IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2581~2594
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2581-2594

Hyperparameter optimization of deep residual recurrent fusion models for facial emotion recognition

Muhammad Munsarif[1,2], Ku Ruhana Ku-Mahamud[3,4]
[1]Computer Science Study Program, Faculty of Engineering and Computer Science, Universitas Muhammadiyah Semarang, Semarang, Indonesia
[2]Department of Information and Communication Technology, Asia e University, Selangor, Malaysia
[3]School of Computing, Universiti Utara Malaysia, Kedah, Malaysia
[4]Faculty of Business Management and Information Technology, Universiti Muhammadiyah Malaysia, Perlis, Malaysia

Article Info

Article history:

Received Feb 18, 2025
Revised Feb 13, 2026
Accepted Apr 20, 2026

Keywords:

Bidirectional long short-term
Deep learning
Facial emotion recognition
Gated recurrent unit
Long short-term memory
Recurrent neural networks
Residual networks

ABSTRACT

Deep learning facial emotion recognition (FER) is widely applied in healthcare, education, and human–computer interaction. However, many deep learning models suffer from suboptimal hyperparameter configurations that reduce accuracy and stability. This study proposes three deep residual recurrent fusion models that integrates residual blocks with recurrent neural networks (bidirectional long short-term memory (BiLSTM), long short-term memory (LSTM), and gated recurrent unit (GRU)) to capture both spatial and temporal features. A systematic hyperparameter optimization strategy was applied, tuning kernel size, filter size, recurrent units, batch size, learning rate, dropout, and weight decay to balance generalization and computational efficiency. The models were evaluated on four benchmark datasets: FER2013, FERPlus, RAF-DB, and CK+. The results show that optimized configurations achieved outstanding accuracy, reaching 99.85% on FER2013, 99.99% on FERPlus, and 100% on RAF-DB and CK+. These findings demonstrate that careful hyperparameter tuning significantly enhances feature extraction, mitigates vanishing gradient and overfitting issues, and improves generalization across diverse datasets. The proposed framework highlights the importance of optimization in advancing robust FER systems for real-world applications.

Corresponding Author:

Muhammad Munsarif
Computer Science Study Program, Faculty of Engineering and Computer Science
Universitas Muhammadiyah Semarang
Semarang, Indonesia
Email: m.munsarif@unimus.ac.id

1. INTRODUCTION

In recent years, automatic facial emotion recognition (FER) has become a major research area in artificial intelligence (AI) and computer vision. This technology has a wide range of applications, such as in education to assess students' emotional responses to the learning process [1], in healthcare for real-time psychological condition monitoring [2], and in security, human-machine interaction, and customer behavior analysis in digital marketing [3]. Given the increasing demand for fast, accurate, and robust FER systems, researchers have explored and developed various models to improve recognition performance and computational efficiency. One such model is the use of long short-term memory (LSTM) unit called deep convolutional bidirectional long short-term memory (BiLSTM) fusion, which combines a deep spatial network (DSN) for spatial feature extraction, a deep temporal network (DTN) to capture expression

dynamics, and BiLSTM to model sequential information. This model has demonstrated high performance on the CK+, Oulu-CASIA, and MMI datasets but has yet to be tested on more complex datasets such as FER2013 [4]. Additionally, the multimodal transformer network, which integrates transformers with electroencephalography (EEG) and facial expression data, has proven robust in emotion recognition. However, it requires EEG data, which is difficult to obtain in real-world scenarios [5].

To enhance video-based facial expression recognition, researchers have developed several deep learning approaches that effectively capture both spatial and temporal features in facial movements. One widely used model is 3D-CNN + ConvLSTM, which integrates 3D convolutional neural networks (CNNs) for spatial feature extraction with convolutional long short-term memory (ConvLSTM) to capture the sequential nature of facial expressions over time. This approach has shown high accuracy on structured datasets such as CK+, SAVEE, and AFEW, where facial expressions are relatively well-aligned and contain minimal noise. However, its generalizability to real-world conditions remains uncertain, as it has been primarily evaluated on controlled datasets rather than on datasets that contain significant variations in pose, occlusions, lighting conditions, and facial misalignment. For example, FER2013 presents substantial challenges due to unconstrained settings, image noise, mislabeling, and diverse facial structures, which can severely impact recognition accuracy. The absence of extensive testing on such complex datasets raises concerns about the model's effectiveness in practical applications where facial expressions occur in dynamic and unpredictable environments [6].

Another model, the local-spatial-global-temporal network [7], adopts a CNN-transformer hybrid approach, combining spatial CNN for spatial feature extraction and temporal transformer (T-former) for processing global temporal dependencies. While this model has achieved high accuracy with low computational costs on the DFEW and FERV39k datasets, it still struggles to detect micro-expressions, which require precise temporal modeling accurately. In addition to CNN and transformer-based models, several studies have explored recurrent neural networks (RNNs) to capture temporal dependencies in sequential data. The gated recurrent unit (GRU) + CNN model within a teacher-student multitasking framework effectively captures temporal dependencies but suffers from high model complexity [8]. Meanwhile, the residual CNN approach leverages residual blocks to mitigate the vanishing gradient problem, thereby improving FER performance [9]. Furthermore, the attention-BiLSTM model, which integrates BiLSTM, multi-head attention, and residual connections, has been used to address network degradation caused by overfitting in the IEMOCAP dataset. However, it has yet to be tested on static facial expression datasets [10].

Several lightweight models have also been developed to enhance computational efficiency, aiming to reduce complexity while maintaining high accuracy. One such model is patch and attention MobileNet (PAtt-Lite), specifically designed to address challenges such as occlusions and facial pose variations across multiple datasets, including CK+, RAF-DB, FER2013, and FERPlus. This model has demonstrated high accuracy on well-structured datasets; however, its performance declines when applied to FER2013, which contains significant noise, misaligned faces, and class imbalance [11]. Similarly, the deep residual recurrent fusion model, which integrates CNN, residual network, and BiLSTM for improved spatial-temporal feature extraction, achieves high accuracy on CK+ but requires a large number of training epochs, leading to longer computational times [2].

Despite advancements in FER models, challenges remain in generalizing models across diverse datasets, improving computational efficiency, and differentiating visually similar expressions. Therefore, this study proposes a residual deep recurrent fusion-based model, which integrates residual blocks, BiLSTM networks, and spatial-temporal fusion techniques to enhance accuracy in FER. The proposed model captures complex spatial information from facial expression images while modeling temporal expression sequences using BiLSTM, resulting in more accurate and stable emotion mapping [10]. Hyperparameter tuning strategies are implemented to optimize the performance of the proposed model, including optimizing convolutional kernel size, BiLSTM units, batch size, and learning rate. This approach aims to balance accuracy and computational efficiency, ensuring the model excels in classification precision and adapts effectively to real-world data conditions [12]. The proposed model was evaluated on benchmark datasets such as FER2013, RAF-DB, and FERPlus, each presenting unique challenges in expression variations, image quality, and class distribution [11].

