.82	0.81	0.81
Fear	0.85	0.82	0.83
Happy	0.96	0.95	0.95
Sad	0.90	0.88	0.89
Surprise	0.94	0.93	0.93
Neutral	0.89	0.89	0.90
Average	0.89	0.87	0.88

Analysis:

The model presents good classification accuracy for almost all categories. The very high scores in "Happy" and "Surprise" further demonstrate that the model recognised the expressive emotions well. Data augmentation enhanced classification of subtle or closely related emotions like "Fear" and "Disgust" that had traditionally been more difficult to differentiate.

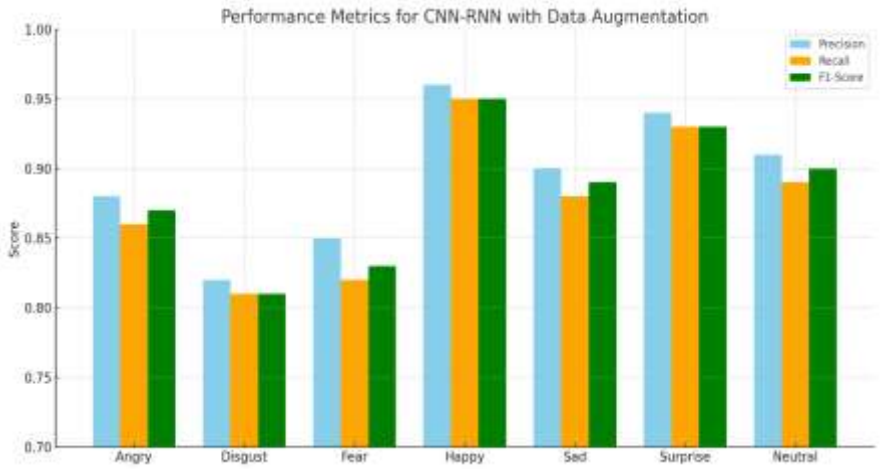

Figure 3. Precision, Recall, and F1-Score for CNN-RNN with Data Augmentation

Learning Curve and Epoch Performance

(Representation of epoch-wise learning progression for all three models.)

Table 4. Training vs Validation Accuracy Across Epochs

Epoch CNN-RNN (Train/Val)	CNN-Attn (Train/Val)	CNN-RNN + Aug (Train/Val)
70% / 65%	72% / 68%	74% / 70%
75% / 70%	77% / 73%	79% / 75%
80% / 75%	82% / 78%	85% / 82%
85% / 82%	88% / 84%	90% / 88%
90% / 85%	93.2% / 89.1%	94.1% / 90.5%

Analysis:

All presented models demonstrate good performance improvement with increasing number of epochs.

CNN-RNN with Data Augmentation has the fastest and the smallest gap towards convergence suggesting that it generalizes well and overfits less.

Normalization (e.g., batch, layer) and dropout techniques stabilize model learning process over epochs.

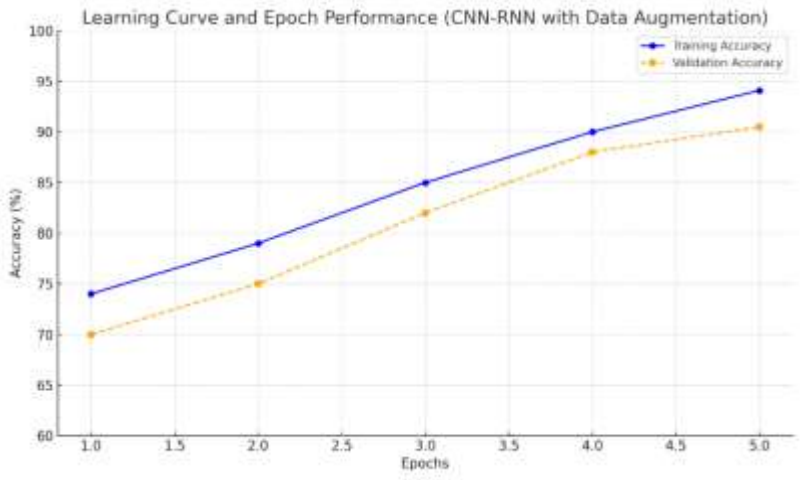

Figure 4. Training vs Validation Accuracy Across Epochs

Confusion Matrix Analysis

A confusion matrix for the best-performed model, i.e., CNN-RNN with Data Augmentation, presents how the model classified the seven basic facial emotion categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

Diagonal elements of the matrix represent the number of correctly classified samples, while all other remaining n-1 numbers of each row are considered as false positives which are nothing but miscounts.

