Journal of Information Technology and Computer Science
Volume 11, Number 2, August 2026, pp. 236-254
Journal Homepage: www.jitecs.ub.ac.id

Multi-Level Ensemble Learning for Facial Expression Recognition on Imbalanced FER2013 Dataset

Muhammad Fajarivan Pratama*[1], Wayan Firdaus Mahmudy[2], Lailil Muflikhah[3]

[1,2,3]Universitas Brawijaya, Malang
[1]mfajarivanp@student.ub.ac.id, [2]wayanfm@ub.ac.id, [3]lailil@ub.ac.id
*Corresponding Author

Received 16 May 2026; accepted 14 June 2026

Abstract. Facial expression recognition remains a challenging task in computer vision, particularly due to class imbalance in datasets such as FER2013, where the Happy class accounts for 25.05% and Disgust only 1.52%, leading to biased predictions. This study proposes a Multi-Level Ensemble approach that integrates data diversity (bootstrap sampling in ELM), model diversity (CNN and CNN-ELM), and classifier diversity (ELM with different random seeds). The method is evaluated using Stratified 5-Fold Cross-Validation on 35,887 FER2013 images. Results show that pure ELM achieves 36.37% accuracy, CNN baseline 66.90%, CNN-ELM 67.28%, and ELM ensemble 67.34%. The proposed method achieves the best performance at 68.23%, improving the CNN baseline by +1.33%. Diversity analysis reports a disagreement rate of 21.2%, Q-statistic of 0.9313, and double-fault of 25.5%. These results indicate that the proposed framework effectively improves FER performance under class imbalance conditions.

Keywords: Facial Expression Recognition, Ensemble Machine Learning, Convolutional Neural Network, Extreme Learning Machine, FER2013

1 Introduction

Facial expressions play a crucial role in human interaction as a primary form of nonverbal communication [1]. Automatic facial expression recognition (FER) has attracted significant attention due to its wide applications in human-computer interaction, security systems, and emotion analysis in healthcare, mental health, and education [2]. Despite this progress, FER remains a challenging task due to variations across individuals (e.g., age, ethnicity, and gender), differences in expression intensity, inconsistent lighting conditions, and occlusions such as glasses or masks, all of which complicate data distribution and model learning [3].

The FER2013 dataset introduced by Goodfellow et al. [4] exemplifies these challenges, particularly in terms of class imbalance. For instance, the Happy class dominates with 8,989 samples (25.05%), while Disgust contains only 547 samples (1.52%). This imbalance can bias models toward majority classes and degrade performance on minority classes. In addition, ambiguity between similar expressions, such as Fear vs. Surprise and Sad vs. Neutral, further complicates classification. These issues have been widely recognized as key factors affecting FER performance [5], [6].

Recent advances in deep learning, particularly Convolutional Neural Network (CNN), have significantly improved FER performance by enabling automatic extraction of hierarchical and discriminative features from images [7], [8], [9].

Architectures such as VGGNet, ResNet, and Inception have been widely adopted and demonstrate strong capability in capturing complex visual patterns [10], [11], [12]. He [13] proposed a lightweight, multi-branch CNN architecture to enhance feature extraction capabilities in FER, while Chen et al. [14] developed a lightweight CNN model utilizing depthwise separable convolution to reduce computational complexity. However, single CNN models still struggle to simultaneously capture global and local features and to distinguish fine-grained expressions [15].

On the other hand, Extreme Learning Machine (ELM) [16], [17] offers highly efficient training by relying on a single-step Moore-Penrose pseudoinverse without iterative backpropagation. This enables significantly faster training compared to conventional methods. Prior studies have demonstrated the effectiveness of ELM in face-related tasks [18], [19]. However, its performance is highly dependent on feature quality, and previous work suggests that combining ELM with deep features and ensemble techniques can improve generalization [20], [21].

Ensemble learning combines multiple base learners to improve predictive performance [22]. In this study, a Multi-Level Ensemble approach is proposed to integrate the feature extraction capability of CNN with the efficiency of ELM. Unlike conventional ensemble methods that rely on homogeneous models and simple averaging, the proposed approach explicitly leverages diversity at three levels: (1) data diversity via bootstrap sampling in ELM, (2) model diversity through the combination of CNN and CNN-ELM, and (3) classifier diversity through random parameter variations in ELM [23], [24], [25]. This design aims to enhance generalization while addressing class imbalance in FER2013.

The main contribution of this study is the development of a Multi-Level Ensemble framework for facial expression recognition that integrates three levels of diversity (data, classifier, and model diversity) within a unified architecture. This contribution is supported by: (a) quantitative diversity analysis using disagreement, Q-statistic, and double-fault metrics; (b) statistical validation using paired t-test ($p < 0.05$); and (c) a fair evaluation protocol based on Stratified 5-Fold Cross-Validation applied consistently across all models [26], [27].

2 Related Works

2.1 Facial Expression Recognition (FER)

FER aims to classify human facial expressions into basic emotion categories, including Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral [1], [2]. FER has been widely studied due to its relevance in various real-world applications. However, it remains a challenging task due to variations in facial appearance and environmental conditions. And one benchmark dataset frequently used in FER research is FER2013, that introduced by Goodfellow et al. [4]. The dataset consists of 35,887 grayscale images with a resolution of 48×48 pixels, categorized into seven emotion classes. The images were collected from the web under diverse conditions, including variations in pose, illumination, and occlusions, making the dataset particularly challenging for model generalization.

A key limitation of FER2013 is its highly imbalanced class distribution, which significantly affects model performance. As shown in Table 1, the Happy class dominates the dataset at 25.05%, whereas the Disgust class is severely underrepresented at 1.52%. This imbalance can lead to biased learning, where models tend to favor majority classes while underperforming on minority classes. Such challenges have been consistently reported in previous studies as a major factor influencing FER performance [5], [6]

Table 1. Distribution of the FER2013 Dataset

Label	Expression	Samples	Percentage
0	Angry	4,953	13.80%
1	Disgust	547	1.52%
2	Fear	5,121	14.27%
3	Happy	8,989	25.05%
4	Sad	6,077	16.93%
5	Surprise	4,002	11.15%
6	Neutral	6,198	17.27%
Total		**35,887**	**100%**

2.2 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is one of the most widely used deep learning approaches in image classification tasks, including facial expression recognition. A typical CNN architecture consists of three main components: convolutional layers for extracting local spatial features, pooling layers for reducing feature dimensionality, and fully connected layers for classification [7], [8], [9]. Through hierarchical feature learning, CNNs are able to capture increasingly complex visual representations from facial images.