2. RELATED WORKS

Hyperparameter optimization is crucial in improving the performance of deep learning models, particularly in CNN and RNN. Various studies have implemented automated search strategies to optimize key hyperparameters such as learning rate, batch size, number of layers, activation functions, dropout rate, kernel size, and optimization algorithms, ensuring improved model accuracy and computational efficiency.

Several optimization methods are employed to refine CNN architectures. Genetic algorithm is applied to fine-tune the learning rate, number of layers, activation function, and optimization algorithm, leading to improved model convergence [13]. Using genetic algorithms, random searches, and grid searches to optimize learning rates, the number of filters, kernel size, batch size, and activation function has enhanced CNN-based classification performance [14]. Genetic algorithm is also successfully utilized to optimize optimizer selection, kernel size, activation function, number of layers, and filter size through an extensive automated search process [15].

Another widely adopted approach is Bayesian optimization, which enables an efficient search for the best hyperparameter configurations. This method is employed to fine-tune the learning rate, batch size, dropout, and convolutional layers, resulting in superior performance after multiple iterations of automated searches [16], [17]. Bayesian optimization has also been used for hyperparameter tuning in CNN-based models, focusing on L2 regularization, batch size, and number of layers, ensuring robust model performance with reduced overfitting [18], [19]. Some studies have explored other metaheuristic approaches, such as particle swarm optimization (PSO) and tree-structured parzen estimation to improve the efficiency of CNN hyperparameter search. The application of tree-structured parzen estimation and PSO has successfully optimized the number of convolutional layers, kernel size, batch size, learning rate, and activation function, demonstrating improved efficiency in hyperparameter selection [20]. The combination of genetic algorithm, PSO, ant colony optimization, and Bayesian optimization was also used to optimize the learning rate, number of layers, kernel size, and activation function, significantly improving the accuracy of CNN, LSTM, and support vector machine (SVM) models [21]. Grid search and random search are also fundamental methods for CNN hyperparameter tuning. Studies have explored asynchronous successive halving algorithm (ASHA), grid search, Bayesian optimization, and random search for optimizing learning rate, dropout, batch size, and optimization methods, successfully identifying optimal hyperparameters through extensive automated searches [22]. The use of grid search to optimize learning rate, batch size, number of convolutional layers, and dropout has also contributed to improved CNN generalization [23], [24]. Hybrid architectures, such as CNN-BiLSTM-AR and 3D–2D CNN, have also been optimized using various hyperparameter search techniques. The application of PSO, genetic algorithm, and random search has successfully optimized the learning rate, batch size, dropout, and convolutional layers, significantly enhancing temporal sequence modeling [25]. Integrating band selection and attention mechanism in hyperparameter tuning for 3D–2D CNN has also proven effective in balancing model complexity and accuracy [26].

For models based on RNNs, LSTM, and BiLSTM, hyperparameter tuning techniques were implemented to enhance sequential data processing. Adaptive learning rate tuning, cyclical learning rate, and hyperparameter transfer dynamics approaches have been used to optimize learning rate, dropout, batch size, and number of layers, significantly improving model performance in time-series applications [27]. Furthermore, neural architecture search was explored as an automated approach for hyperparameter optimization. This method was applied to refine U-Net architectures, ensuring an adaptive and automated hyperparameter tuning process [28]. The integration of AutoML with Bayesian optimization and grid search was also applied in tuning CNN-autoencoder models, achieving high efficiency in feature extraction and regularization [19]. Despite the significant advancements in hyperparameter optimization, challenges remain in balancing model complexity, training time, and accuracy. Most studies rely on automated search processes involving hundreds of cases, yet selecting the best hyperparameter configuration still requires extensive computational resources. Further research is needed to develop more adaptive and computationally efficient hyperparameter tuning strategies, particularly for large-scale datasets and real-time applications [28].

3. METHOD

This study develops a deep residual recurrent fusion model to enhance the performance of FER through the integration of residual, spatial, and temporal features. Figure 1 illustrates the methodology workflow. This workflow includes data preprocessing, the development and fusion of residual–recurrent networks, systematic hyperparameter tuning, and model performance evaluation.

3.1. Datasets

The proposed model is evaluated using multiple benchmark datasets commonly used in FER research to ensure robustness and generalizability. The datasets used in this study include FER2013, FERPlus, RAF-DB, and CK+, which provide a diverse range of facial expressions and variations in real-world conditions as shown in Table 1. FER2013 is a widely used dataset containing 35,887 grayscale images of facial expressions categorized into seven emotion classes. FERPlus extends FER2013 by incorporating improved labels through crowdsourcing, increasing its robustness with eight emotion classes. RAF-DB consists of 15,339 facial images collected from real-world scenarios, annotated with seven basic emotion

classes. Lastly, CK+ is a controlled dataset containing 981 facial expression sequences transitioning from neutral to peak emotions, making it suitable for training models in recognizing dynamic facial expressions.

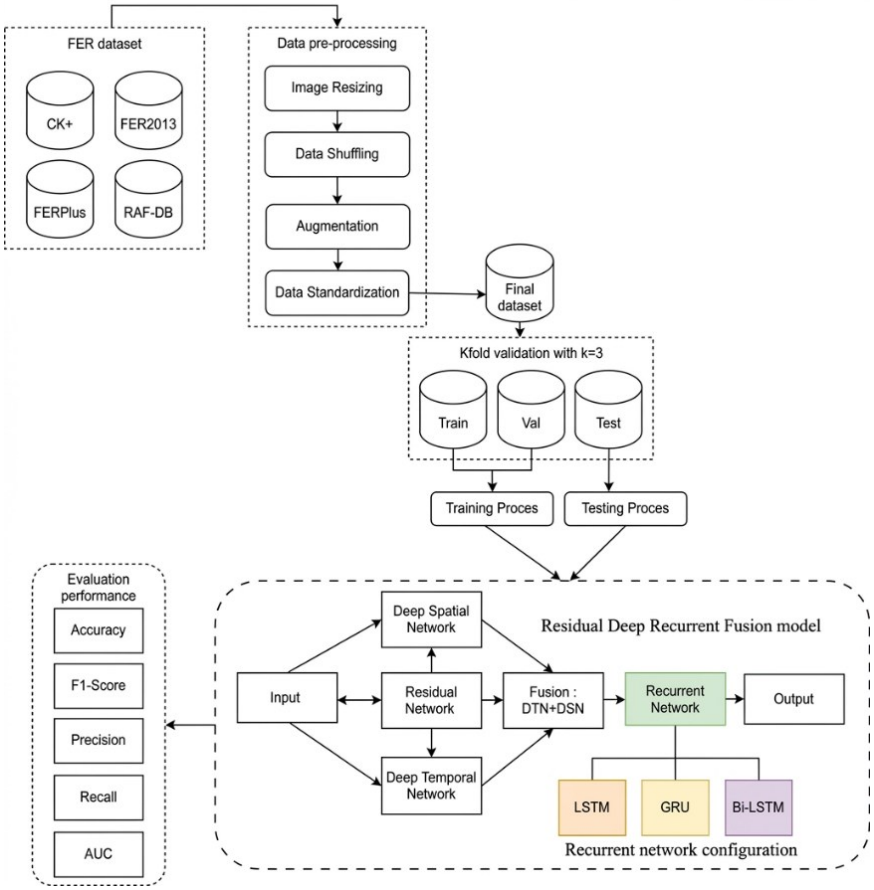

Figure 1. Workflow of methodology

Table 1. Allocation for experiment (in-the-wild databases)

