Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/100/2025.20426

# *Enhancing Facial Expression Recognition with Robust CNN Architectures and Adaptive Preprocessing Techniques*

**Ao Guo**[1,a,*]

[1]*Cockrell School of Engineering, The University of Texas at Austin, Austin, Texas, United States of America*
*a. davidguo@utexas.edu*
**corresponding author*

***Abstract:*** Facial expression recognition (FER) is an essential technology at the intersection of artificial intelligence (AI), computer vision, and psychology. This study proposes a novel framework for FER, aiming to improve system robustness and generalization, especially under variable real-world conditions. Using the FER2013 dataset, this research combines an adaptive preprocessing pipeline with a custom Convolutional Neural Network (CNN) architecture. Key preprocessing steps include normalization, rotation, and flipping to improve data quality and diversity. The CNN architecture combines regularization methods, including dropout and L2 regularization. Dynamic hyperparameter tuning and early stopping optimize performance and prevent overfitting. Normalized confusion matrix indicating strong recognition for well-represented emotions, such as happiness with 86% accuracy, and challenges with underrepresented categories like disgust. This research aims to contribute to the ongoing development of facial expression recognition systems by enhancing their robustness and generalization. While further refinement is needed, this work provides a step toward more accurate and adaptable FER models, with the potential to support advancements in human-computer interaction and various real-world applications.

***Keywords:*** Facial Expression Recognition, Convolutional Neural Networks, Adaptive Preprocessing, Regularization Techniques, Emotion Classification.

## 1. Introduction

In today's digital epoch, the ability to understand and interpret human emotions has become increasingly crucial for technological advancement. Facial expression recognition (FER) is the intersection of artificial intelligence (AI), computer vision, and psychology. It is a critical tool in diverse applications ranging from human-computer interaction to mental health diagnosis [1, 2]. FER encourages the recognition of emotions in healthcare and driver safety applications while also finding use in personalized marketing and surveillance systems [3, 4]. By analyzing facial muscle movements, this technology promotes more intuitive and emotionally aware interactions, potentially bridging human-machine communication gaps [3].

Facial expressions represent a fundamental aspect of human communication, with certain basic expressions like happiness, anger, and sadness being universally recognized [2]. This universality was first documented in psychological studies by Ekman and Friesen. This universality provides the theoretical foundation for automated FER systems [5]. However, while these expressions are

DOI: 10.54254/2755-2721/100/2025.20426

biologically rooted, cultural and individual differences present variability that complicates their analysis [2, 5]. Additionally, real-world applications face challenges such as occlusions, lighting variations, and pose changes, all of which impact system performance [4][6]. These issues prevent the creation of systems that can generalize effective results beyond controlled environments.

Previous research has explored both traditional machine learning methods and deep learning approaches to address these challenges. Early FER systems relied on handcrafted feature extraction techniques such as Local Binary Patterns (LBP), Optical Flow, and Gabor filters [5, 6]. While effective in static and controlled settings, these methods often needed help with the variability inherent in dynamic, real-world environments [4]. With the advent of deep learning, Convolutional Neural Networks (CNN) have become a cornerstone for FER development. By automatically learning hierarchical features directly from raw image data, CNN-based models have demonstrated remarkable improvements in accuracy and robustness [1, 3]. These methods have been successfully applied to large-scale datasets like FER2013, comprising over 35,000 grayscale images across seven emotional categories [4].

While deep learning approaches have advanced the field considerably, several fundamental challenges persist in FER development. Overfitting due to limited labeled data, data imbalance, and variability in human expressions remain significant barriers to achieving robust system performance [1, 5, 6]. To address these limitations, researchers have employed strategies such as data augmentation, dropout layers, and normalization, which have shown promise in mitigating these issues [3, 4]. Normalization plays a crucial role by ensuring feature comparability across samples, improving numerical stability during training, and facilitating faster model convergence [7]. Nevertheless, there remains a pressing need for innovative methodologies to enhance the generalizability and applicability of FER systems in diverse real-world scenarios.

This study offers FER by developing a novel deep learning framework that explores three critical challenges: overfitting in large-scale datasets, class imbalance among emotion categories, and performance degradation under variable real-world conditions.

