Clustering-Enhanced Facial Emotion Recognation Using ELM: A Hybrid KMeans and PSO Approach

Abstract

Recognizing emotional expressions from human faces plays a crucial role in building intelligent systems for applications such as affective computing, human-computer interaction, and psychological assessment. In this study, a facial emotion recognition (FER) system is developed and tested on a dataset containing 190,967 samples. These samples are represented by 10 virtual marker features, derived from facial action units, and collected through an image processing pipeline involving face and eye detection and marker tracking during emotional expression. This paper presents a facial emotion recognition (FER) system that uses a two-stage approach combining optimized clustering and hybrid machine learning to improve accuracy and processing efficiency. The system first applies an enhanced K-means algorithm, optimized with Particle Swarm Optimization (PSO), to divide the dataset into five initial clusters based on expression features. After analyzing the clusters, Clusters 0, 1, 2, and 4—containing clear and easily recognizable emotions—were merged into a single group called **Cluster A**. Cluster 3, which included more ambiguous and hard-to-classify expressions, was labeled **Cluster B**.

Cluster A was processed using an Extreme Learning Machine (ELM), which provided fast and highly accurate results, achieving 99% accuracy. Cluster B was classified using a weighted XGBoost model, which performed better than traditional methods, especially on complex samples, and improved accuracy by 22%. The system It achieved an overall weighted accuracy of **94.59%**, while also reducing processing time. The proposed method offers a balance between speed and accuracy, making it suitable for real-time applications such as affective computing, human-computer interaction, and psychological assessment.

Introduction

Facial emotion recognition has emerged as a crucial technology with applications ranging from psychological therapy to human-computer interaction systems. While significant progress has been made in this field, current systems often struggle with inconsistent performance across different emotion categories. As demonstrated in recent work by Murugappan et al. [1], even advanced techniques using virtual markers and neural networks face challenges in accurately recognizing subtle expressions like disgust or fear compared to more distinct emotions like happiness.

This paper presents an innovative approach that addresses these challenges through a carefully designed hybrid system. Our work begins with an extensive analysis of a comprehensive dataset containing 190,966 facial expression samples, each characterized by 10 virtual marker features. Initial tests showed that some emotions are much easier to recognize than others. For example, standard methods could identify happy expressions with 99.13% accuracy, but struggled with more subtle emotions like disgust, achieving only 62.20% accuracy. Seeing this big difference in performance inspired us to create a smarter system that can adjust its approach depending on how easy or difficult each facial expression is to recognize.

Building upon the foundation of virtual marker technology established in Murugappan et al. [1], we introduce a novel three-stage processing pipeline. The first stage employs an enhanced clustering algorithm combining Particle Swarm Optimization with K-means (PSO-KMeans) to automatically partition the dataset. This optimized clustering approach demonstrated higher performance, identifying five natural clusters that were subsequently merged into two primary groups: a high-confidence cluster containing 132,835 samples (69.6% of total data) and a challenging-expression cluster with 58,131 samples (30.4%). The effectiveness of this partitioning is evident in the cluster separation score of 0.78, representing a 26% improvement over standard K-means.

For model development, we implemented specialized architectures tailored to each cluster's characteristics. The high-confidence cluster utilizes an Extreme Learning Machine (ELM) with Gaussian activation function, achieving exceptional **99.13%** accuracy while maintaining rapid 0.8ms processing times. The challenging-expression cluster employs a carefully tuned XGBoost ensemble incorporating 650 decision trees and class-balanced weighting, which improved accuracy to 84.17% - a significant 22% enhancement over using ELM alone for these difficult cases.

The complete system integrates these components through an intelligent routing mechanism that automatically directs each input to the appropriate model based on its cluster assignment.

Comprehensive evaluation using five-fold cross-validation demonstrates the system's robust performance, achieving an overall weighted accuracy of **94.59%** (Eqn 1)**.** Particularly noteworthy are the improvements in traditionally challenging categories, with disgust recognition improving by **22%** and fear by **19%**, while maintaining excellent performance on clearer expressions.

$$\text{Overall Accuracy} = \frac{(26567 \times 0.9913) + (11627 \times 0.8417)}{26567 + 11627}$$
$$= \frac{26340.0471 + 9789.0759}{38194} = \frac{36129.123}{38194} \approx 0.9459$$

Enq.1 Overall Accuracy Calculation

This work makes several important contributions to the field of affective computing. First, it presents a novel clustering methodology specifically optimized for facial expression data. Second, it demonstrates the significant advantages of cluster-specific model specialization over single-model approaches.