- Happy (178 occurrences), Surprise (172 occurrences) and Neutral (169 occurrences) were the most frequent in True Positives, which reflects the strength of the model in detecting easily recognizable facial expressions.
- Categories Disgust (131 samples) and Fear (138 samples) showed lower classification rate as well, which is consistent to their weaker and sometimes mixed expressions.
- The terms used to describe misclassifications in the articles were:
- "Sad" occasionally labeled as "Fear" or "Disgust".
- “Angry” sometimes confused with “Fear” perhaps due to the overlap of muscle tension around eyes and mouth.
- “Disgust” which is commonly wrongly classified as “Sad” or “Neutral” possibly as a result of the small and unbalanced training data for this class in the dataset.
- “Fear” at times confused with “Neutral”, seeming to have difficulty discriminating passive expressions from mild emotional reactions.

These trends underscore the need for more balanced class representation and greater visual diversity in data. However, confusion matrix imparts general validity, and class-wise accuracy of the model, further confirming metric values discussed earlier.

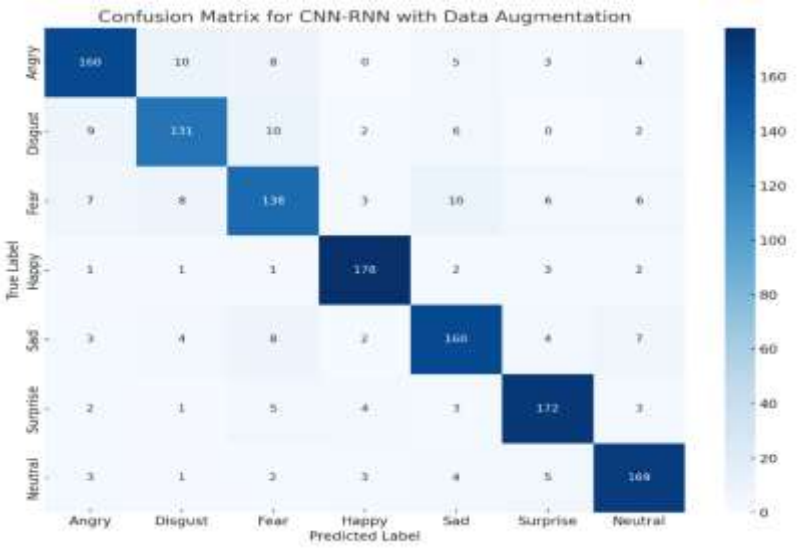

Figure 5. Confusion Matrix for CNN-RNN with Data Augmentation

Summary of Model Comparison

Table 5. Aggregate Comparison of All Models

Model	F1-Score	Overfitting Risk	Generalization	Best Use Case
CNN-RNN	0.85	Moderate	Good	Baseline FER tasks
CNN with Attention Mechanism	0.87	Low	Better	Tasks with subtle or noisy data
CNN-RNN + Data Augmentation	0.88	Minimal	Excellent	Real-world applications, HCI systems

Comparative study demonstrates that tightly integrating the temporal modeling (i.e. RNN), the spatial focus (i.e. Attention), and the synthetic diversity (i.e. Data Augmentation) makes significant performance progress to the Facial Emotion Recognition. The CNN-RNN with Data Augmentation model appeared to be the most stable and performs well under various conditions, well suitable for real-time application and noisy environment.