The approach introduces several critical enhancements to advanced FER systems:

(1) an adaptive preprocessing pipeline that enhances data quality and feature diversity through techniques like normalization, rotation, and flipping.

(2) an optimized CNN architecture with advanced regularization methods such as dropout and L2 regularization.

(3) an efficient training strategy featuring dynamic hyperparameter tuning and early stopping. These strategies aspire to enhance classification accuracy and computational efficiency.

## 2. DataSets

### 2.1. Data collection and description

The dataset used in this study is the FER2013 dataset, a widely recognized benchmark for FER tasks. Figure 1 shows 32,298 grayscale images of human faces, each resized to 48x48 pixels, distributed across seven emotional categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset is divided into training (28,709 images) and test (3,589 images) subsets.

DOI: 10.54254/2755-2721/100/2025.20426

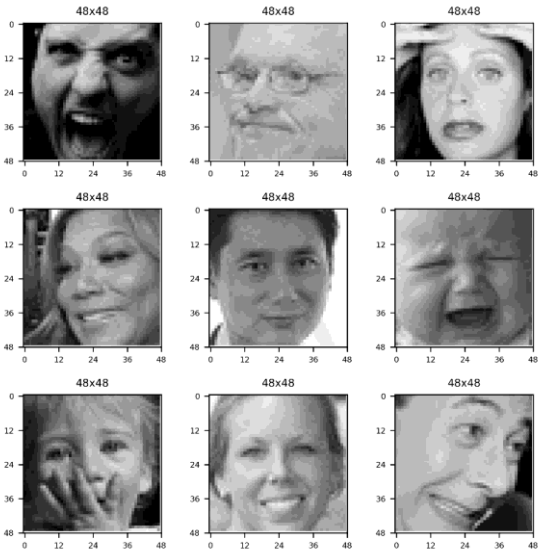


Figure 1: Sample Images from the FER2013 Dataset.

FER2013 provides the scale and diversity needed for training deep learning models in FER. Its pose, lighting, and demographic feature variability offer a challenging standard for developing robust and generalizable FER systems. Additionally, its comprehensive representation of emotional states aligns with the goal of addressing real-world variability and enhancing model performance across diverse scenarios. However, as shown in Figure 2, the dataset exhibits notable class imbalance, with happiness accounting for approximately 25% of the training samples and disgust comprising only about 1.5%, offering a significant challenge for model training.

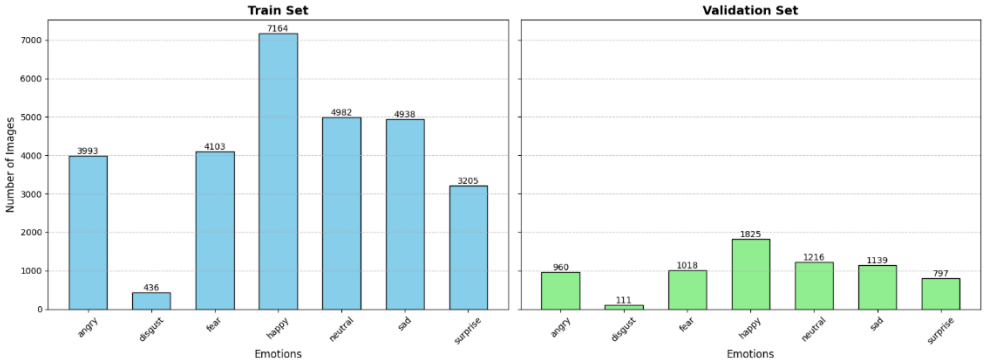


Figure 2: Emotion class distribution in the FER2013 dataset for train and validation sets.

## 2.2. Data Pre-processing

Effective data preprocessing is essential for ensuring the quality and consistency of inputs in FER tasks, where noise and variability can significantly impact model performance. All images were converted to grayscale and resized to a uniform dimension of 48x48 pixels, reducing computational complexity while preserving essential features for emotion recognition. Pixel intensity normalization followed, rescaling all pixel values to the range [0,1]. The raw pixel intensities are scaled consistently across the dataset through this normalization process, reducing the influence of varying brightness levels or contrast in the images. Furthermore, data augmentation was applied to address the class imbalance and increase training data diversity, using random rotations, horizontal flipping, width and

DOI: 10.54254/2755-2721/100/2025.20426

height shifts, and zooming techniques. These transformations simulate real-world variations, such as changes in orientation and perspective, improving the model's ability to generalize to unseen data.