Recent studies have demonstrated the effectiveness of CNN-based approaches for FER. He [13] proposed a multi-branch attention CNN and reported an accuracy of 69.49% on the FER2013 dataset. Chen et al. [28] introduced a lightweight CNN model (LWER) designed to reduce computational complexity while maintaining competitive performance. Ye et al. [15] developed a dual-branch CNN with an attention mechanism, achieving 68.50% accuracy on FER2013, whereas Liu et al. [29] proposed a multi-scale residual attention network (Ms-RAN) to improve feature representation at different scales. In addition, hybrid approaches combining CNN with other classifiers have also been explored. Duan et al. [30], for example, introduced a CNN-ELM framework for age and gender classification tasks.

As illustrated in Figure 1, the baseline CNN architecture used in this study consists of three convolutional blocks arranged hierarchically with 64, 128, and 256 filters, respectively. Each block includes two convolutional layers with 3×3 kernels followed by batch normalization. Max pooling and dropout layers are subsequently applied to reduce the feature dimensions while helping the model avoid overfitting.

Following the feature extraction stage, the generated feature maps are flattened and fed into a fully connected layer containing 256 units, which is also combined with batch normalization and dropout. The network is finalized with a softmax output layer for classification into seven facial expression categories. Overall, the model contains 3,510,215 trainable parameters.

2.3 Extreme Learning Machine (ELM)

ELM is a learning algorithm designed for Single Hidden Layer Feedforward Neural Networks (SLFNs), originally introduced by Huang et al. [16], [17]. One of the main advantages of ELM lies in its fast training process. Unlike conventional neural network methods that rely on iterative optimization, the input weights and hidden layer biases in ELM are randomly initialized and kept fixed during training. As a result, only the output weights need to be computed, which can be obtained analytically using the Moore-Penrose pseudoinverse in a single step [18]. In ELM, the hidden layer representation is formulated at Equation (1).

$$H = \sigma(XW_{input} + b) \quad (1)$$

Where σ is the activation function, X denotes the input data, W_{input} represents the input weights, and b is the hidden layer bias. The output weights are then estimated using regularized least squares as

$$\beta = (H^T H + I/C)^{-1} H^T Y \quad (2)$$

where H is the hidden layer output matrix, Y is the target label matrix, I denotes the identity matrix, and C is the regularization parameter [24].

Several previous studies have demonstrated the potential of ELM in face-related classification tasks. Alessawi [18], for instance, reported an accuracy of 96.7% for face recognition using the Yale Face Database. Meanwhile, Agarwal et al. [19] proposed a hybrid DCNN-ELM approach for masked face recognition, where the combination of ResNet152 and ELM achieved an accuracy of 60.9%.

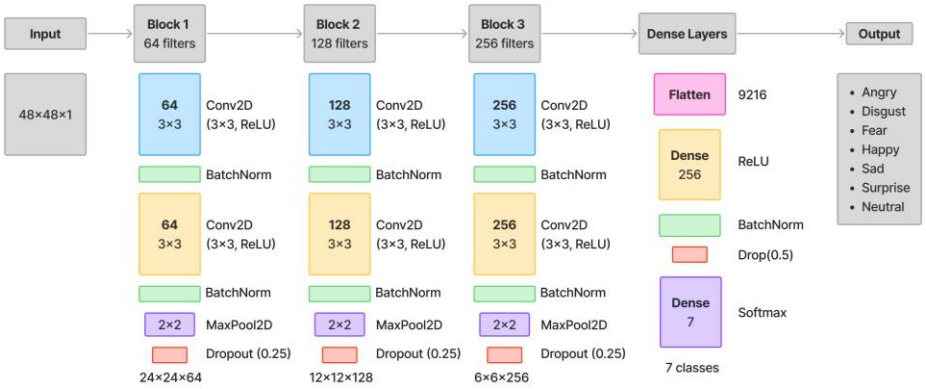

Figure 1. CNN Baseline Architecture

2.4 Ensemble Learning and Diversity Metrics

Ensemble learning combines multiple classifiers to generate predictions that are often more robust and reliable than those produced by a single model [25]. In facial expression recognition (FER), ensemble approaches have attracted considerable attention because they can improve generalization and reduce prediction bias. Yu et al. [27], for instance, proposed MoVE-CNNs (Model Averaging Ensemble of CNNs) for the FER2013 dataset and achieved an accuracy of 77.70%. Lawpanom et al. [2] introduced a homogeneous ensemble CNN (HoE-CNN) with an accuracy of 75.51%, while Hunafa et al. [1] developed a weighted ensemble approach based on the prisoner's dilemma concept for FER tasks.

A major factor affecting ensemble performance is the level of diversity between classifiers. In general, diversity can be explored at three different levels. The first is data diversity, commonly implemented through bagging, where multiple classifiers are trained using different subsets of training data [25], [31]. The second is model diversity, which combines models with different architectures or learning characteristics to capture complementary feature representations. The third is classifier diversity, where identical classifiers are initialized with different random parameters, allowing each model to produce slightly different decision boundaries. Combining these diversity strategies can improve model generalization, particularly when dealing

with complex and imbalanced datasets such as FER2013.

To evaluate the relationship between ensemble members, [31] introduced several diversity metrics that are widely used in ensemble analysis. These metrics are useful for measuring whether classifiers produce complementary predictions rather than identical outputs. The diversity measures adopted in this study are summarized in Table 2.

Table 2 Diversity Metrics [31]

Metric	Formula	Interpretation
Disagreement	$(N_{10} + N_{01})/N\_total$	The higher the disagreement, the better the ensemble
Q-statistic	$(N_{11}N_{00} - N_{10}N_{01})/(N_{11}N_{00} + N_{10}N_{01})$	The Q value ranges from -1 to 1. Q=0 means the classifier is independent
Double-fault	N_{00}/N_total	Lower values indicate better diversity

The diversity metrics are computed from a pairwise contingency table, where N_{11} denotes the number of instances correctly classified by both classifiers, N_{00} denotes the number of instances misclassified by both classifiers, N_{10} denotes instances correctly classified only by classifier i, and N_{01} denotes instances correctly classified only by classifier j. N_{total} represents the total number of evaluated instances.