Database	Training	Validation	Testing	Total
RAF-DB	12,271		3,068	15,339
FER2013	28,729		7,158	35,887
FERPlus	66,379	8,341	3,573	78,293
CK+	785		196	981

3.2. Pre-processing dataset

The data pre-processing stage is carried out to ensure that facial image data are consistent and ready for use in the training process of the FER model. Images from the CK+ dataset are resized to 48×48 pixels, while images from the RAF-DB, FER2013, and FERPlus datasets are resized to 100×100 pixels in accordance with their respective characteristics. All images are then normalized to have zero mean and unit variance in order to stabilize the training process and reduce computational complexity. Several data augmentation techniques are applied to enhance data diversity and improve the model's generalization capability, including random rotations between -30° and 30°, horizontal flipping, random cropping followed by resizing, and random adjustments to brightness, contrast, saturation, and hue. These augmentation strategies help the model become more robust to variations in facial expressions, lighting conditions, and viewing angles, thereby improving its performance on unseen data.

3.3. Development of the proposed deep residual recurrent fusion model

The residual network is a key component of the proposed model, designed to address the vanishing gradient problem and improve training stability and efficiency. This study integrates the residual network

into two main components: enhanced DTN and enhanced DSN. One of the challenges in sequential data analysis is overfitting, especially when dealing with complex facial expression patterns. The residual blocks are implemented in DTN to retain crucial information from previous layers and enhance training stability. The model must also accurately extract spatial features, so DSN is reinforced with residual blocks, enabling deeper exploration of facial patterns and improving the accuracy of facial expression classification. By incorporating residual blocks into DTN and DSN, the model can retain essential input information, enhance generalization, and prevent accuracy degradation in deeper networks. The deep residual recurrent fusion model integrates spatial and temporal networks to optimize FER. This model is developed in three main variants based on BiLSTM, LSTM, and GRU units.

3.3.1. Deep residual bidirectional long short-term memory fusion model

Figure 2 shows the flowchart of the proposed deep residual BiLSTM fusion network, while Figure 3 illustrates the complete architecture of this model. The model is a sophisticated architecture designed for FER that combines spatial and temporal features through layers of DSN, DTN, and RNN using BiLSTM. The model workflow begins with image preprocessing, followed by feature extraction using CNNs, which recurrent networks then process. The residual blocks ensure smooth gradient flow during training, and finally, the processed features are fed into a fully connected layer for final classification. The main advantage of this model lies in its ability to combine spatial and temporal information while preventing overfitting, leading to improved performance in facial expression recognition.