## 2.3. Model selection

Among various machine learning approaches, CNN has emerged as a leading solution for the FER problem due to its ability to automatically learn hierarchical features directly from raw image data, eliminating manual feature extraction, and showcasing remarkable robustness and adaptability across diverse datasets and conditions [8]. This study employs a customized CNN architecture tailored to address FER challenges, incorporating techniques to enhance performance and generalizability. The overall structure of the proposed architecture is illustrated in Figure 3.

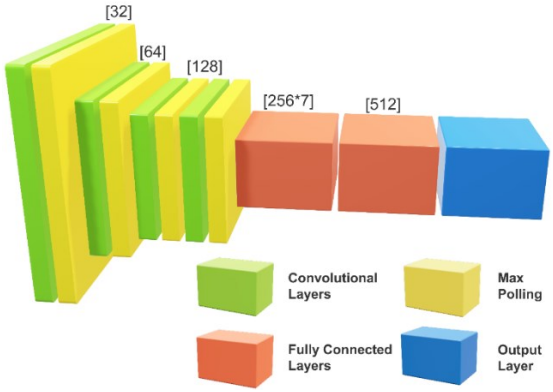


Figure 3: Proposed CNN architecture for facial expression recognition.

The architecture begins with an input layer designed for grayscale images with dimensions of 48x48 pixels. This size is computationally efficient while retaining critical details necessary for emotion classification. Following the input layer, the model employs a series of convolutional layers, each coupled with batch normalization, ReLU activation, and max-pooling. Batch normalization, combined with dropout layers strategically placed throughout the network, helps to reduce overfitting and ensures robust generalization to unseen data [9]. This sequential structure allows the network to progressively extract increasingly complex features, from simple edges and textures to more abstract patterns representing facial expressions. The convolutional layers are configured with increasing filters 32, 64, and 128 across successive layers. Each layer uses a kernel size of 3x3, a standard choice for balancing feature granularity and computational efficiency. Batch normalization is applied after each convolution to stabilize learning by normalizing the activations, while ReLU activation introduces non-linearity, enabling the model to capture intricate features. Max-pooling layers reduce spatial dimensions, minimizing computational requirements while maintaining essential features. In the final convolutional layer, a dropout layer with a rate of 0.5 is applied, further reducing overfitting by randomly deactivating neurons during training. Following the feature extraction phase, the architecture incorporates two fully connected (FC) layers with 256*7 and 512 neurons, respectively. Both layers are equipped with dropout layers with a rate of 0.5, randomly deactivating a fraction of neurons during training. L2 regularization penalizes large weights, enhancing the model's robustness. The first FC layer transforms the flattened feature map from the final convolutional layer into a dense representation with 256*7 dimensions, capturing relationships across all extracted features. The second FC layer further processes these features, refining the abstraction and enabling effective classification before the output layer. The final layer in the network is a dense layer with seven neurons, corresponding to the seven emotional categories in the FER2013 dataset. A softmax

DOI: 10.54254/2755-2721/100/2025.20426

activation function normalizes the outputs, creating a probability distribution over the emotion classes, which allows the model to make accurate predictions.

The proposed CNN training protocol focused on optimizing performance and generalizability. The Adam optimizer was employed with an initial learning rate of 0.001. A batch size of 128 was used to balance memory efficiency and training speed. Early stopping was implemented to monitor validation accuracy, halting training when no improvement was observed for ten epochs. Model checkpointing saved the best-performing weights during training, confirming the retention of optimal parameters. The training was conducted for a maximum of 100 epochs, with all epochs evaluated against the validation set to track generalization performance.