Previous studies have demonstrated the effectiveness of ensemble learning in FER tasks. Yu et al. [27] reported 77.70% accuracy on FER2013 using MoVE-CNNs, whereas [2] achieved 75.51% using HoE-CNN. Hunafa et al. [1] further improved performance through a weighted ensemble strategy, obtaining 83.25% accuracy on RAF-DB and 64.73% on FER2013. Beyond FER applications, Gupta et al. [26] showed that voting and bagging classifiers can improve face recognition performance, while Pang et al. [32] proposed Instance Euclidean Distance (IED) as a diversity metric for imbalanced ensemble learning scenarios.

3 Methodology

This study utilizes the FER2013 dataset [4], which has been described in detail in Section 2.1. Briefly, the dataset contains 35,887 grayscale facial images with a resolution of 48×48 pixels representing seven facial expression categories, namely Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The dataset exhibits an imbalanced class distribution, where the Happy class represents the largest portion of the samples (25.05%), while Disgust contains only 1.52% of the total data. This imbalance condition makes FER2013 a challenging benchmark for facial expression recognition tasks**.**

3.1 Overview of the Proposed Method

This study proposes a Multi-Level Ensemble approach to improve the performance of Facial Expression Recognition by integrating three forms of diversity, namely data diversity, classifier diversity, and model diversity [25]. The overall framework of the proposed method is illustrated in Figure 2.

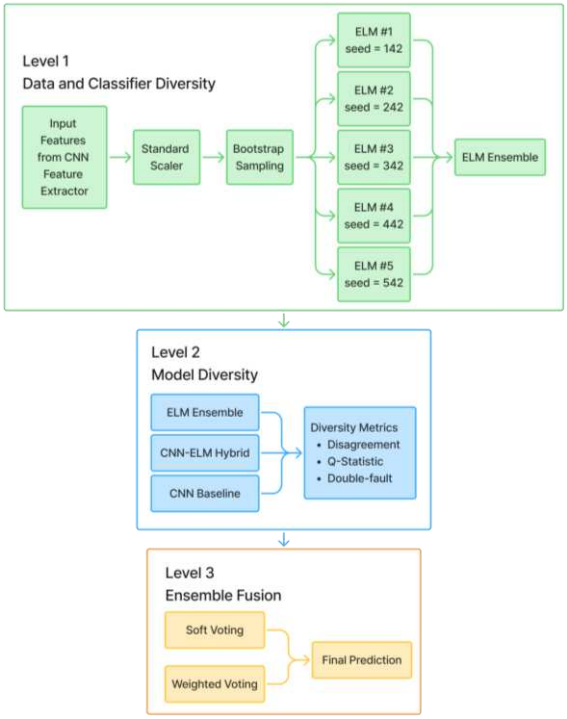

Figure 2. Proposed Multi-Level Ensemble Framework

The framework consists of three parallel branches: Branch 1 (CNN Baseline) performs end-to-end classification directly from input images using a softmax classifier. Branch 2 (CNN-ELM Hybrid) first extracts 256-dimensional features through the CNN feature extractor, then these features are fed into an ELM classifier (1500 hidden neurons, sigmoid activation). Branch 3 (ELM Ensemble) uses the same CNN feature extractor as Branch 2, but applies bootstrap sampling to create five different training subsets, each used to train an independent ELM classifier with different random seeds (142,242,342,442,542). The predictions from all three branches are combined using soft voting to produce the final output.

At the first level, data diversity is introduced through a bagging mechanism applied to feature representations extracted from the CNN model. The CNN architecture first learns facial representations from FER2013 images and produces 256-dimensional feature vectors obtained from the fully connected feature layer. Prior to the learning process, these features are normalized using StandardScaler to obtain zero-mean and unit-variance distributions.

To generate diverse training subsets, bootstrap sampling with replacement is applied to the normalized feature vectors. Each bootstrap subset is then used to train an independent ELM classifier. This strategy allows different ELM models to observe slightly different data distributions during training, thereby increasing data-level variation within the ensemble framework.

In addition to data diversity, classifier diversity is introduced by employing multiple ELM classifiers with different random initializations. In this study, five ELM branches are constructed using different random seeds while maintaining the same

architecture configuration. Each ELM utilizes 1,500 hidden neurons with sigmoid activation and computes the output weights analytically using a regularized pseudoinverse formulation. Since ELM randomly initializes hidden layer parameters, different random seeds produce different decision boundaries even when trained on similar data distributions. The probability outputs generated by all ELM branches are subsequently combined using an average ensemble strategy to produce the final ELM ensemble prediction.

At the second level, model diversity is achieved by combining three learning models with different characteristics, namely the CNN baseline model, the CNN-ELM hybrid model, and the ELM ensemble model. The CNN baseline performs end-to-end feature learning and classification directly from image inputs, while the CNN-ELM model utilizes CNN as a feature extractor followed by ELM as the classifier. Meanwhile, the ELM ensemble model introduces additional variability through bagging and multiple randomized ELM classifiers. Since each model learns feature representations and decision patterns differently, combining them is expected to improve generalization capability and reduce prediction bias.

At the final level, an ensemble fusion strategy is applied to combine prediction probabilities generated by all models from the previous stage. Two fusion mechanisms are evaluated in this study, namely soft voting and weighted voting. In the soft voting scheme, the final probability is obtained by averaging the prediction probabilities from all models. Meanwhile, weighted voting assigns larger contribution weights to models with higher validation accuracy. The final facial expression label is determined based on the highest probability value using the argmax operation. Overall, the proposed multi-level ensemble framework is designed to simultaneously exploit data variation, classifier variability, and heterogeneous model learning strategies in order to obtain more stable and accurate facial expression recognition performance on the FER2013 dataset.

3.2 Data Preprocessing and Augmentation

The FER2013 dataset used in this study was provided in CSV format, where each row contains an emotion label and a sequence of pixel intensity values representing a grayscale facial image. Unlike conventional image datasets stored as separate image files, FER2013 stores facial images as flattened pixel arrays consisting of 48×48=2304 pixel values. This representation facilitates direct numerical processing and simplifies the conversion of image data into tensor format for deep learning experiments. During preprocessing, these pixel arrays are reshaped into two-dimensional image representations before being used for model training [2].

In the preprocessing stage, the pixel sequences from the CSV file were first converted into numerical arrays and reshaped into 48×48×1 image representations to match the input dimension required by the CNN architecture. Since FER2013 consists of grayscale facial images, a single-channel configuration was maintained throughout the experiments. Prior to model training, pixel intensity values were normalized from the original range of 0-255 into a floating-point range of 0-1. This normalization process helps stabilize gradient updates and accelerates convergence during optimization. In addition, categorical emotion labels corresponding to the seven facial expression classes were transformed into one-hot encoded vectors before the classification stage.