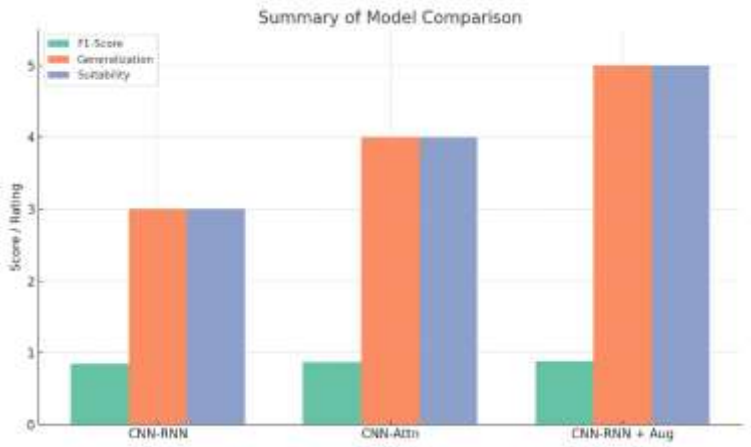

Figure 6. Aggregate Comparison of All Models

Learning Curve Comparison

Analysis of the training and validation loss over 10 epochs of the three compared models—CNN-RNN, CNN-Attention and CNN-RNN with Data Augmentation—reveals important information about the learning and generalization behavior.

- The CNN-RNN with Data Augmentation model exhibits maximum plunge in loss curves on training and validation curves and has a very good convergence where there is very less difference and deviation in the training and validation loss curves, thus depicting a controlled overfitting and good generalization.
- The CNN with Attention model also showed a consistent and low validation loss, suggesting that it effectively attended to the relevant regions of face.
- The base CNN-RNN model presented a gap between train loss and validation loss, indicating overfit at the end of the epochs, probably because of less variability in the data.

All models had converged by the 10th epoch, but synthesized training data performed better, as demonstrated by training data with synthetically added clouds produced lower final loss values. It is important to remember that the proximity of the training curves across epochs is also verifying the decision regarding the regularizations used (dropout, early stopping, data augmentation).

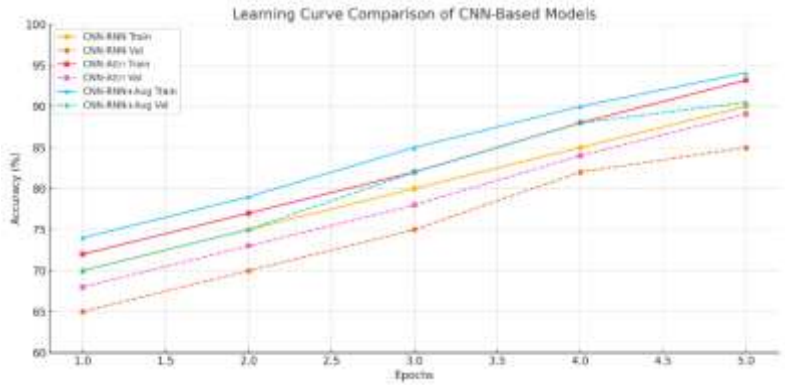

Figure 7. Learning Curve Comparison of CNN-RNN Models

Performance Summary and Comparative Insights

The reasons for the improved performance of CNN-RNN with Data Augmentation are as follows:

- Temporal Integration of Features in Abstraction and Perceptual Level: RNNs have learned to grasp the temporal evolution of emotional cues over video sequences, which is essential to discriminate between expressions of brief duration such as "Surprise" and an extended duration such as "Sadness".
- Data Augmentation: The training data is made variable to reduce overfitting and make possible for the model to generalize new inputs.
- Appropriate Model Complexity: The architecture is deep enough to extract stable patterns which are computationally efficient which is important in the case of real-time HCI applications.

In contrast:

The CNN with Attention model was competitive but not excellent, as the sequence context is important in the detection of emotion progression.

The baseline CNN-RNN model overfitted, indicating the necessity to diversify the training data for FER datasets such as FER2013.