## 3. Results and Discussion

### 3.1. Experimental Setup

The experimental setup employed in this study was devised to evaluate the proposed CNN performance for FER rigorously. This process included carefully selecting optimizers, hyperparameters, and evaluation strategies, which were underpinned by findings from comparative studies on optimization algorithms. Optimization algorithms are vital for achieving efficient and precise model training. In this study, the Adam optimizer was selected for its proven effectiveness in addressing challenges such as sparse gradients and non-stationary objectives. A comparative study on optimization techniques demonstrated that Adam consistently outperforms other algorithms, including Stochastic Gradient Descent (SGD) and Root Mean Square Propagation (RMSProp), in terms of training speed and accuracy across various computer vision tasks [10]. Grid search experiments were conducted to identify the optimal hyperparameters for training by varying the learning rate, batch size, and dropout rate. Learning rates were tested at [0.001, 0.0005, 0.0001] to find a balance between stable convergence and efficient learning. Batch sizes of [32, 64, 128, 256] were evaluated to identify a configuration that maximized performance while remaining computationally feasible. Similarly, dropout rates of 0.5 and 0.25 were explored to determine the most effective approach for mitigating overfitting without impairing the network's learning capacity. The combination of a 0.0001 learning rate, a batch size of 128, and a dropout rate of 0.5 appeared as the optimal configuration through these experiments. Early stopping was applied, halting training after ten consecutive epochs with no improvement in validation accuracy. This strategy addresses overfitting by regulating the number of epochs based on performance trends and ensuring computational efficiency [11]. Early stopping is superior in maintaining convergence with minimal error rates, making it an indispensable component of regularization strategies in deep learning. Additionally, model checkpointing saved the best-performing model weights based on validation accuracy. The model was implemented using TensorFlow and trained on a GPU-equipped system to accelerate computations. Metrics, including training loss, validation loss, and accuracy, were recorded after each epoch to evaluate model performance and ensure convergence. The experimental protocol also included visualizations of loss and accuracy trends, providing senses into training and allowing for informed adjustments to hyperparameters when necessary.

### 3.2. Performance Metrics

In deep learning, performance metrics are essential for evaluating model efficiency, convergence, and generalization. These metrics provide a quantitative foundation for assessing a model's suitability for a given task. For FER, the primary metrics include training loss, validation loss, training accuracy, and validation accuracy. The proposed CNN architecture achieved exemplary performance in these metrics using a configuration of a learning rate of 0.0001, a batch size of 128, and a dropout rate of 0.5. Training loss rapidly declined during the initial epochs and plateaued at a low value, indicating

DOI: 10.54254/2755-2721/100/2025.20426

effective learning. Validation loss mirrored this behavior, stabilizing without significant oscillations, which is critical for demonstrating convergence and robustness. By the 25th epoch, the validation loss had leveled off. The accuracy metrics underline the strength of this configuration. Training accuracy rapidly increased to exceed 55% within the first ten epochs and continued to improve, reaching 65% by the 40th epoch. Validation accuracy, a vital indicator of the model's real-world applicability, followed a parallel trend, stabilizing at approximately 62%. This consistent alignment between training and validation metrics reflects the model's ability to generalize effectively. The selected configuration provided a well-balanced approach, achieving robust convergence and generalization, highlighting its suitability for FER tasks. Figure 4 shows the training and validation loss and accuracy trends for the model.

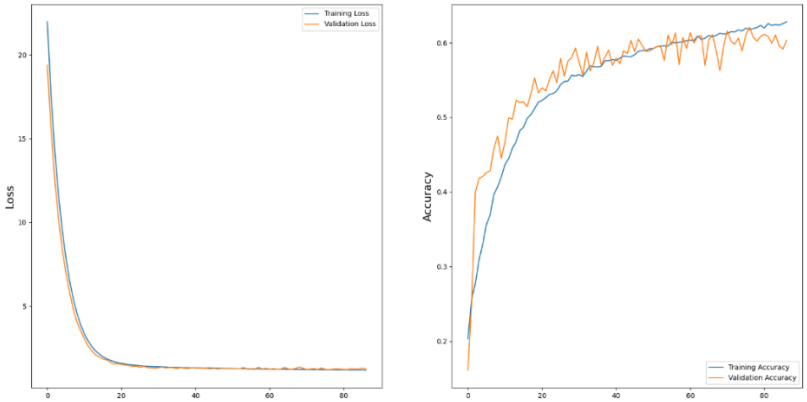


Figure 4: Training and validation loss and accuracy trends for the proposed CNN architecture.