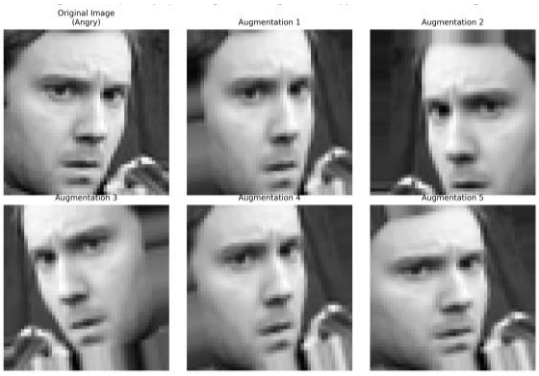

Figure 3. Augmentation applied to FER2013

Figure 3 presents several examples of the preprocessing and augmentation procedures applied in this study. The visualization includes the reconstructed grayscale image obtained from the CSV pixel sequence alongside several augmented samples generated during training. These augmented samples were produced using random rotation, translation, horizontal flipping, and zoom transformations to simulate appearance variations commonly encountered in real-world facial expression scenarios.

To improve model generalization and reduce overfitting, data augmentation was applied exclusively to the training data. Several augmentation techniques were employed, including random rotation within a range of $\pm 15^{\circ}$, horizontal and vertical shifts of up to 10% of the image size, horizontal flipping, and random zoom transformations within a range of 10%. Empty regions generated during augmentation were handled using the nearest-neighbor filling strategy (fill mode = nearest). The augmentation process was performed dynamically using an image generator during training, allowing the CNN model to receive slightly different image variations at each iteration. This strategy increases data diversity and helps the proposed framework learn more robust facial representations under varying pose and appearance conditions.

3.3 CNN Architecture

Table 3 CNN Architecture

Layer	Configuration	Output Shape
Input	48×48×1 grayscale image	(48,48,1)
Block 1	Conv2D(64, 3×3) → BN → Conv2D(64, 3×3) → BN → MaxPooling → Dropout(0.25)	(24,24,64)
Block 2	Conv2D(128, 3×3) → BN → Conv2D(128, 3×3) → BN → MaxPooling → Dropout(0.25)	(12,12,128)
Block 3	Conv2D(256, 3×3) → BN → Conv2D(256, 3×3) → BN → MaxPooling → Dropout(0.25)	(6,6,256)
Feature Layer	Flatten → Dense(256, ReLU) → BN → Dropout(0.5)	256
Output Layer	Dense(7, Softmax)	7
	**CNN baseline only*	

The CNN architecture used in this study was adapted from the hybrid CNN-ELM framework proposed by Duan et al. [30] for age and gender classification, with modifications to suit the FER2013 dataset's characteristics (48×48 input resolution and seven-class classification). The design follows the conventional CNN structure consisting of three convolutional blocks with 64, 128, and 256 filters, which has been widely adopted in FER tasks [8], [11], [14], [15]. The overall network contains approximately 3.51 million parameters, consisting of 3,507,911 trainable parameters with a total memory size of approximately 13.39 MB. The detailed CNN configuration used in this study is presented in Table 3.

For the CNN baseline model, the final output layer consisted of seven neurons with softmax activation corresponding to the seven facial expression categories in the FER2013 dataset. In contrast, for the CNN-ELM and ELM ensemble frameworks, the 256-dimensional feature vectors extracted from the feature layer were subsequently used as inputs for the ELM classifier. Model training was performed using the Adam optimizer with a learning rate of 0.0003. The categorical cross-entropy loss function with label smoothing of 0.1 was employed to improve generalization and reduce overconfidence during prediction. To address class imbalance in FER2013, balanced class weights computed from the training set were applied during optimization. To prevent overfitting, early stopping was implemented by monitoring validation loss with a patience value of 15 epochs. In addition, adaptive learning rate reduction was employed using the ReduceLROnPlateau strategy with a reduction factor of 0.5, patience of 5 epochs, and a minimum learning rate threshold of 1×10^{-6}.

3.4 ELM Classifier

ELM classifier used in this study was implemented using a fixed parameter configuration obtained from the best-performing pure ELM baseline experiment. The selected configuration was adopted based on previous studies demonstrating the effectiveness of ELM for facial expression recognition and image classification tasks by [17], [20], [24], [33]. The parameter configuration used throughout the experiments is summarized as follows 1500 number of hidden neurons (n_{hidden}), sigmoid activation function, and 1.0 regularization coefficient (C).

In this study, the ELM classifier receives 256-dimensional feature vectors extracted from the CNN feature layer as input representations. The combination of CNN-based feature extraction and ELM-based classification was designed to exploit the strong representation learning capability of CNN while maintaining the computational efficiency of ELM. Furthermore, the use of multiple ELM classifiers with different random initializations contributes to classifier diversity within the proposed ensemble framework.

3.5 Multi-Level Ensemble Framework

The proposed framework adopts a multi-level ensemble strategy to improve facial expression recognition performance by integrating several sources of diversity within a unified learning framework. Inspired by ensemble learning theory introduced by Leo Breiman and Zhi-Hua Zhou [23], [25], the proposed method combines data diversity, classifier diversity, and model diversity to encourage complementary learning behavior among classifiers. By combining these diversity mechanisms within a unified architecture, the proposed method aims to improve classification robustness and reduce

prediction bias on the FER2013 dataset. The overall structure of the framework consists of three sequential ensemble levels.

A. Level 1: Data Diversity through Bootstrap Sampling

The first level focuses on data diversity using a bagging mechanism applied to deep feature representations extracted from the CNN feature extractor. After the CNN processes the FER2013 facial images, 256-dimensional feature vectors are obtained from the fully connected feature layer. These features are subsequently normalized using StandardScaler to produce zero-mean and unit-variance distributions before being used in the ELM learning stage.

To increase variability within the training process, bootstrap sampling with replacement is applied to the normalized feature vectors. Each bootstrap subset maintains the same size as the original training set but contains different sample compositions due to random resampling. As a result, each ELM branch is exposed to slightly different data distributions during training. Five bootstrap subsets are generated in this study, and each subset is used to train an independent ELM classifier. This bagging strategy encourages the ensemble to learn more diverse decision patterns and helps reduce model variance during prediction.