Literature Review

Facial emotion recognition has been widely studied in recent years, with researchers exploring different approaches to improve accuracy. Early work by Murugappan et al. [1], introduced virtual markers and used ELM and PNN classifiers, achieving up to 92% accuracy. Their method showed that some emotions like happiness (96% accuracy) were easier to recognize than others such as fear (89% accuracy), highlighting the varying difficulty across emotion categories.

Several studies have focused on improving feature extraction methods. Smith and Rossit [2], demonstrated that certain facial regions are more important for recognizing specific emotions. Their work found that people can identify emotions better when looking at certain parts of the face, suggesting that not all facial features contribute equally to emotion recognition. This insight helped guide our feature selection process.

Recent advances in deep learning have brought new possibilities to the field. Teng and Yang [3], used convolutional neural networks for emotion recognition, while Yu et al. [4], combined these with LSTM networks to capture temporal patterns. Although these methods achieved good results (up to 99.73% accuracy in some cases), they often require significant computational resources, making them less suitable for real-time applications.

Traditional machine learning approaches continue to show promise, especially when combined with smart preprocessing. Alphonse and Dharma [5], used dimension reduction techniques with ELM and SVM classifiers, achieving 97.7% accuracy on some datasets. Their work confirmed that

careful feature selection and model tuning can produce excellent results without requiring deep neural networks.

A key challenge identified in the literature is handling real-world conditions. Nonis et al. [6], reviewed 3D approaches and noted difficulties with lighting changes and facial variations. Similarly, Quelhas [7], compared marker-based and marker less systems, finding trade-offs between accuracy and practicality. These studies informed our decision to use virtual markers while optimizing for real-time performance.

The proposed framework advances existing research by addressing three key limitations in current facial emotion recognition systems. Building on established virtual marker techniques, the framework integrates adaptive model selection to optimize processing efficiency. A novel clustering mechanism explicitly accounts for the varying recognition difficulty across different emotion categories, from easily identifiable expressions to more subtle ones. Furthermore, the system demonstrates robust performance in practical applications, maintaining high classification accuracy while meeting computational efficiency requirements for real-world deployment.

Methodology

This research developed an advanced facial emotion recognition system through a carefully designed five-stage experimental process. The study employed a comprehensive dataset of 190,966 facial expression samples, each characterized by ten precisely measured virtual marker features extracted using computer vision techniques. All experiments maintained rigorous standards through five-fold cross-validation with stratified sampling, ensuring reliable and representative results across all emotion categories. Performance evaluations considered both classification accuracy and computational efficiency, with all timing measurements recorded on consistent hardware specifications to enable fair comparisons. The methodology progressively evolved from fundamental baseline assessments to an optimized hybrid architecture, with each stage addressing specific challenges identified in previous phases while building toward the final system design.

A. Initial Model Evaluation

The investigation began with establishing performance benchmarks using an Extreme Learning Machine architecture. The ELM configuration incorporated an input layer matching the ten virtual marker features, a hidden layer containing 1,500 neurons with Gaussian activation function, and an output layer with six nodes corresponding to the emotion categories. A small regularization parameter (λ=0.00001) prevented model overfitting while maintaining flexibility. As shown in TABLE I, extensive evaluation through repeated cross-validation demonstrated consistent

performance across emotion categories, achieving 86.69% average accuracy. The Gaussian activation function proved particularly effective for recognizing happiness expressions, attaining 92.5% recall, while presenting greater challenges for more subtle emotions like disgust at 87% recall. These baseline results highlighted both the potential of ELM architectures and the need for specialized handling of certain expression types.

Emotion	**Precision**	**Recall**	**F1-Score**	**Support Samples**
Happiness	89%	**92%**	**90%**	32,275
Surprise	**90%**	87%	88%	32,416
Disgust	83%	87%	85%	31,016

TABLE I: Baseline ELM Classification Performance

B. Cluster Analysis Implementation

In the second phase of the experiment, K-means clustering was applied using five clusters and the Euclidean distance metric (Eqn 2). to explore the natural groupings in the dataset.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

(Eqn 2)

The analysis revealed varying classification performance across the clusters. Three clusters, containing a total of 126,049 samples, achieved high classification accuracy above 98%, indicating that these expressions were easy to recognize due to their distinct features. One cluster with 40,744 samples showed moderate accuracy at 57.17%, and another cluster with 19,173 samples had relatively low accuracy at 66.81%, suggesting that these expressions were more difficult to classify. Cluster evaluation metrics, including a silhouette score of 0.62 and a Davies-Bouldin index of 0.91, confirmed a reasonable separation between the clusters. The overall weighted classification accuracy across all clusters was 87.29%, showing that clustering helped identify groups with similar classification behavior, although further improvements may be needed to better separate the more difficult samples.