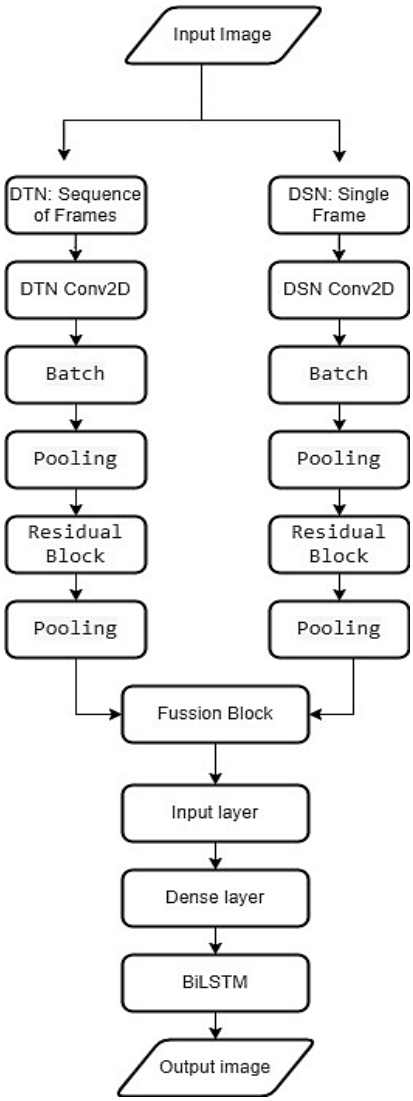

Figure 2. Flowchart of deep residual BiLSTM fusion model

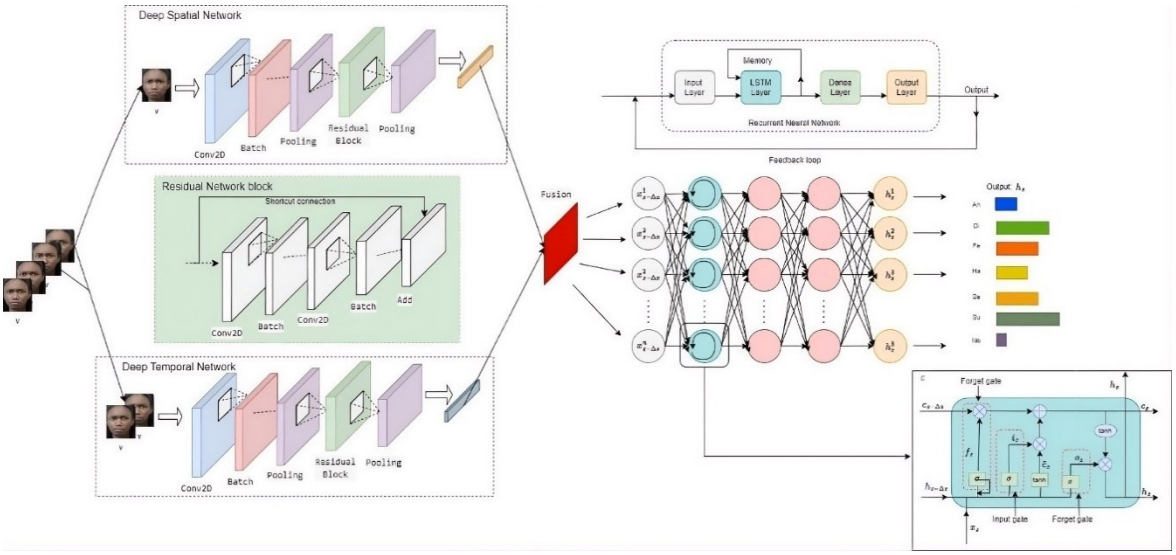

Figure 3. Architecture of deep residual BiLSTM fusion model

The deep residual recurrent fusion model focuses on effectively integrating spatial and temporal feature extraction, enhancing the model's ability to recognize facial expressions accurately. In this architecture (refer to Figure 3), using convolutional layers, the DSN is responsible for extracting spatial features from static images, such as facial structures and critical points like the eyes, nose, and mouth. On the other hand, DTN processes sequences of images or video frames to capture changes in facial expressions over time, focusing on temporal dynamics. By utilizing residual blocks in DSN and DTN, the model ensures that essential information is preserved across layers, preventing critical data loss during deep network processing. The DSN is used to extract spatial features from facial images. At the same time, the DTN captures temporal dynamics from image sequences, adding residual blocks to maintain gradient stability and ensure the flow of essential information. After the spatial and temporal features are extracted, they are merged in the fusion layer, which combines the strengths of both networks. This fusion allows the model to create a richer and more comprehensive feature representation, considering the spatial layout and the temporal evolution of facial expressions. The merged features are then passed through an LSTM network. The LSTM is crucial for capturing long-term dependencies, allowing the model to track and understand the sequential flow of facial expressions over time. This spatial and temporal information combination enhances the model's ability to recognize subtle expression changes, which is essential for tasks like FER. The integration of residual blocks plays a vital role in maintaining stability and ensuring that critical features are not lost during the learning process. The architecture is designed to tackle challenges such as overfitting and the vanishing gradient problem, making it highly effective for handling complex datasets and generating precise facial expression predictions. This approach makes deep residual recurrent fusion a powerful solution for various applications, such as pattern recognition in images, video analysis, or natural language processing, with high efficiency and accuracy in handling complex data.

3.3.2. Deep residual long short-term memory fusion model

The deep residual recurrent fusion network is an architecture designed for FER that combines spatial and temporal features processed in layers through DSN, DTN, and RNN using LSTM. This process ensures that critical information is preserved and analyzed optimally through multiple stages, ultimately leading to accurate expression predictions. Figure 4 illustrates the complete architecture of this model. In the first stage, the DSN extracts spatial features from the input image. Spatial features such as facial structure, lip contours, and eyebrow positions are identified using convolutional layers (Conv2D), followed by batch normalization and pooling. These features are then passed through a residual block, which plays a crucial role in preserving critical information by applying a shortcut connection that allows the original input x to be passed along with the transformed output $\mathcal{F}(x)$. The basic equation for this residual block operation is in (1). Once all spatial features are extracted, the output from the DSN is forwarded to the next stage, which is combined with the output from the DTN.

$$y = \mathcal{F}(x) + x \tag{1}$$

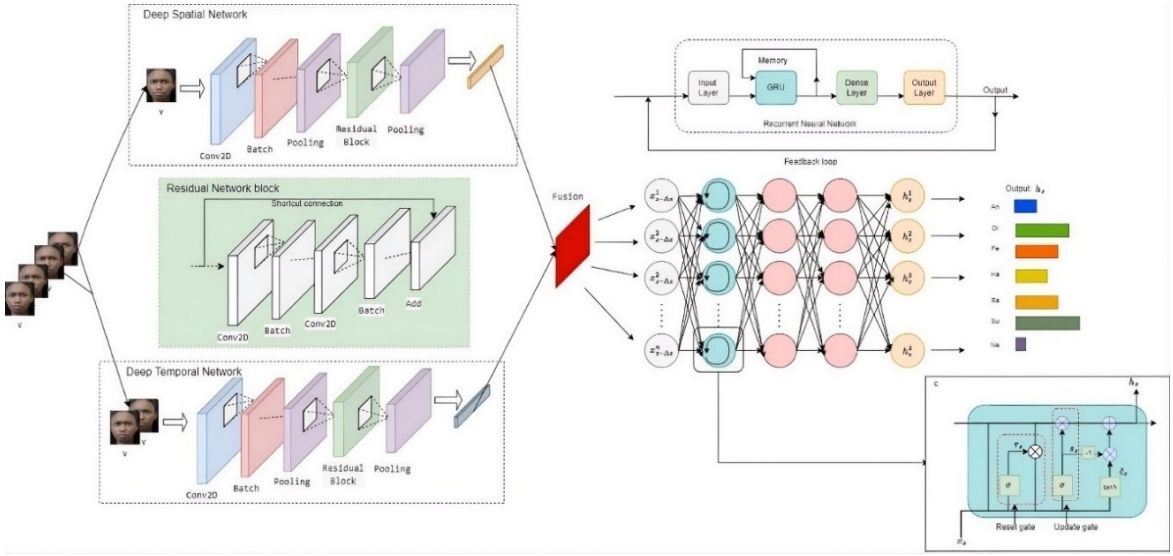

Figure 4. Architecture of deep residual LSTM fusion model

3.3.3. Deep residual gated recurrent unit fusion model

The deep residual recurrent fusion network still integrates spatial and temporal features for FER, but this section uses GRU instead of LSTM for greater efficiency (refer Figure 5). As in the previous model, the DSN extracts spatial features such as facial contours, lip positions, and eyebrow shapes using Conv2D layers, batch normalization, pooling, and residual blocks. The extracted spatial features are then merged with temporal features captured by the DTN, which tracks changes in facial expressions over time. The LSTM model's similarity lies in the fusion of spatial and temporal features through the fusion layer, which produces a richer representation before being passed to the recurrent structure. In this model, however, GRU replaces LSTM to simplify computation while effectively capturing temporal dependencies and ensuring accurate facial expression recognition. The DTN focuses on capturing temporal changes in facial expressions, such as the transition from neutral to smiling or the subtle shifts in eye movements. This network processes each frame using a structure similar to the DSN, involving Conv2D layers, batch normalization, pooling, and residual blocks but emphasizing temporal dynamics. The combined spatial and temporal features are merged in the fusion layer, producing a richer and more comprehensive feature map fed into the GRU layer.

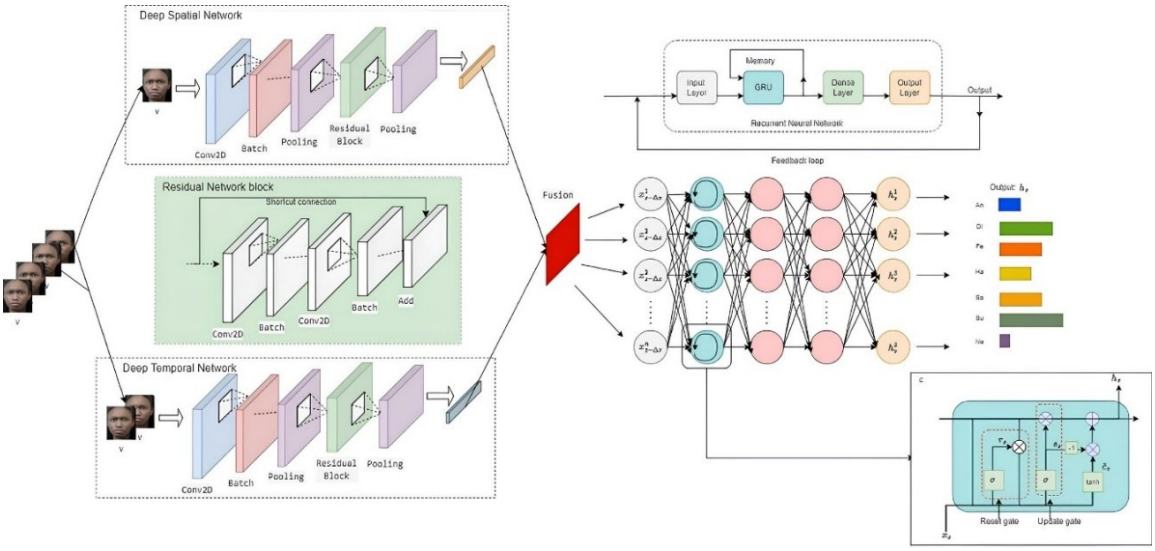

Figure 5. Architecture of deep residual GRU fusion model

3.4. Hyperparameter tuning configuration

Hyperparameters are crucial in improving model performance, requiring a tuning process to find the best configuration. The key steps in hyperparameter tuning include identifying key hyperparameters such as kernel size, filter size, LSTM/BiLSTM/GRU units, batch size, learning rate, dropout rate, and optimizer type as displayed in Table 2. To determine the optimal settings, various configurations are explored by testing multiple hyperparameter value combinations in two scenarios (i.e case I and case II). Performance analysis is conducted based on metrics such as validation accuracy, training loss, and convergence speed, and the results are further refined to enhance training stability and efficiency. This hyperparameter tuning ensures that the deep residual recurrent fusion model balances high accuracy and computational efficiency.

