Facial Expression Recognition Using CNNs and Attention Mechanisms on Modified FER2013 Dataset

Original Scientific Paper

Aqeel Majeed Breesam*
Institute of Medical Technology/ Baghdad
Middle Technical University
Baghdad, Iraq
aqeelmajeed@mtu.edu.iq

Roaa Safi Abed Alah
Department of Intelligent Medical Systems
University of Information Technology and Communications
Baghdad, Iraq
rouaa.safi@uoitc.edu.iq

Mina Faris Ali Alnaimy
College of Engineering, Department of Media and Communications Technology
University of Information Technology and Communications
Baghdad, Iraq
mina.alnaimy@uoitc.edu.iq

*Corresponding author

Abstract *– Facial Expression Recognition (FER) is an important component of affective computing and human-computer interaction, yet it is difficult to apply in practice because of the variations in illumination, pose variation, occlusions, low-resolution images, and skew in the classes. To overcome these issues, this paper introduces FER-CBAMNet, a small and attention-based convolutional neural network that uses a lightweight residual network with Convolutional Block Attention Modules (CBAM) to focus on discriminative facial regions, including the eyes, eyebrows, and mouth, and ignore irrelevant background information. The model is trained with a powerful pipeline that includes a large amount of data augmentation, normalization, and class-balanced regularization on the FER-2013 dataset under a modified evaluation protocol. Experimental evidence shows that FER-CBAMNet has a mean accuracy of 96.58% and a macro F1-score of 0.963, with low inter-run variability, significantly surpassing current state-of-the-art models on the same test set, such as standard CNNs, lightweight CNN-KNN hybrids (~75%), and transformer-based models (~71\). Ablation experiments reveal that the addition of CBAM adds significant performance improvement, enhances accuracy, and balances the recognition of all seven emotion categories. Also, Grad-CAM visualizations demonstrate that the model makes decisions based on semantically meaningful parts of the face, which are interpretable and diagnostically meaningful. These results suggest that attention-enhanced lightweight CNNs are highly accurate, robust, and transparent, and FER-CBAMNet is highly suitable for real-world and resource-limited applications.*

Keywords: *facial expression recognition, convolutional neural network, deep learning, FER2013*

Received: February 11, 2026; Received in revised form: May 5, 2026; Accepted: May 6, 2026

1. INTRODUCTION

To communicate effectively and naturally, it is imperative to understand human emotions since facial expressions communicate important details other than verbal messages. Visual cues are a very important part of interpersonal interaction as they help see intentions, empathy, and mind states. Facial Expression Recognition (FER) (an automatic detecting method of facial expressions) has therefore become an important field of study in computer vision, affective computing, and artificial intelligence, allowing its use in human-computer interaction, medical diagnostics, driver monitoring, and intelligent surveillance systems [1-3].

The ability to detect facial emotions is a challenge, even though deep learning (DL) advances at a very rapid pace. The real-world images are incredibly variable in terms of illumination, pose, occlusion, and background mess, as well as in age, gender, and ethnicity. Conventional feature-based methods that were done manually were unable to handle these inconsistencies. FER was revolutionized by Convolutional Neural Networks (CNNs), which learned hierarchical visual representations directly on the data, and were better than classical pipelines [4, 5]. Nonetheless, conventional CNNs would consider all spatial locations identically, although some corners of the face, such as those of the eyes, mouth, and eyebrows, are the ones that bear most of the discriminative information to detect emotions.

Although recent research has proposed attention mechanisms to alleviate this drawback, current FER methods tend to have at least one of three unsolved problems: (i) dependence on deep and computationally intensive architectures that are difficult to scale out to, (ii) weakened robustness to in-the-wild datasets with noise and class-imbalance like FER-2013, and (iii) lack of transparency, since most of the successful models are black boxes, which do not give interpretable evidence to their predictions. Therefore, there is a noticeable gap in research regarding FER models that can accurately recognize emotions, operate efficiently, and provide explanations suitable for real-world applications.

Recently, attention mechanisms have become a promising tool to deal with the homogeneous processing of spatial features by selectively focusing on salient regions and inhibiting irrelevant background information. The Convolutional Block Attention Module (CBAM) [6] is one of them, providing an appealing tradeoff between representational capacity and computational costs through the sequential application of channel-wise and spatial attention. Nevertheless, little has been done on integrating lightweight attention modules into small FER architectures combined with systematic interpretability analysis, especially on noisy and imbalanced datasets, like FER-2013.

Parallel to this, due to the lack of interpretability of deep FER models, questions have been raised about their use in safety-critical and ethically sensitive uses. Even though post-hoc explanation methods like Gradient-weighted Class Activation Mapping (Grad-CAM) [7] have been introduced, many of these methods are not considered analytical or diagnostic instruments of a model's behavior and failure modes but instead seen as auxiliary visualizations. This also gives the incentive to have FER frameworks in which attention mechanisms and interpretability are reciprocally used to increase predictive reliability and transparency.