B. Level 2: Classifier Diversity through Randomized ELM Branches

The second level introduces classifier diversity through multiple randomized ELM classifiers. Although all ELM branches employ the same architecture configuration, each classifier is initialized using a different random seed, namely 142, 242, 342, 442, and 542. Each ELM uses 1,500 hidden neurons with sigmoid activation and computes the output weights analytically using a regularized pseudoinverse formulation. Since the hidden layer weights and biases in ELM are initialized randomly, different random seeds generate different hidden feature mappings and decision boundaries even when trained on similar data distributions. The activation process maps the input features into a hidden representation that can be expressed using Equation (1). Furthermore, the output weights are determined analytically through a regularized least-squares solution, as presented in Equation (2). This analytical computation enables ELM to achieve fast training performance while maintaining good generalization capability. The probability outputs generated by all ELM branches are then combined using probability averaging:

$$P_{ensemble} = \frac{P_1 + P_2 + P_3 + P_4 + P_5}{5} \tag{3}$$

This combination strategy helps stabilize prediction results while preserving classifier variability among ELM branches.

C. Level 3: Model Diversity through Heterogeneous Learning Models

At the third level, model diversity is introduced by combining three heterogeneous learning models that employ different representation and classification mechanisms. The first model, M1 or the CNN Baseline, uses an end-to-end CNN architecture with a Softmax classifier, enabling direct image-based classification. The second model, M2 or CNN-ELM, integrates a CNN feature extractor with an ELM classifier, thereby combining deep feature representation with an analytical

classification approach. Meanwhile, the third model, M3 or the ELM Ensemble, consists of an ensemble of five bagged ELM classifiers that leverage bagging and randomized ensemble learning to improve classification robustness and diversity.

The CNN baseline model performs both feature extraction and classification directly from image inputs using a Softmax output layer. In contrast, the CNN-ELM model separates the feature extraction and classification stages by utilizing CNN for deep representation learning and ELM for classification. Meanwhile, the ELM ensemble model emphasizes ensemble learning through bootstrap sampling and multiple randomized ELM branches. Because each model learns facial representations using different optimization characteristics and decision mechanisms, combining them is expected to produce complementary prediction behavior and improve generalization performance.

D. Ensemble Fusion Strategy

After obtaining prediction probabilities from all models, an ensemble fusion mechanism is applied to generate the final facial expression prediction. Two fusion strategies are evaluated in this study, namely soft voting and weighted voting. In the soft voting approach, the prediction probabilities from all models are averaged equally:

$$P_{final} = \frac{P_{cnn} + P_{cnn\_elm} + P_{elm\_ensemble}}{3} \tag{4}$$

The final predicted class is determined using the argmax operation:

$$\hat{y} = \text{argmax}(P_{final}) \tag{5}$$

In addition, a weighted voting mechanism is also explored by assigning larger weights to models with higher validation accuracy. The total accuracy score is first computed as:

$$accuracy_{total} = acc_{cnn} + acc_{cnn\_elm} + acc_{elm\_ensemble} \tag{6}$$

The contribution weight for each model is then calculated proportionally:

$$w_{cnn} = \frac{acc_{cnn}}{accuracy_{total}} \tag{7}$$

The contribution weight for each model is then calculated proportionally:

$$P_{final} = w_{cnn}P_{cnn} + w_{cnn\_elm}P_{cnn\_elm} + w_{elm\_ensemble}P_{elm\_ensemble} \tag{8}$$

Overall, the proposed multi-level ensemble framework is designed to simultaneously exploit variations in training data, randomized classifier behavior, and heterogeneous model learning strategies in order to achieve more stable and accurate facial expression recognition performance on the FER2013 dataset.

4 Result and Discussion

4.1 Model Performance Summary

Data augmentation was applied uniformly to all models throughout the experiments. Hence, the accuracy gain achieved by the proposed ensemble (+1.33%) reflects the contribution of the ensemble mechanism itself, not the augmentation process. The proposed framework was evaluated using Stratified 5-Fold Cross Validation on the FER2013 dataset, which consists of 35,887 facial images distributed across seven emotion classes. The use of stratified validation ensures that the class distribution remains consistent across all folds, allowing a fair comparison between models under imbalanced data conditions. Table 4 summarizes the performance obtained by all evaluated models in terms of accuracy, F1-score, and total training time for the five-fold evaluation process.

Table 4. Performance Comparison of All Models Using 5-Fold Cross Validation

Model	Accuracy	F1-Score	Training Time (5-Fold)
Pure ELM	36.37%	33.37%	3.0 minutes
CNN Baseline	66.90%	66.72%	120.3 minutes
CNN-ELM	67.28%	67.10%	135.4 minutes
ELM Ensemble	67.34%	67.10%	113.5 minutes
*Soft Voting Ensemble (Proposed)	68.23%	68.03%	113.5 minutes

The experimental results indicate that the Pure ELM model produced the lowest performance among all evaluated methods. Although ELM offers very fast training time due to its analytical learning mechanism, the model struggled to learn discriminative facial representations directly from raw pixel inputs. The obtained accuracy was only 36.37%, with extremely poor recognition performance for minority and visually ambiguous classes such as Disgust and Fear. In several evaluation folds, the recall value for the Disgust class even approached 0%, while Fear recognition remained very limited. These findings suggest that shallow classifiers operating directly on raw grayscale pixels are insufficient for handling the complex intra-class variations present in FER2013.

The CNN baseline model significantly improved recognition performance by learning hierarchical spatial representations from facial images. Using convolutional feature extraction combined with Softmax classification, the CNN achieved an accuracy of 66.90% and an F1-score of 66.72%. This result confirms the effectiveness of deep convolutional representations for facial expression recognition tasks. Further improvement was observed when the CNN feature extractor was combined with the ELM classifier in the CNN-ELM architecture. The hybrid approach achieved 67.28% accuracy, outperforming the CNN baseline by 0.38%. Although the numerical improvement appears relatively small, several emotion categories experienced noticeable recall gains, particularly Fear and Happy. The analytical learning mechanism of ELM appears to complement the discriminative deep features generated by CNN, allowing the hybrid model to produce slightly more generalized decision boundaries.

The ELM Ensemble model also demonstrated competitive performance with an accuracy of 67.34%. By integrating bootstrap sampling and multiple randomized ELM

branches, the ensemble was able to reduce prediction variance and improve classification stability compared to a single ELM classifier. Interestingly, the ELM Ensemble slightly outperformed the CNN-ELM model while requiring lower overall training time than the CNN baseline.