Table 2. Hyperparameter configuration

Parameters	Case I	Case II
Block DTN (kernel)	64	128
Block DTN (filter)	4×4	5×5
Block DSN (kernel)	32	128
Block DSN (filter)	32	128
Block LSTM (units)	100	228
Optimizer Adam	Yes	Yes
Batch size	32	64

3.5. Performance evaluation

The model's performance is evaluated using several key metrics: accuracy, precision, recall, F1-score, and confusion matrix. These metrics are computed as (2) to (5) [4].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{5}$$

Additionally, confusion matrix provides an in-depth view of classification performance by displaying number of correct and incorrect predictions for each class. The confusion matrix is structured as in Table 3 [11].

True positives (TP) occur when the model correctly identifies a positive class, while false positives (FP) indicate incorrect positive predictions. True negatives (TN) reflect correctly identified negative cases, whereas false negatives (FN) represent missed positive cases. To ensure fair benchmarking, the model's performance is compared with recent recurrent network-based models (2017-2023), focusing on models that utilize spatial-temporal fusion, attention mechanisms, and recurrent network structures. These comparisons help validate the effectiveness of the proposed approach in real-world FER tasks.

Table 3. Confusion matrix

Actual/predicted	Positive prediction	Negative prediction
Positive class	TP	FN
Negative class	FP	TN

4. RESULTS AND DISCUSSION

The proposed model was evaluated on the FER2013, FERPlus, RAF-DB, and CK+ datasets. The model performance is measured using accuracy, precision, recall, and F1-score. Table 4 shows the proposed models' performance on the FER2013 dataset. The deep residual GRU model achieved the highest accuracy of 99.86% in both cases, followed closely by the deep residual BiLSTM and the proposed residual LSTM. The second case for deep residual BiLSTM shows a slight decrease in accuracy, indicating a minor sensitivity to hyperparameter changes.

Table 4. Evaluation performance proposed model using FER2013

Model	Case	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Deep residual BiLSTM	I	99.85	99.85	99.85	99.85
	II	99.79	99.79	99.79	99.79
Proposed residual LSTM	I	99.83	99.83	99.83	99.83
	II	99.83	99.83	99.83	99.83
Deep residual GRU	I	99.86	99.86	99.86	99.86
	II	99.86	99.86	99.86	99.86

Table 5 presents the evaluation results on the FERPlus dataset. The deep residual BiLSTM model achieves the highest accuracy of 99.99% in case I, while all models maintain stable and high performance in case II. These results suggest that the residual learning mechanism enhances feature extraction, leading to better classification results.

Table 5. Evaluation performance proposed model using FERPlus

Model	Case	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Deep residual BiLSTM	I	99.99	99.99	99.99	99.99
	II	99.79	99.97	99.97	99.97
Proposed residual LSTM	I	99.96	99.96	99.96	99.96
	II	99.97	99.97	99.97	99.97
Deep residual GRU	I	99.97	99.97	99.97	99.97
	II	99.97	99.97	99.97	99.97

Table 6 shows that all models reached perfect accuracy on the RAF-DB dataset. Similarly, Table 7 confirms the same performance on the CK+ dataset. These results indicate that the models recognize facial expressions in these datasets with zero classification errors. Each model's consistency across both cases demonstrates its robustness in learning and generalizing facial features.

Table 6. Evaluation performance proposed model using RAF-DB

Model	Case	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Deep residual BiLSTM	I	100	100	100	100
	II	100	100	100	100
Proposed residual LSTM	I	100	100	100	100
	II	100	100	100	100
Deep residual GRU	I	100	100	100	100
	II	100	100	100	100

Table 7. Evaluation performance proposed model using CK+

Model	Case	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Deep residual BiLSTM	I	100	100	100	100
	II	100	100	100	100
Proposed residual LSTM	I	100	100	100	100
	II	100	100	100	100
Deep residual GRU	I	100	100	100	100
	II	100	100	100	100

The comparison of model performance across different datasets provides insightful observations. On the FER2013 dataset, the deep residual GRU model performed slightly better than the others, although all models exhibited exceptional accuracy above 99.79%. The deep residual BiLSTM model showed a slight drop in accuracy in case II, suggesting its sensitivity to changes in hyperparameter settings. Meanwhile, the proposed residual LSTM maintained consistent performance across both cases, indicating its robustness in handling variations within the dataset. For the FERPlus dataset, the deep residual BiLSTM model achieved the highest accuracy of 99.99% in case I, surpassing the other models. However, a slight decline was observed in case II, whereas the proposed residual LSTM and deep residual GRU models remained stable. These results suggest that the residual learning mechanism in these models facilitates enhanced feature extraction, contributing to improved recognition capabilities. Regarding the RAF-DB and CK+ datasets, all models attained 100% accuracy in both cases. These results indicate that the models could perfectly classify facial expressions in these structured datasets. The results demonstrate the effectiveness of

residual-based architecture in deep learning models for facial expression recognition. The absence of classification errors suggests that the proposed models effectively capture the essential spatial and temporal patterns required for accurate expression recognition. Overall, the deep residual BiLSTM model excels in the FERPlus dataset, whereas the deep residual GRU model shows slight superiority in FER2013. However, all three models perform exceptionally well, highlighting the effectiveness of integrating residual connections into LSTM and GRU architectures for facial expression recognition. The results confirm that combining spatial and temporal feature extraction mechanisms enhances recognition performance across various facial expression datasets.

Figure 6 presents the summary of the accuracy achieved by the deep residual BiLSTM, proposed residual LSTM, and deep residual GRU models across the FER2013, FERPlus, RAF-DB, and CK+ datasets under case I and case II configurations. The differences observed between the two configurations indicate the models' sensitivity to hyperparameter changes, particularly on the FER2013 and FERPlus datasets, where BiLSTM exhibits a performance decrease in case II, while GRU remains stable. From a practical perspective, the GRU variant performs better due to its simpler and more efficient gating structure, which enables effective temporal dependency modeling with a lower risk of overfitting, whereas BiLSTM is more suitable for scenarios requiring bidirectional context modeling but is more sensitive to suboptimal hyperparameter settings.