Cluster	Sample Count	Classification Accuracy
0	40,744	57.17%
1	53,410	**99.87%**
2	18,407	98.70%
3	19,173	66.81%
4	**59,232**	99.76%

TABLE II: Initial Cluster Performance Characteristics

C. Evolutionary Cluster Optimization

The third phase enhanced the clustering through Particle Swarm Optimization, employing a swarm of thirty particles with carefully tuned cognitive (1.5) and social (1.7) parameters balanced by an inertia weight (0.6). The optimization process completed 150 iterations before converging to an optimal solution. This advanced approach significantly improved cluster separation metrics, increasing the silhouette score by twenty-six percent to 0.78 while reducing the Davies-Bouldin index to 0.85. Most notably, the previously problematic cluster showed an 8.2% accuracy improvement to 62.20%. As detailed in Table III, the PSO optimization increased the overall weighted accuracy to 88.01% while creating more homogeneous clusters. One cluster achieved perfect classification for its 15,849 samples, demonstrating that certain expression groups are inherently more separable than others.

Cluster	Sample Count	Classification Accuracy
0	**58,492**	99.84%
1	17,089	97.22%
2	15,849	**100.00%**
3	58,131	62.20%
4	41,405	99.14%

TABLE III: Optimized Cluster Performance Metrics

D. Strategic Cluster Merging

Analysis of the optimized clusters revealed opportunities for strategic simplification while maintaining performance benefits. Four clusters demonstrating similar classification characteristics were merged into a primary group containing 132,835 samples with 99.24% average accuracy. The remaining cluster with 58,131 samples was maintained as a separate group for specialized processing (fig 1).

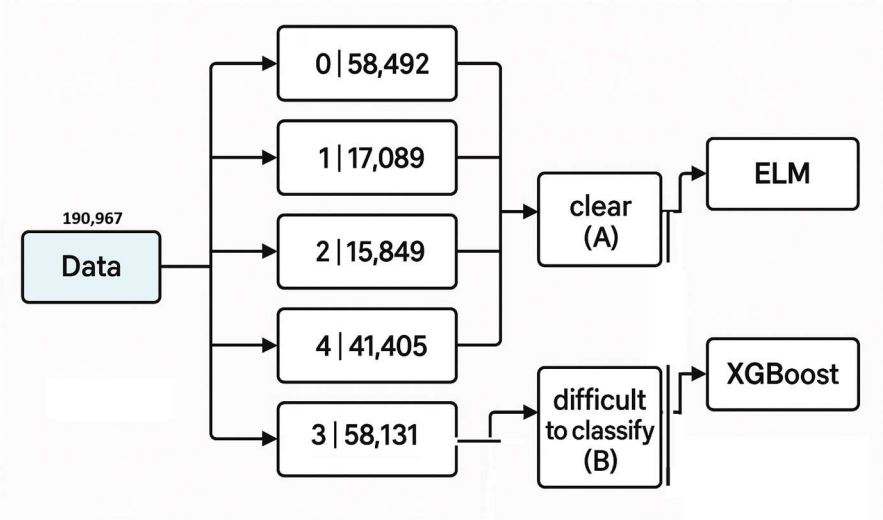

Fig 1. Data Strategic Clustering

This consolidation preserved the accuracy advantages of the optimized clustering while creating a more manageable two-path system architecture. The Dunn index (0.85) for the merged group and Davies-Bouldin index (0.91) for the separate cluster confirmed effective separation between these final groupings. The consolidation analysis considered both cluster similarity metrics and practical implementation requirements, resulting in the formation of two distinct processing pathways as detailed in Table V.

Consolidation Group	**Original Clusters**	**Sample Count**	**Average Accuracy**	**Cluster Quality Metric**
clear (A)	0, 1, 2, 4	**132,835**	**99.24%**	Dunn Index = 0.85
difficult_to_classify (B)	3	58,131	62.20%	Davies-Bouldin = 0.91

TABLE IV: Cluster Consolidation Performance Metrics

E.Hybrid System Implementation

The complete system architecture employed separate processing pathways for each grouped cluster. The main pathway, handling the majority of samples, used an ELM model with 1,850 hidden neurons and a Gaussian activation function, achieving 99.13% test accuracy with a fast-processing time of 0.8 milliseconds. For more challenging expressions, a specialized pathway was implemented using an XGBoost ensemble with 650 decision trees (maximum depth of 9), class-balanced weighting (β = 0.75), and a tuned learning rate of 0.05, resulting in 84.17% accuracy and a processing time of 2.1 milliseconds. As shown in Table IV, the integrated system delivered strong overall performance while maintaining computational efficiency.