Among all evaluated methods, the proposed Soft Voting ensemble achieved the best overall performance with an accuracy of 68.23% and an F1-score of 68.03%. Compared to the CNN baseline, the proposed framework improved accuracy by 1.33%. This result indicates that combining heterogeneous models with complementary learning characteristics can produce more robust predictions than relying on a single classifier architecture. The experimental findings demonstrate that the proposed multi-level ensemble framework successfully benefits from the combination of deep feature learning, randomized classifier behavior, and ensemble fusion strategies. The integration of data diversity, classifier diversity, and model diversity contributes to more stable facial expression recognition performance under the challenging conditions of the FER2013 dataset.

4.2 Diversity Metrics Analysis

To further analyze the effectiveness of the proposed ensemble framework, several diversity metrics were calculated between the CNN baseline and CNN-ELM models. Diversity analysis is important in ensemble learning because ensemble performance generally improves when participating models produce different prediction behaviors and error patterns.Three commonly used diversity metrics were evaluated in this study, namely disagreement measure, Q-statistic, and double-fault measure. The results are summarized in Table 5.

Table 5. Diversity Metrics Between CNN Baseline and CNN-ELM

Metric	Value	Interpretation
Disagreement	21.2%	Ideal diversity range (20–30%)
Q-statistic	0.9313	High positive correlation, but not identical
Double-fault	25.5%	Both models simultaneously fail on one-quarter of samples

The disagreement value of 21.2% indicates that the CNN baseline and CNN-ELM models produced different predictions on approximately one-fifth of the evaluated samples. According to ensemble learning theory, disagreement values within the range of 20–30% are generally considered beneficial because they indicate sufficient prediction diversity without causing excessive instability. This suggests that the two models learn complementary decision patterns from the same facial expression data.

Meanwhile, the obtained Q-statistic value of 0.9313 indicates a strong positive correlation between both models. Although the predictions remain highly correlated, the value still shows that the models are not completely identical. This condition is desirable in ensemble systems because extremely low correlation may indicate unstable behavior, while perfectly identical predictions would provide little ensemble benefit.

The double-fault value of 25.5% shows that both models simultaneously misclassified approximately one out of four samples. This result implies that there remains a substantial portion of difficult facial expression samples within FER2013, particularly for visually overlapping emotion categories. However, because the remaining prediction errors are not fully shared between models, the ensemble mechanism still has opportunities to correct incorrect predictions produced by

individual classifiers.

Overall, the diversity analysis supports the effectiveness of the proposed multi-level ensemble framework. The combination of moderately diverse prediction behavior and strong individual classifier performance enables the ensemble model to achieve higher overall recognition accuracy compared to single-model approaches.

4.3 Per-Class Performance Analysis

To obtain a more detailed understanding of the proposed framework, class-wise recall analysis was conducted for each facial expression category in FER2013. Recall was selected because the dataset contains class imbalance, making overall accuracy alone insufficient to fully describe model behavior across minority and majority classes. Table 6 presents the recall comparison between the CNN baseline and the proposed Soft Voting ensemble.

Table 6. Recall Comparison for Each Facial Expression Class

Emotion Class	CNN Baseline	Soft Voting	Improvement
Angry	62.57%	62.63%	+0.06%
Disgust	71.85%	69.84%	−2.01%
Fear	43.68%	45.44%	+1.76%
Happy	83.55%	85.45%	+1.90%
Sad	51.46%	53.49%	+2.48%
Surprise	80.46%	81.38%	+0.92%
Neutral	71.35%	71.93%	+0.58%

The results show that six out of seven emotion classes experienced recall improvement after applying the proposed ensemble framework. The most noticeable gains were observed for the Sad and Fear categories, which improved by 2.48% and 1.76%, respectively. These two classes are generally considered more difficult to recognize because their facial characteristics often overlap with other expressions such as Neutral and Angry. The improvement suggests that the ensemble framework was able to capture complementary prediction patterns that were not fully learned by the CNN baseline alone.

The Happy and Surprise classes also achieved relatively high recall values exceeding 80%. These expressions typically contain more distinctive facial patterns, such as smiling and widened eyes, making them easier for deep learning models to distinguish consistently.

Meanwhile, the Disgust class was the only category that experienced a performance decrease, dropping by 2.01% compared to the CNN baseline. This decline may be associated with the severe class imbalance present in FER2013, where Disgust contains substantially fewer training samples than other emotion categories. As a result, the ensemble fusion process may shift decision boundaries toward more dominant classes during probability averaging.

Although the decrease in Disgust recall should be acknowledged as a limitation, the overall results still indicate that the proposed framework improves recognition consistency across most facial expression categories, particularly for difficult and ambiguous emotions.

4.4 Statistical Significance Analysis

To verify whether the observed performance improvement was statistically meaningful, a paired t-test was conducted between the CNN baseline and the proposed

Soft Voting ensemble model. The test evaluates whether the performance differences obtained across the five validation folds occurred consistently rather than by random variation. The statistical test results are summarized in Table 7.

Table 7. Paired t-Test Results

Metric	Value
Mean Difference	+1.33%
t-statistic	-16.86%
p-value	0.000073

The obtained p-value is substantially lower than the significance threshold of 0.05, indicating that the null hypothesis can be rejected. Therefore, the performance improvement produced by the proposed ensemble framework is statistically significant.

Although the numerical improvement in accuracy appears moderate, the statistical analysis demonstrates that the gain was consistently observed across all validation folds. This finding strengthens the reliability of the proposed method and suggests that the observed improvement is not caused by random sampling effects.

The result also supports the effectiveness of combining multiple diversity mechanisms within a single ensemble framework. The integration of data diversity, classifier diversity, and heterogeneous model learning contributes to more stable prediction behavior and improved generalization performance on FER2013.

4.5 Comparison with Previous Studies

To evaluate the effectiveness of the proposed framework, we compared our results with several recent FER studies on the FER2013 dataset, as summarized in Table 8.