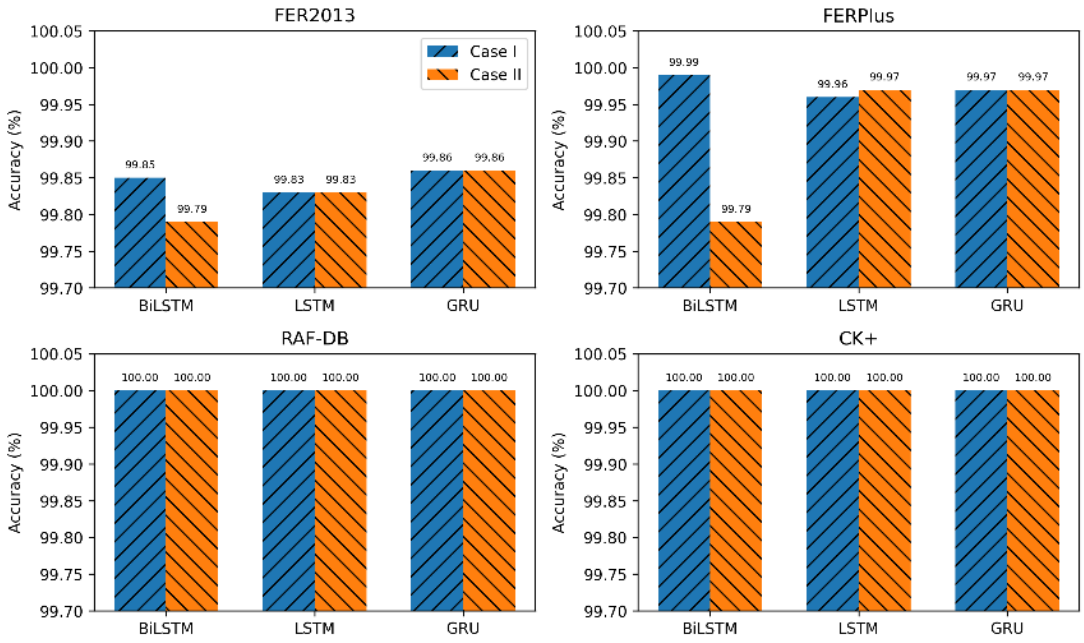

Figure 6. Accuracies of the proposed models

The confusion matrices provide deeper insights into the classification performance of the models. Each confusion matrix represents the percentage of correctly and incorrectly classified emotions. Figure 7 illustrates the confusion matrices for different models. The first confusion matrix as shown in Figure 7(a) demonstrates high accuracy across all emotion categories, with minimal misclassification occurring primarily between similar emotions like fear and surprise or happiness and neutral. The highest classification accuracy is observed for anger (99.72%) and surprise (99.85%), whereas slight misclassifications are found in the happiness and fear classes. The second confusion matrix as shown in Figure 7(b) represents the CNN-based model, which achieves 100% classification accuracy in almost all categories except for neutral (99.92%) and sadness (99.77%), where minor misclassifications are observed. The additional contempt class introduces slight challenges, but the model still performs remarkably well. The third and fourth confusion matrices as shown in Figures 7(c) and 7(d) exhibit perfect classification performance (100%) across all emotion categories, indicating that the models effectively generalize across different facial expressions without errors. This highlights the robustness of the deep residual recurrent fusion model in handling complex FER tasks.

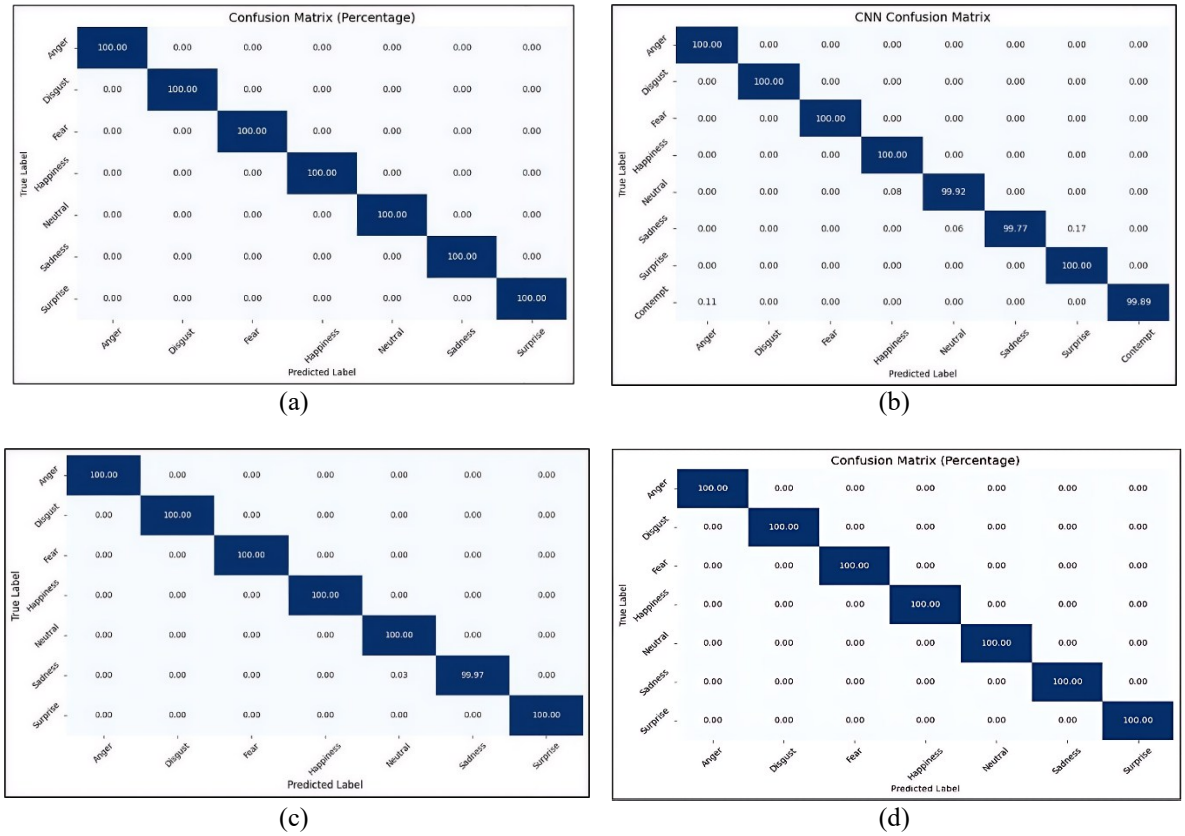

Figure 7. Confusion matrix of proposed models based on various datasets of (a) FER2013, (b) FERPlus, (c) RAF-DB, and (d) CK+

Hyperparameter tuning improves model performance by enabling an appropriate balance between network depth, dropout rate, and learning rate, thereby allowing the feature extraction process to be effective and stable during training. Proper regulation of network depth prevents the model from becoming overly complex, while the use of dropout reduces inter-neuron dependency and an appropriate learning rate ensures optimal convergence. However, this study has limitations due to the relatively small size of the FER datasets, as near-perfect accuracy may increase the risk of overfitting and limit generalization to real-world conditions. Nevertheless, the results demonstrate strong potential for applying FER in practical scenarios, such as emotion-based health monitoring and the development of adaptive learning systems that respond to users' emotional states.