Figure 5 shows that the normalized confusion matrix provides detailed insights into the model's classification performance across the seven emotional categories. This matrix highlights the strengths and areas for improvement in the model's ability to generalize across diverse facial expressions and serves as a comprehensive tool for evaluating classification metrics. Confusion matrices are significant in visualizing the performance of classifiers by offering a clear breakdown of correct. The matrix reveals that the model performed exceptionally well in recognizing the emotion of happiness, achieving a classification accuracy of 86% for this category. This result aligns with the observation that the FER2013 dataset contains a relatively balanced representation of the "happy" class, making it easier for the model to generalize. Similarly, the model exhibited a strong performance for surprise, with an accuracy of 66%, and neutral, with an accuracy of 69%, demonstrating its ability to handle classes with less ambiguity and moderate representation. However, the matrix also highlights challenges in recognizing certain emotions. For example, disgust was the most difficult category to classify, with an accuracy of only 18%. A significant proportion of "disgust" samples were misclassified as angry (48%), reflecting the similarity in facial features associated with these emotions, as well as the dataset's severe imbalance for the "disgust" class. Similarly, fear and sadness achieved lower accuracies of 26% and 55%, with frequent misclassifications into adjacent categories such as angry and neutral. These results suggest that while the model effectively distinguishes well-represented and visually distinct emotions, it needs help with underrepresented and visually overlapping categories.

DOI: 10.54254/2755-2721/100/2025.20426

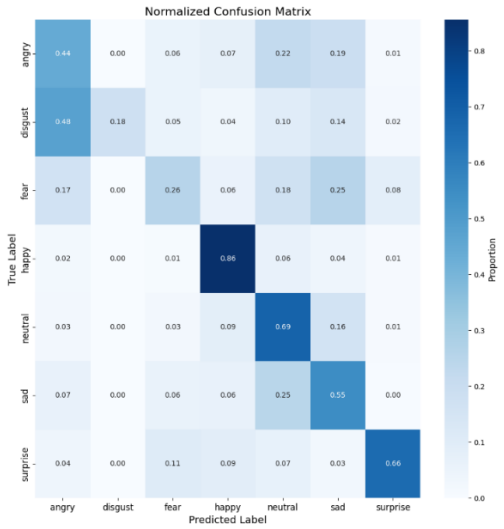


Figure 5: Normalized confusion matrix for the proposed CNN model's classification performance.

## 3.3. Analysis of Results

This section provides a detailed analysis of the impact of hyperparameter choices on the performance of the CNN model for facial expression recognition. By systematically varying the learning rate, batch size, and dropout rate, the study identifies the trade-offs and advantages of each configuration, supported by quantitative comparisons. The analysis is structured into three subsections: the impact of learning rate, batch size effects, and dropout rate analysis.

### 3.3.1. Learning Rate Impact

The configuration with a learning rate of 0.001 demonstrated a rapid decrease in training and validation loss during the initial epochs, indicative of accelerated convergence. However, as training progressed, significant oscillations in both metrics were observed. These fluctuations suggest that the higher learning rate caused the optimizer to overshoot local minima, leading to unstable updates and suboptimal performance in later epochs. In contrast, the slower learning rate of 0.0001 facilitated a smoother and more consistent convergence, avoiding erratic updates and ensuring superior performance stability.

### 3.3.2. Batch Size Effects

Similarly, the choice of batch size had a notable impact on the efficiency and accuracy of the model. Larger batch sizes, such as 256, demonstrated more stable learning but at the cost of reduced accuracy due to insufficient gradient variability. The batch size 128 provided a practical middle ground, balancing computational efficiency with accurate gradient estimation, promoting the model to achieve high validation accuracy while maintaining smooth training curves.