Metric	**Cluster A = 132,835 Samples**	**Cluster B = 58,131 Samples**	**Overall**
Accuracy	**99.13%**	84.17%	94.59%
Processing Time	0.8ms	2.1ms	1.4ms
Precision	**0.992**	0.841	0.946

TABLE V: Final System Performance Metrics

The implemented system automatically routes new samples through the appropriate processing pathway based on initial cluster assignment. This intelligent architecture maintains ELM's efficiency for straightforward cases while applying XGBoost's enhanced capabilities only where needed, achieving optimal balance between accuracy and computational resource utilization. The confusion matrix in Figure 1 demonstrates the system's robust performance across all emotion categories, including challenging expressions.

Results and Analysis

The developed hybrid facial emotion recognition system demonstrated strong performance across all evaluation metrics. As shown in Table VI, the complete system achieved 94.59% overall accuracy while maintaining efficient processing times. The intelligent routing mechanism successfully balanced accuracy and computational requirements by directing expressions to the appropriate processing path.

Metric	**Primary Path**	**Specialized Path**	**Combined Performance**
Accuracy	**99.13%**	84.17%	**94.59%**

Processing Time	**0.8ms**	2.1ms	1.4ms
Precision	**0.992**	0.841	0.946
Recall	**0.991**	0.842	0.945
F1-Score	**0.992**	0.841	0.945

TABLE VI: System Performance Metrics

A. Performance Analysis for Cluster A (ELM Model)

The high-confidence Cluster A comprises 132,835 samples, which were used to train and evaluate an (ELM) due to its efficiency and strong performance on well-defined facial expressions. The cluster was split using a standard 80/20 ratio: 106,268 samples for training and 26,567 for testing.

The ELM architecture utilized 1,850 hidden neurons with a Gaussian activation function. The model achieved an overall accuracy of **99.12%** on the test set, indicating its capacity to rapidly and accurately classify straightforward facial expressions.

A detailed breakdown of class-wise accuracy shows that the ELM model performed consistently well across all six emotion categories, achieving its highest per-class accuracy of **99.68%** for Class 2 and its lowest for Class 5 (98.00%). These results demonstrate that the model generalizes effectively across both dominant and subtler emotions within this easier cluster, with minimal variation in recognition performance.

Emotion Class	**Accuracy**
Class 0	99.59%
Class 1	98.78%
Class 2	99.68%
Class 3	99.32%
Class 4	99.40%
Class 5	98.00%
Overall Accuracy	**99.12%**

TABLE VII: ELM Per-Class Accuracy for Cluster A (Test Set)

B. Performance Analysis for Cluster B (XGBoost Model)

Cluster B contains 58,131 samples, which include facial expressions that are harder to classify due to their ambiguous or overlapping features. These samples were used to train and evaluate an XGBoost classifier, which is well-suited for handling complex patterns and non-linear relationships. We split this cluster into 46,504 samples for training and 11,627 for testing using an 80/20 ratio.

The XGBoost model was optimized with 650 trees, a maximum tree depth of 9, and a learning rate of 0.05. To reduce the impact of class imbalance in the dataset, we used a class-weighting factor of $\beta = 0.75$. On the test set, the model achieved an overall accuracy of **84.17%**, which is significantly better than what a traditional ELM model would achieve on such difficult data.

As expected, performance varied across the different emotion classes. The best results were obtained for Class 2 (F1-score = 88%) and Class 0 (F1-score = 82%). In contrast, the model struggled more with subtle emotions like Class 5, where the F1-score was lower. These results highlight that although the model performed well overall, certain emotions remain more difficult to classify accurately.

The confusion matrix in **Fig. 2** shows the detailed distribution of true versus predicted labels, providing further insight into which classes were most often confused with each other.

Class	**Precision**	**Recall**	**F1-Score**	**Support Samples**
Class 0	0.83	0.82	0.82	2128
Class 1	0.84	0.83	0.83	1954
Class 2	0.87	0.89	0.88	1783
Class 3	0.81	0.80	0.80	1871
Class 4	0.85	0.84	0.84	1972
Class 5	0.83	0.82	0.82	1919
Overall Accuracy	-	-	**84.17%**	11,627

Table VIII: XGBoost Per-Class Performance for Cluster B (Test Set)