Table 8 presents the performance comparison between the proposed models and existing methods for facial expression recognition. The results show that previous models such as PAtt-Lite [11] achieved 92.50% on FER2013, 95.55% on FERPlus, and 95.05% on RAF-DB, while other approaches like visual transformers with feature fusion (VTFF) [29] and region attention network (RAN) [30] yielded accuracies under 90% for FERPlus and RAF-DB. Compared to these existing methods, the proposed deep residual BiLSTM, proposed residual LSTM, and deep residual GRU models significantly outperform previous works. Specifically, the proposed models achieve near-perfect classification results, reaching 99.85% on FER2013, 99.99% on FERPlus, and 100% on RAF-DB and CK+. These improvements highlight the effectiveness of residual connections in enhancing feature extraction and optimizing learning stability. Moreover, previous models such as a novel spatio-channel attention net (SCAN)-complementary context information (CCI) [31] and TransFER [32] reported lower performance on RAF-DB and FERPlus, with accuracy ranging from 86.90% to 97.31%. In contrast, the proposed models achieve consistent performance across all datasets, eliminating the performance gap observed in earlier methods. The comparison further reveals that even state-of-the-art models like DSN + BiLSTM [4] only reached 99.4% accuracy on CK+, whereas the proposed models achieve 100% accuracy consistently on CK+ and RAF-DB. This consistency suggests that the proposed approach generalizes better across different facial expression datasets and is less susceptible to overfitting or misclassification. These results confirm that integrating deep residual networks with recurrent units (LSTM, GRU, and BiLSTM) enhances the learning capability of deep models, leading to a substantial

performance improvement over existing method. The ability to capture spatial and temporal features effectively contributes to superior accuracy, making the proposed models highly suitable for real-world facial expression recognition applications.

Table 8. Existing models for FER2013, FERPlus, RAF-DB, and CK+

Reference	Dataset	Model	Result
[11]	FER2013, FERPlus, RAF-DB, CK+	PAtt-Lite	FER2013: 92.50%, FERPlus: 95.55%, RAF-DB: 95.05%, CK+: 100%
[29]	FERPlus, RAF-DB	VTFF	FERPlus: 88.81%, RAF-DB: 88.14%
[31]	FERPlus, RAF-DB, CK+	SCAN-CCI	FERPlus: 89.42%, RAF-DB: 89.02%, CK+: 97.31%
[32]	FERPlus, RAF-DB	TransFER	FERPlus: 90.83%, RAF-DB: 90.91%
[4]	CK+	DSN+BiLSTM	99.4%
	CK+	DSN+DTN	98.2%
	CK+	DTN+BiLSTM	93.5%
This work	FER2013	Deep residual BiLSTM	99.85%
		Proposed residual LSTM	99.83%
		Deep residual GRU	99.86%
	CK+	Deep residual BiLSTM	100%
		Proposed residual LSTM	100%
		Deep residual GRU	100%
	RAF-DB	Deep residual BiLSTM	100%
		Proposed residual LSTM	100%
		Deep residual GRU	100%
	FERPlus	Deep residual BiLSTM	99.99%
		Proposed residual LSTM	99.97%
		Deep residual GRU	99.97%

5. CONCLUSION

This study proposes three deep residual recurrent fusion models where tuning was applied to the hyperparameters to enhance the robustness of FER systems. The integration of residual connections with recurrent architectures, namely BiLSTM, LSTM, and GRU, enables effective modeling of spatial and temporal features while reducing vanishing gradient and overfitting issues. Experimental evaluations on the FER2013, FERPlus, RAF-DB, and CK+ datasets demonstrate consistently high performance, confirming the effectiveness of the residual–recurrent fusion approach in improving accuracy and generalization. The main novelty of this work lies in the systematic hyperparameter tuning applied across multiple RNN variants, which reveals differences in performance stability among architectures and identifies GRU as the most robust configuration under varying model complexities. Future research may extend this framework by incorporating multimodal datasets, integrating transformer-based architectures, and adding explainability approaches such as Shapley additive explanations (SHAP) or gradient-weighted class activation mapping (Grad-CAM) to improve interpretability and readiness for real-world deployment.

FUNDING INFORMATION

There is no funding for this study.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muhammad Munsarif	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Ku Ruhana Ku-Mahamud		✓		✓		✓				✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

There is no conflict of interest.

DATA AVAILABILITY

The datasets used in this study, including FER2013, FERPlus, RAF-DB, and CK+, are publicly available and can be accessed through their respective sources. The details and access links for each dataset are as follows:

- RAF-DB at https://drive.google.com/drive/folders/195ncd0QpxOXit5Jtaq3J46-SX8FOWe8Z?usp=sharing.
- FER2013 at https://drive.google.com/drive/folders/1-wjzROwEU_uJ1jYBwtBjmFCj7a4lI1fq.
- FERPlus at https://drive.google.com/drive/folders/1gkkKRtJAPARxOkgdG2CzkIHMl6WJvrMh?usp=sharing.
- CK+ at https://drive.google.com/drive/folders/1htoePSrPTseUL_fMZL08DIDmBnpr4sjR?usp=sharing.

REFERENCES

[1] N. Ahmed, Z. Al Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intelligent Systems with Applications*, vol. 17, Feb. 2023, doi: 10.1016/j.iswa.2022.200171.

[2] M. Sajjad *et al.*, "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," *Alexandria Engineering Journal*, vol. 68, pp. 817–840, Apr. 2023, doi: 10.1016/j.aej.2023.01.017.

[3] Y. Wang *et al.*, "A systematic review on affective computing: emotion models, databases, and recent advances," *Information Fusion*, vol. 83–84, pp. 19–52, Jul. 2022, doi: 10.1016/j.inffus.2022.03.009.

[4] D. Liang, H. Liang, Z. Yu, and Y. Zhang, "Deep convolutional BiLSTM fusion network for facial expression recognition," *The Visual Computer*, vol. 36, no. 3, pp. 499–508, Mar. 2020, doi: 10.1007/s00371-019-01636-3.

[5] X. Jin, J. Xiao, L. Jin, and X. Zhang, "Residual multimodal transformer for expression-EEG fusion continuous emotion recognition," *CAAI Transactions on Intelligence Technology*, vol. 9, no. 5, pp. 1290–1304, Oct. 2024, doi: 10.1049/cit2.12346.

[6] R. Singh, S. Saurav, T. Kumar, R. Saini, A. Vohra, and S. Singh, "Facial expression recognition in videos using hybrid CNN & ConvLSTM," *International Journal of Information Technology*, vol. 15, no. 4, pp. 1819–1830, Apr. 2023, doi: 10.1007/s41870-023-01183-0.

[7] J. Wang and Z. Zhang, "Facial expression recognition in online course using light-weight vision transformer via knowledge distillation," in *PRICAI 2023: Trends in Artificial Intelligence*, Singapore: Springer, 2024, pp. 247–253, doi: 10.1007/978-981-99-7025-4_22.