### 3.3.3. Dropout Analysis

The dropout rate also played a pivotal role in determining the model's generalization capabilities. Configurations with lower dropout rates, such as 0.3, showed reduced regularization effects, leading to overfitting. Higher dropout rates excessively hindered learning by deactivating too many neurons, resulting in underfitting and lower accuracy metrics. The dropout rate of 0.5 struck an ideal balance, mitigating overfitting while retaining the model's capacity to learn complex features. Beyond these

DOI: 10.54254/2755-2721/100/2025.20426

individual hyperparameters, their combined effect shapes the model's overall performance. For example, while a learning rate of 0.0005 showed promise with smoother training curves, its combination with lower dropout rates of 0.25 yielded suboptimal validation accuracy, peaking at 58%. Highlights the importance of carefully balancing hyperparameters to achieve robust generalization and high performance. This study systematically analyzes performance across different settings, highlighting the critical role of hyperparameter tuning in deep learning models. It establishes the optimal configuration as a benchmark for advancing FER systems. These findings emphasize the need for balanced, data-driven parameter selection in developing scalable and effective AI solutions.

## 4. Conclusion

This study proposed an alternative approach to improving critical FER challenges, including overfitting, class imbalance, and variability in real-world conditions. The model achieved significant progress in robustness and generalization by introducing a novel framework incorporating advanced preprocessing techniques, a custom CNN architecture, and dynamic hyperparameter optimization. The optimal configuration demonstrated strong performance, achieving a validation accuracy of 62% with consistent alignment between training and validation metrics. The normalized confusion matrix further highlighted the model's strengths in accurately identifying well-represented emotions, such as happiness and surprise, while identifying areas of improvement for underrepresented or overlapping categories like disgust and fear.

This study provides an alternative pathway for advancing FER systems by combining architectural optimization with strategic preprocessing and regularization methods. This approach offers a flexible and scalable solution for improving FER performance in real-world applications, particularly in healthcare, human-computer interaction, and education, where accurate emotion recognition is critical. The methodology outlined here underscores the potential for systematic tuning and integration of advanced techniques to enhance the overall reliability and effectiveness of FER systems. Nevertheless, limitations remain. The class imbalance inherent in the FER2013 dataset and the model's difficulty distinguishing visually similar expressions highlight areas for further exploration. Future work could mitigate these challenges by incorporating advanced data augmentation, attention mechanisms, or ensemble learning strategies. By building upon this study's insights, future research can refine and extend this alternative approach to create more inclusive, adaptive, and accurate FER solutions.

## References

[1] *Kopalidis, T., Solachidis, V., Vretos, N., et al. (2024). Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets. Information, 15(3), 135.*

[2] *Ali, G., Ali, A., Ali, F., Draz, U., Majeed, F., Yasin, S., ... & Haider, N. (2020). Artificial neural network based ensemble approach for multicultural facial expressions analysis. Ieee Access, 8, 134950-134963.*

[3] *Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. IEEE transactions on affective computing, 13(3), 1195-1215.*

[4] *Bhagat, D., Vakil, A., Gupta, R. K., & Kumar, A. (2024). Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN). Procedia Computer Science, 235, 2079-2089.*

[5] *Canedo, D., & Neves, A. J. (2019). Facial expression recognition using computer vision: A systematic review. Applied Sciences, 9(21), 4678.*

[6] *Sajjad, M., Ullah, F. U. M., Ullah, M., Christodoulou, G., Cheikh, F. A., Hijji, M., ... & Rodrigues, J. J. (2023). A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. Alexandria Engineering Journal, 68, 817-840.*

[7] *Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1), 91-99.*

DOI: 10.54254/2755-2721/100/2025.20426

[8] *Vyas, A. S., Prajapati, H. B., & Dabhi, V. K. (2019, March). Survey on face expression recognition using CNN. In 2019 5th international conference on advanced computing & communication systems (ICACCS) (pp. 102-106). IEEE.*

[9] *Ogundokun, R. O., Maskeliunas, R., Misra, S., & Damaševičius, R. (2022, July). Improved CNN based on batch normalization and adam optimizer. In International Conference on Computational Science and Its Applications (pp. 593-604). Cham: Springer International Publishing.*

[10] *Hassan, E., Shams, M. Y., Hikal, N. A., & Elmougy, S. (2023). The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study. Multimedia Tools and Applications, 82(11), 16591-16633.*

[11] *Sitaula, C., & Ghimire, N. (2017). An analysis of early stopping and dropout regularization in deep learning. International Journal of Conceptions on Computing and Information Technology, 5(1), 17-20.*