Table 8. Comparison of Accuracy using FER2013 Studies

Study	Method	Accuracy
He, 2023[13]	Multi-branch Attention CNN	69.49%
Nguyen et al., 2022 [34]	Multi-level CNN	73.03%
Nguyen et al., 2022 [34]	Ensemble of 3 MLCNNs	74.09%
Renda et al., 2019 [35]	Ensemble of 9 CNNs (Pretraining)	72.25%
Lawpanom & Songpan, 2024 [2]	HoE-CNN (Homogeneous Ensemble)	75.51%
Hunafa et al., 2025 [1]	Weighted Ensemble	64.73%
*Proposed Method, 2026	Multi-Level Ensemble (Soft Voting)	68.23%

As shown in Table 8, the highest accuracy (75.51%) is achieved by Lawpanom and Songpan [2] using a homogeneous ensemble of seven CNN architectures. Nguyen et al. [34] achieved 74.09% with an ensemble of three multi-level CNN variants, while Renda et al. [35] obtained 72.25% using a pretraining strategy with nine CNNs. He [13] achieved 69.49% with a multi-branch attention CNN, and Hunafa et al. [1] achieved 64.73% with a weighted ensemble. Our method achieves 68.23%, which is higher than Hunafa et al. [1] and comparable to He [13], but lower than the top-performing ensemble-based methods.

This performance gap is primarily attributed to the use of a simpler CNN containing only 3.51 million parameters, without transfer learning, face alignment, or additional pretraining datasets. Nevertheless, the simplicity of the proposed architecture offers potential for further performance improvement through the integration of more advanced feature extractors in future work.

From a computational perspective, the proposed method remains lightweight. The total training time for a complete 5-fold cross-validation procedure is only 113.5 minutes on a single Tesla T4 GPU, indicating that the model can be trained efficiently using modest computational resources. Although the training times of the competing methods are not reported, their architectures are substantially larger. For example, Nguyen et al. [34] utilizes an architecture with 20.8 million parameters, while Lawpanom and Songpan [2] employ an ensemble of seven different architectures, including VGG16, which alone contains approximately 138 million parameters. Even Renda et al. [35] achieves higher accuracy by combining nine different networks. These observations suggest that the proposed method offers a favorable trade-off between recognition accuracy and model complexity, demonstrating that competitive performance can be achieved using a considerably lighter architecture suitable for resource-constrained environments.

Nevertheless, our method offers three distinct contributions not present in these previous studies: First, we explicitly analyze ensemble diversity using disagreement (21.2%), Q-statistic (0.9313), and double-fault (25.5%) metrics. None of the compared studies [1], [2], [13], [34], [35] provided such quantitative diversity analysis to explain why their ensembles work. Second, we integrate three diversity mechanisms simultaneously (data diversity via bootstrap sampling, classifier diversity via random ELM seeds, and model diversity via heterogeneous architectures). Renda et al. [35]and Lawpanom and Songpan [2]used homogeneous CNN ensembles, whereas our method combines CNN, CNN-ELM, and ELM architectures. Third, we provide statistical validation using paired t-test ($p=0.000073$), confirming that our improvement is statistically significant. This analysis is absent in all compared studies. In summary, although our accuracy is lower than state-of-the-art ensemble methods, our contributions in diversity analysis, heterogeneous ensemble design, and statistical validation provide valuable insights into understanding why ensemble learning works for FER tasks.

5 Conclusion

This study proposed a Multi-Level Ensemble framework for Facial Expression Recognition using the FER2013 dataset. The proposed approach integrates data diversity, classifier diversity, and model diversity within a unified ensemble architecture to improve recognition performance under class imbalance conditions.

Experimental results obtained using Stratified 5-Fold Cross Validation showed that the Pure ELM model performed poorly on raw pixel inputs, achieving only 36.37% accuracy. This limitation motivated the use of CNN-based feature extraction combined with ELM classification. The CNN-ELM model slightly outperformed the CNN baseline, improving accuracy from 66.90% to 67.28%.

The best performance was achieved by the proposed Soft Voting ensemble, which obtained 68.23% accuracy and 68.03% F1-score, representing a 1.33% improvement over the CNN baseline. Paired t-test analysis confirmed that the improvement was statistically significant with a p-value of 0.000073.

The effectiveness of the proposed framework is supported by the diversity analysis results. A disagreement value of 21.2% indicates that the participating models produced complementary prediction patterns, while the bagging strategy and randomized ELM initialization helped improve prediction stability and generalization performance. Class-wise evaluation further showed recall improvements in six out of seven emotion categories, particularly for Sad and Fear expressions.

Overall, the results demonstrate that combining multiple diversity strategies

within a single ensemble framework can improve FER performance on imbalanced datasets. Future work may focus on imbalance-aware learning methods and evaluation on additional FER benchmark datasets such as FER+, RAF-DB, and CK+.

References

1. M. H. Hunafa, M. Luthfi Ramadhan Mgs, A. W. Ramadhan, M. F. Rachmadi, and W. Jatmiko, "Weighted Ensemble Based on Prisoner Dilemma for Facial Expression Recognition," *IEEE Access*, vol. 13, pp. 109200–109218, 2025, doi: 10.1109/ACCESS.2025.3582643.
2. R. Lawpanom, W. Songpan, and J. Kaewyotha, "Advancing Facial Expression Recognition in Online Learning Education Using a Homogeneous Ensemble Convolutional Neural Network Approach," *Appl. Sci. 2024, Vol. 14, Page 1156*, vol. 14, no. 3, p. 1156, Jan. 2024, doi: 10.3390/APP14031156.
3. T. D. Pham, M. T. Duong, Q. T. Ho, S. Lee, and M. C. Hong, "CNN-Based Facial Expression Recognition with Simultaneous Consideration of Inter-Class and Intra-Class Variations," *Sensors 2023, Vol. 23, Page 9658*, vol. 23, no. 24, p. 9658, Dec. 2023, doi: 10.3390/S23249658.
4. I. J. Goodfellow *et al.*, "Challenges in Representation Learning: A Report on Three Machine Learning Contests," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8228 LNCS, no. PART 3, pp. 117–124, 2013, doi: 10.1007/978-3-642-42051-1_16.
5. B. Nemade, V. Bharadi, S. S. Alegavi, and B. Marakarkandy, "A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons," 2023. [Online]. Available: www.ijisae.org
6. B. Luís Coimbra Maia *et al.*, "Class-Based Adaptive Training for Facial Expression Recognition Using a Deep Convolutional Network," *21st Int. Conf. Comput. Vis. Theory Appl.*, 2026, doi: 10.5220/0014338500004084.
7. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
8. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, Accessed: May 05, 2026. [Online]. Available: http://code.google.com/p/cuda-convnet/
9. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/NATURE14539;SUBJMETA.
10. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, Sep. 2014, Accessed: May 05, 2026. [Online]. Available: https://arxiv.org/pdf/1409.1556
11. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.1109/CVPR.2016.90.
12. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2818–2826, Dec. 2016, doi: 10.1109/CVPR.2016.308.
13. Y. He, "Facial Expression Recognition Using Multi-Branch Attention Convolutional Neural Network," *IEEE Access*, vol. 11, pp. 1244–1253, 2023, doi: 10.1109/ACCESS.2022.3233362.
14. Q. Chen, X. Jing, F. Zhang, and J. Mu, "Facial Expression Recognition Based on A Lightweight CNN Model," *IEEE Int. Symp. Broadband Multimed. Syst. Broadcast. BMSB*, vol. 2022-June, 2022, doi: 10.1109/BMSB55706.2022.9828739.
15. F. Ye, J. Ju, M. Lin, and J. Tang, "Research on Dual-Branch CNN Model for Facial Expression Recognition with Attention Mechanism," *2025 7th Int. Conf. Electron. Commun. Netw. Comput. Technol. ECNCT 2025*, pp. 268–274, 2025, doi: 10.1109/ECNCT66493.2025.11172422.
16. G.-B. Huang Ã, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, pp. 489–501, 2006, doi:

10.1016/j.neucom.2005.12.126.
17. G. Bin Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012, doi: 10.1109/TSMCB.2011.2168604.
18. K. S. Alessawi, "A Robust Face Recognition Framework Based on CapsNet–PCA–ELM Integration," *J. Technol. Res.*, vol. 3, no. 2, pp. 155–163, Dec. 2025, doi: 10.26629/JTR.2025.18.
19. C. Agarwal, A. Mishra, and C. Bhatnagar, "Deep Transfer Learning for Masked Face Reconstruction and Hybrid DCNN-ELM Framework for Recognition," *Int. J. Commun. Networks Inf. Secur.*, pp. 170–192, Sep. 2024, Accessed: May 05, 2026. [Online]. Available: https://mail.ijcnis.org/index.php/ijcnis/article/view/6882
20. D. Ghimire and J. Lee, "Extreme Learning Machine Ensemble Using Bagging for Facial Expression Recognition," *J. Inf. Process. Syst.*, vol. 10, p. 443~458, May 2014, doi: 10.3745/JIPS.02.0004.
21. R. A. Cahya, F. A. Bachtiar, and W. F. Mahmudy, "Comparison of Bagging Ensemble Combination Rules for Imbalanced Text Sentiment Analysis," *J. Inf. Technol. Comput. Sci.*, vol. 6, no. 1, pp. 33–49, Apr. 2021, doi: 10.25126/JITECS.202161206.
22. L. Muflikhah, N. Widodo, W. F. Mahmudy, and Solimun, "Prediction of Liver Cancer Based on DNA Sequence Using Ensemble Method," *2020 3rd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2020*, pp. 37–41, Dec. 2020, doi: 10.1109/ISRITI51436.2020.9315341.
23. L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/BF00058655/METRICS.
24. G. G. Wang, M. Lu, Y. Q. Dong, and X. J. Zhao, "Self-adaptive extreme learning machine," *Neural Comput. Appl. 2015 272*, vol. 27, no. 2, pp. 291–303, Mar. 2015, doi: 10.1007/S00521-015-1874-3.
25. Z. H. Zhou, *Ensemble methods: Foundations and algorithms*, 1st ed. New York: CRC Press, 2012. doi: https://doi.org/10.1201/b12207.
26. A. Gupta, S. Sriram, and V. Nivethitha, "Harnessing Diversity in Face Recognition: A Voting and Bagging Ensemble Approach," *2024 Int. Conf. Autom. Comput. AUTOCOM 2024*, pp. 249–255, 2024, doi: 10.1109/AUTOCOM60220.2024.10486188.
27. J. X. Yu, K. M. Lim, and C. P. Lee, "MoVE-CNNs: Model aVeraging Ensemble of Convolutional Neural Networks for Facial Expression Recognition," *IAENG Int. J. Comput. Sci.*, vol. 48, no. 3, pp. 1–5, 2021, Accessed: May 05, 2026. [Online]. Available: https://research.nottingham.edu.cn/en/publications/move-cnns-model-averaging-ensemble-of-convolutional-neural-networ/
28. X. Chen, S. Yang, L. Shen, and X. Pang, "A Distributed Training Algorithm of Generative Adversarial Networks with Quantized Gradients," 2020, [Online]. Available: http://arxiv.org/abs/2010.13359
29. D. Liu, L. Wang, Z. Wang, and L. Chen, "Novel multi-scale deep residual attention network for facial expression recognition," *J. Eng.*, vol. 2020, no. 12, pp. 1220–1226, Dec. 2020, doi: 10.1049/JOE.2020.0183.
30. M. Duan, K. Li, C. Yang, and K. Li, "A hybrid deep learning CNN–ELM for age and gender classification," *Neurocomputing*, vol. 275, pp. 448–461, Jan. 2018, doi: 10.1016/J.NEUCOM.2017.08.062.
31. L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, May 2003, doi: 10.1023/A:1022859003006/METRICS.
32. Y. Pang, L. Peng, H. Zhang, Z. Chen, and B. Yang, "Imbalanced ensemble learning leveraging a novel data-level diversity metric," *Pattern Recognit.*, vol. 157, p. 110886, Jan. 2025, doi: 10.1016/J.PATCOG.2024.110886.
33. Y. Tian, J. Zhang, L. Chen, Y. Geng, and X. Wang, "Selective Ensemble Based on Extreme Learning Machine for Sensor-Based Human Activity Recognition," *Sensors 2019, Vol. 19, Page 3468*, vol. 19, no. 16, p. 3468, Aug. 2019, doi: 10.3390/S19163468.
34. H. D. Nguyen, S. H. Kim, G. S. Lee, H. J. Yang, I. S. Na, and S. H. Kim, "Facial Expression Recognition Using a Temporal Ensemble of Multi-Level Convolutional Neural Networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 226–237, 2022, doi:

10.1109/TAFFC.2019.2946540.
35. A. Renda, M. Barsacchi, A. Bechini, and F. Marcelloni, “Comparing ensemble strategies for deep learning: An application to facial expression recognition,” *Expert Syst. Appl.*, vol. 136, pp. 1–11, Dec. 2019, doi: 10.1016/J.ESWA.2019.06.025.