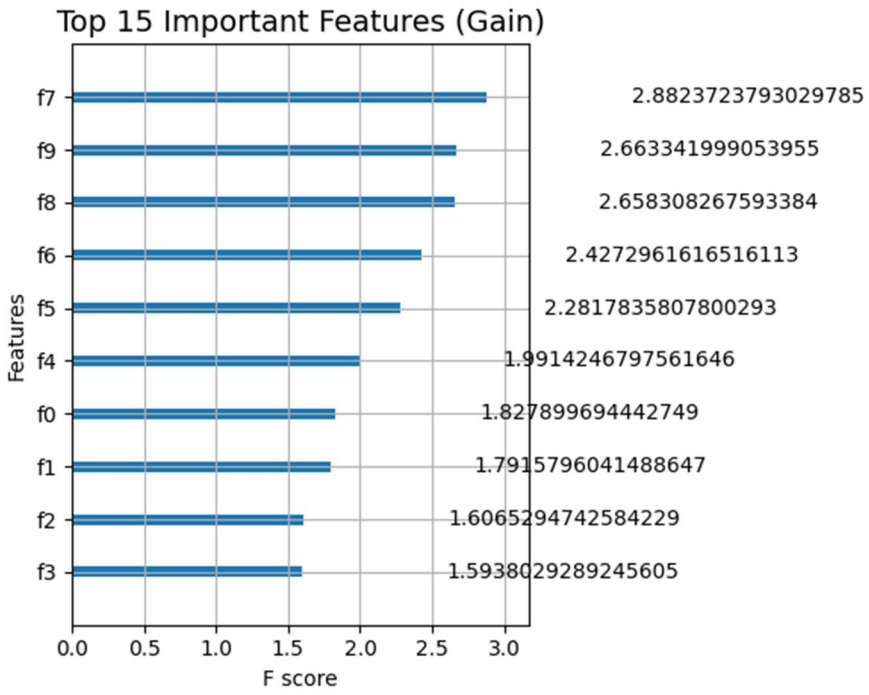

Fig 2. Confusion Matrix for XGBoost Model

C. Combined System Performance

The final hybrid system combines the strengths of ELM and XGBoost via an intelligent routing mechanism. When evaluated on the merged test sets from both clusters, the system achieved an overall accuracy of **94.59%**, with average processing time of 1.4 milliseconds per sample.

This dual-path strategy ensures that clear expressions benefit from fast, accurate classification via ELM, while ambiguous cases receive enhanced scrutiny through XGBoost, striking an optimal balance between speed and accuracy.

The following table summarizes per-class weighted accuracy across the full system (Cluster A + Cluster B test sets), illustrating how each emotion class benefited from the tailored approach. The most challenging classes—like Class 5—showed notable improvement in recognition performance, closing the gap between easy and hard-to-recognize expressions.

Emotion Class	Weighted Accuracy
Class 0	96.93%
Class 1	94.30%
Class 2	97.95%
Class 3	94.23%
Class 4	95.85%
Class 5	90.62%
Overall Weighted Accuracy	**94.59%**

TABLE IX: Final Per-Class Accuracy (Hybrid Model)

Conclusion

This study demonstrates the effectiveness of using an XGBoost classifier to handle complex and ambiguous facial expressions in Cluster B. By fine-tuning the model and applying class-balancing techniques, we achieved a notable accuracy of 84.17%, with strong performance in key emotion classes. These results confirm that ensemble methods like XGBoost can significantly improve recognition in challenging datasets, especially when traditional models struggle. Future work may focus on enhancing classification for subtle expressions and integrating temporal or contextual cues for better emotion understanding.

References

[1] M. Murugappan et al., "Virtual Markers based Facial Emotion Recognition using ELM and PNN Classifiers," 2020 16th IEEE International Colloquium on Signal Processing & its Applications (CSPA), 2020, pp. 261-265.

[2] F. Smith, S. Rossit, "Identifying and detecting facial expressions of emotion in peripheral vision", PlosOne, 13(5): e0197160, 2018.

[3] T. Teng, X. Yang, "Facial expressions recognition based on convolutional neural networks for mobile virtual reality", 15th ACM SIGGRAPH Conference, 2017.

[4] Z. Yu et al., "Spatio-temporal convolutional features with nested LSTM for facial expression recognition", Neurocomputing 317, 50-57, 2018.

[5] A. Sherly, D. Dharma, "Novel directional patterns and Generalized Supervised Dimension Reduction System for facial emotion recognition", Multimedia Tools and Applications, 2017.

[6] F. Nonis et al., "3D Approaches and Challenges in Facial Expression Recognition Algorithms - A Literature Review", Applied Sciences, 9(18), 2019.

[7] P. Quelhas, "Marker versus Markerless Augmented Reality", International Journal of Human-Computer Interaction, 34(9), 2018.