[8] M. T. Vu, M. B.-Aimar, and S. Marchand, "Multitask multi-database emotion recognition," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2021, pp. 3630–3637, doi: 10.1109/ICCVW54120.2021.00406.

[9] I. Bah and Y. Xue, "Facial expression recognition using adapted residual based deep neural network," *Intelligence & Robotics*, vol. 2, no. 1, pp. 78–88, 2022, doi: 10.20517/ir.2021.16.

[10] Y. Gao and C. Xu, "Residual learning with Bi-LSTM and multi-head attention for multi-modal emotion recognition," in *2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA)*, Aug. 2023, pp. 409–413, doi: 10.1109/ICIPCA59209.2023.10257779.

[11] J. L. Ngwe, K. M. Lim, C. P. Lee, T. S. Ong, and A. Alqahtani, "PAtt-Lite: lightweight patch and attention MobileNet for challenging facial expression recognition," *IEEE Access*, vol. 12, pp. 79327–79341, 2024, doi: 10.1109/ACCESS.2024.3407108.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[13] S. Lee, J. Kim, H. Kang, D.-Y. Kang, and J. Park, "Genetic algorithm based deep learning neural network structure and hyperparameter optimization," *Applied Sciences*, vol. 11, no. 2, Jan. 2021, doi: 10.3390/app11020744.

[14] N. M. Aszemi and P. D. D. Dominic, "Hyperparameter optimization in convolutional neural network using genetic algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, 2019, doi: 10.14569/IJACSA.2019.0100638.

[15] M. Munsarif, E. Noersasongko, P. N. Andono, and M. A. Soeleman, "Improving convolutional neural network based on hyperparameter optimization using variable length genetic algorithm for english digit handwritten recognition," *International Journal of Advances in Intelligent Informatics*, vol. 9, no. 1, pp. 66–78, Mar. 2023, doi: 10.26555/ijain.v9i1.881.

[16] G. Atteia, A. Alhussan, and N. Samee, "BO-ALLCNN: Bayesian-based optimized CNN for acute lymphoblastic leukemia detection in microscopic blood smear images," *Sensors*, vol. 22, no. 15, Jul. 2022, doi: 10.3390/s22155520.

[17] M. A. Amou, K. Xia, S. Kamhi, and M. Mouhafid, "A novel MRI diagnosis method for brain tumor classification based on CNN and Bayesian optimization," *Healthcare*, vol. 10, no. 3, Mar. 2022, doi: 10.3390/healthcare10030494.

[18] A. F. Chavarro, D. Renza, and D. M. Ballesteros, "Influence of hyperparameters in deep learning models for coffee rust detection," *Applied Sciences*, vol. 13, no. 7, Apr. 2023, doi: 10.3390/app13074565.

[19] D. Passos and P. Mishra, "A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks," *Chemometrics and Intelligent Laboratory Systems*, vol. 223, Apr. 2022, doi: 10.1016/j.chemolab.2022.104520.

[20] M. A. K. Raiaan *et al.*, "A systematic review of hyperparameter optimization techniques in convolutional neural networks," *Decision Analytics Journal*, vol. 11, Jun. 2024, doi: 10.1016/j.dajour.2024.100470.

[21] M. A. Judge, V. Franzitta, D. Curto, A. Guercio, G. Cirrincione, and H. A. Khattak, "A comprehensive review of artificial intelligence approaches for smart grid integration and optimization," *Energy Conversion and Management: X*, vol. 24, Oct. 2024, doi: 10.1016/j.ecmx.2024.100724.

[22] M. Wojciuk, Z. S.-Chadaj, K. Siwek, and A. Gertych, "Improving classification accuracy of fine-tuned CNN models: Impact of hyperparameter optimization," *Heliyon*, vol. 10, no. 5, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26586.

[23] P. Bachhal, V. Kukreja, and S. Ahuja, "Real-time disease detection system for maize plants using deep convolutional neural networks," *International Journal of Computing and Digital Systems*, vol. 14, no. 1, pp. 10263–10275, Oct. 2023, doi: 10.12785/ijcds/140199.

[24] J. Fu, J. Liu, R. Zhao, Z. Chen, Y. Qiao, and D. Li, "Maize disease detection based on spectral recovery from RGB images," *Frontiers in Plant Science*, vol. 13, Dec. 2022, doi: 10.3389/fpls.2022.1056842.

[25] H. Mubarak *et al.*, "Day-ahead electricity price forecasting using a CNN-BiLSTM model in conjunction with autoregressive modeling and hyperparameter optimization," *International Journal of Electrical Power & Energy Systems*, vol. 161, Oct. 2024, doi: 10.1016/j.ijepes.2024.110206.

[26] Y. Jia, Y. Shi, J. Luo, and H. Sun, "Y–Net: identification of typical diseases of corn leaves using a 3D–2D hybrid CNN model combined with a hyperspectral image band selection module," *Sensors*, vol. 23, no. 3, Jan. 2023, doi: 10.3390/s23031494.
[27] D. Vidyabharathi and V. Mohanraj, "Hyperparameter tuning for deep neural networks based optimization algorithm," *Intelligent Automation & Soft Computing*, vol. 36, no. 3, pp. 2559–2573, 2023, doi: 10.32604/iasc.2023.032255.
[28] N. Saeedizadeh, S. M. J. Jalali, B. Khan, P. M. Kebria, and S. Mohamed, "A new optimization approach based on neural architecture search to enhance deep U-Net for efficient road segmentation," *Knowledge-Based Systems*, vol. 296, Jul. 2024, doi: 10.1016/j.knosys.2024.111966.
[29] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1236–1248, Apr. 2023, doi: 10.1109/TAFFC.2021.3122146.
[30] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020, doi: 10.1109/TIP.2019.2956143.
[31] D. Gera and S. Balasubramanian, "Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition," *Pattern Recognition Letters*, vol. 145, pp. 58–66, May 2021, doi: 10.1016/j.patrec.2021.01.029.
[32] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognition Letters*, vol. 119, pp. 49–61, Mar. 2019, doi: 10.1016/j.patrec.2017.10.022.

BIOGRAPHIES OF AUTHORS

Muhammad Munsarif received his master's degree in Computer Science from Dian Nuswantoro University, Semarang, Indonesia, and completed his Ph.D. in Information and Communication Technology from Asia e-University, Selangor, Malaysia. He is currently a lecturer in the Computer Science Study Program at Universitas Muhammadiyah Semarang. His research interests include computer vision, artificial intelligence, and data science. He can be contacted at email: m.munsarif@unimus.ac.id.

Ku Ruhana Ku-Mahamud holds a bachelor's degree in Mathematical Sciences and a masters degree in Computing. Her Ph.D. degree is in Computer Science. Her research interests include ant colony optimization, pattern classification, and vehicle routing problem. Currently, she is attached to Universiti Utara Malaysia as an emeritus professor and Universiti Muhammadiyah Malaysia as professor. She can be contacted at email: ruhana@uum.edu.my.