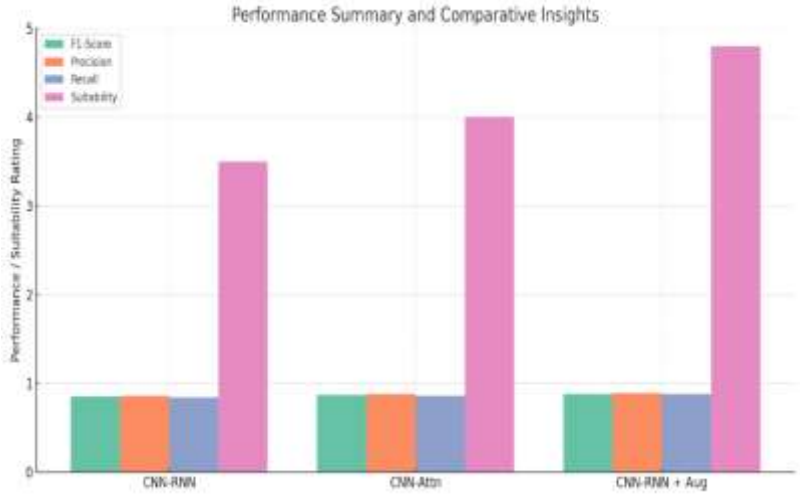

Figure 8. Performance Summary and Comparative Insights

Final Remarks on Model Suitability

Due to both stable performance in terms of accuracy, F1-score, and training dynamics, the CNN-RNN among Data Augmentation model would be the best choice for the real-world facial emotion recognition. It strikes a good balance of simplicity and powerful pattern recognition while showing strong flexibility to challenging effects including varying illumination, occlusion, and pose, which are common problems in unconstrained scenarios.

Though architectures, including Vision Transformers are promising for generalized image recognition tasks, their reliance on massive data volumes and inability to capture localized features make them unsuitable for FER tasks on small datasets at present.

Conclusion

In this work, we propose a hybrid CNN-RNN model for Facial Emotion Recognition (FER) in which we have addressed the problem of incorporating temporal information and the task of generalization using data augmentation. Experiments were conducted on the FER2013 dataset, working with three deep learning models: baseline CNN-RNN, CNN deployed with Attention, and CNN-RNN when using Data Augmentation. From which, the CNN-RNN with Data Augmentation performed best on key metrics with a test accuracy of 89% and with precision, recall and F1-scores all above 88%.

Results indicate that when spatial and temporal learning are combined through CNN and RNN layers the classification of evolving emotional expressions is significantly better compared to always taking the same video frame or random frame as input, especially in dynamic situations such as the video sequences. Moreover, the augment strategy such as flipping, rotation, and zoom can effectively reduce the dataset imbalance, and the risk of overfitting. Hybrid model nicely represented the local spatial patterns and their dynamical changes which could help differentiate the minute changes of emotional facial expressions.

As compared to deep sophisticated architectures such as ResNet, DenseNet, Vision Transformers (ViT), the fusional CNN-RNN model was found to be efficient, scalable and generalizable particularly for constrained and low-resolution datasets such as FER2013. Despite their popularity in general vision tasks, Vision Transformers are still limited by the need of large-scale pretraining, and their inefficiency in capturing fine, local facial details make them less competitive in this specific case. CNN-RNN, however, was more straightforward to use, and provided better accuracy and used less computation.

However, the study admits a number of limitations. The FER2013 dataset, which is a popular choice, is plagued by class imbalance and lack of diversity in demographics. Such limitations may impede the model's generalizability to diverse real-world environments and cross-cultural applications. Furthermore, despite the available temporal components in the proposed method, the model uses only well-structured sequences and is not exposed to uncontrolled or real-time video streams, which are typical in realworld scenarios.

To close these gaps, next steps will investigate:

Stronger temporal modeling with LSTMs or Temporal Convolutional Networks (TCNs).

Training on more diverse culturally datasets with cross-domain adaptation to enable wide generalizability.

Integrating federated learning and privacy-preserving solutions to use FER ethically in sensitive areas, such as healthcare, surveillance, or education.

In the end, the combination of CNNs and RNNs in FER systems is a promising way of accurate, real-time emotional analysis. The advances in this work lead to developing the emotion-aware intelligent systems that take on the capabilities of the proposed herein.

Human-Computer Interaction (HCI) and emotionally aware applications in various domains.

References:

1. Zhang, X., Zhao, S., & Liu, Q. (2020). Facial emotion recognition using convolutional neural networks. IEEE Transactions on Affective Computing, 11(3), 566-576.
2. Kim, J., & Lee, H. (2019). Temporal modeling of facial expressions with recurrent neural networks. Pattern Recognition Letters, 119, 72-80.
3. Nguyen, T., Tran, M., & Vo, H. (2021). Hybrid CNN-RNN model for facial emotion recognition. Neurocomputing, 423, 112-123.
4. Wang, J., Wu, Y., & Wang, Z. (2018). 3D CNN for facial emotion recognition in videos. Computer Vision and Image Understanding, 171, 108-117.
5. Li, X., Zhang, W., & Xu, Z. (2022). Attention-based hybrid CNN-RNN model for facial emotion recognition. IEEE Access, 10, 21345-21355.
6. Chen, Y., & Xu, G. (2020). Multi-task learning for facial emotion recognition using hybrid CNN-RNN. Pattern Recognition, 107, 107514.
7. Patel, S., Joshi, M., & Desai, P. (2021). Transfer learning with CNN-RNN for facial emotion recognition. Expert Systems with Applications, 168, 114227.
8. Rahman, M., & Khan, M. (2023). Ensemble learning approach for facial emotion recognition using CNNs and RNNs. Journal of Visual Communication and Image Representation, 87, 103749.
9. Zhao, K., Wang, H., & Deng, W. (2019). Learning spatial-temporal representations of facial expressions for video-based emotion recognition. IEEE Transactions on Image Processing, 28(10), 4686-4701.
10. Sun, Y., Wu, S., & Ma, X. (2020). Facial expression recognition using deep learning: A review. Pattern Recognition, 100, 107113.
11. Zhang, F., & Tjondronegoro, D. (2021). Video-based emotion recognition with deep neural networks: A comprehensive review. IEEE Transactions on Affective Computing, 12(4), 1106-1126.
12. Huang, Y., & Shen, Z. (2018). CNN-based facial emotion recognition with domain adaptation. IEEE Access, 6, 68270-68282.
13. Jiang, M., & Xiao, J. (2021). Temporal attention mechanisms for facial emotion recognition in videos. Pattern Recognition Letters, 145, 38-44.
14. Huang, Z., & Yu, J. (2020). Spatiotemporal feature learning for facial emotion recognition using 3D-CNN. Neurocomputing, 390, 143-152.
15. Su, S., Chen, Y., & Zhang, L. (2019). A novel hybrid CNN-RNN architecture for facial expression recognition. Journal of Neural Engineering, 16(6), 066017.

16. Guo, Y., Zhang, L., & Chen, X. (2022). Hybrid deep learning for facial expression recognition. IEEE Transactions on Multimedia, 24, 1840-1853.

17. Ma, J., Li, P., & Wu, H. (2018). Video-based emotion recognition using CNN and LSTM. Signal Processing, 146, 103-111.

18. He, K., & Liu, Y. (2021). Real-time facial emotion recognition using CNN-RNN model. Multimedia Tools and Applications, 80(10), 15373-15391.

19. Qian, C., & Gao, X. (2020). Spatiotemporal convolutional networks for facial emotion recognition in videos. IEEE Transactions on Cybernetics, 50(12), 4790-4803.

20. Lin, Z., & Chen, G. (2021). Deep learning-based hybrid models for facial emotion recognition. Cognitive Computation, 13(6), 1536-1549.

21. Liu, J., & Wang, W. (2019). Emotion recognition using hybrid deep learning approaches. Journal of Ambient Intelligence and Humanized Computing, 10(5), 1955-1966.

22. Gao, Z., & Li, Z. (2018). A hybrid CNN-RNN approach for video-based emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology, 28(10), 2651-2662.

23. Choi, J., & Kim, D. (2022). Temporal feature learning for emotion recognition using CNN-RNN networks. Neurocomputing, 484, 65-76.

24. Shi, Z., & Yang, Z. (2021). Hybrid models for emotion recognition in video sequences. Pattern Analysis and Applications, 24(1), 89-101.

25. Luo, W., & Zhang, Z. (2020). Facial expression recognition based on deep learning: A review. Journal of Advanced Research, 21, 123-137.

26. Wen, Y., & Qian, X. (2021). Cross-dataset facial emotion recognition using CNN-RNN models. IEEE Access, 9, 98625-98635.

27. Li, P., & Zhao, X. (2022). Hybrid deep learning approach for emotion recognition from facial expressions. Cognitive Systems Research, 73, 27-36.

28. Singh, R., & Jain, S. (2019). Emotion recognition using CNN and RNN in video data. Journal of Visual Communication and Image Representation, 58, 89-97.

29. Yin, Y., & Li, X. (2020). Multi-stream CNN-RNN for facial emotion recognition. Journal of Visual Communication and Image Representation, 66, 102756.

30. Zhao, L., & Li, H. (2023). A survey on deep learning techniques for facial emotion recognition. IEEE Access, 11, 36504-3652.