Even with the recent advances, literature shows that there is still a gap in methodology in recognizing facial expressions in realistic conditions. Numerous current FER models enhance the accuracy of classification at the cost of computational efficiency and practical deployability by adding depth to the architecture and increasing the number of parameters. Meanwhile, models that are trained on in-the-wild datasets like FER-2013 are susceptible to label noise, class imbalance, pose variation, occlusion, and illumination variation. Another weakness is that interpretability is often seen as an optional add-on to model design and evaluation, instead of being viewed as a core component of the design and evaluation process. Such constraints drive the necessity of an FER framework which is not only precise but also lightweight, robust and transparent. In line with this, this research aims to train and test an attention-based lightweight CNN to recognize facial expressions that can be effectively used on FER-2013 and give interpretable visual features to its predictions.

The following research questions are answered to guide the study. RQ1: Does a lightweight residual CNN with CBAM enhance the performance of facial expression recognition on FER-2013 without compromising its computational efficiency? RQ2: Does the addition of attention mechanisms enhance the model to focus on the discriminative facial areas that are relevant in emotion recognition? RQ3: Does Grad-CAM-based visual analysis have any meaningful insight into correct and incorrect predictions to enhance interpretability and diagnostic value of the proposed framework? The answers to these questions are that this work will help place performance, efficiency and explainability as collectively optimized design objectives instead of individual assessment criteria.

To address the abovementioned gaps, the current paper introduces a small-sized attention-enhanced CNN model, also known as FER-CBAMNet, that is trained on the unconstrained FER-2013 dataset and is complemented with Grad-CAM-based visual explanations.

The main contributions of this work are summarized as follows:

- A lightweight residual CNN enhanced with Convolutional Block Attention Modules (CBAM) is proposed to selectively focus on emotion-relevant facial regions while suppressing irrelevant background information, improving robustness to pose, illumination, and occlusion variations.
- An effective training scheme is used to reduce the impact of label noise and class imbalance that is intrinsic to FER-2013, which allows consistent performance across all types of emotions.
- Visual explanations with Grad-CAM are scientifically examined to comprehend both accurate and incorrect predictions, which is informative about the workings of a model and justifies clear and reliable application.

The remaining parts of this paper will be organized as follows. Section 2 is a review of the work in question. Section 3 contains the strategy proposed. The findings of Grad-CAM analysis and experiments are shown in

Section 4. In part 5, there is a discussion of the analogy to the state-of-the-art models. The conclusion of the paper is in Section 6.

2. RELATED WORK

Facial Expression Recognition (FER) has come a long way in the last ten years, where handcrafted pipelines are replaced with deep learning (DL) models that can learn discriminative affective representations directly based on data. However, the field still struggles to handle unconstrained imaging, such as changes in lighting, occlusions, poses, low-resolution inputs, and label noise, especially on in-the-wild benchmarks, such as FER-2013 [8]. To put the suggested FER-CBAMNet into perspective, the section provides a review of existing literature in the thematic area and points out limitations on our design decisions.

2.1. TRADITIONAL AND BASIC CNN-BASED FER METHODS

Early FER systems used manually created descriptors, including SIFT, HOG, and LBP, with a traditional classifier. Although these techniques worked quite well in a controlled environment, they tend to be non-generalizable in real-world scenarios because they are sensitive to nuisance effects (e.g., illumination and occlusion) and have a low ability to capture the fine muscle patterns on the face. The introduction of convolutional neural networks (CNNs) was also a significant step toward FER since it was possible to learn feature end-to-end and extract hierarchical representations of the original raw facial images, which resulted in superior recognition compared to classical pipelines [9], [10]. Based on this, transfer learning has been extensively used to take advantage of large-scale pretraining, and research has shown good performance by fine-tuning deep backbones (e.g., VGG, ResNet, DenseNet) to emotion classification [11, 12].

Previous visual recognition research is also helpful in the context of the development of facial analysis techniques. An example is the intelligent fused method of face recognition by Meethongjan *et al*. [13], which uses Voronoi diagrams and wavelet moment invariants, which represents the wider pre-deep-learning focus on handcrafted and fused facial representation descriptors. Similarly, Sharif *et al*. [14] showed that complementary local descriptors like SIFT and BRISK can enhance robustness in the presence of clutter, rotation, and change in illumination in image retrieval tasks. These studies were not intended to be used with FER, but they highlight a common problem that is still pertinent to the analysis of expression: strong visualization in uncontrolled imaging conditions. In more recent times, transfer learning has emerged as a paradigm in affective image understanding. Meena and Mohbey examined transfer-learning-based image sentiment analysis on CK+, FER2013, and JAFFE and demonstrated that the performance of the model is highly contingent on the properties of the dataset and the choice of the backbone. Meena *et al*. [15] also used an InceptionV3-based transfer-learning model to analyze images based on sentiment and achieved 73% accuracy on FER2013, which once again confirms the usefulness of pretrained deep visual encoders in expression-related tasks. Moreover, the FSTL-SA framework further expanded this line of research to few-shot transfer learning on facial-expression-based sentiment analysis, emphasizing the role of transferability and label efficiency in the case of annotated affective data being scarce [16].

Recent studies have also explored hybrid and multi-branch architectures to improve feature representation in facial expression recognition tasks. For instance, Bordjiba *et al*. [17] proposed a jointly learned dual-branch network that combines VGG-based feature extraction with EfficientNet-based transfer learning, followed by feature fusion for classification. Their approach demonstrates that integrating complementary feature representations can enhance recognition performance, particularly when dealing with limited or imbalanced datasets such as FER-2013. Similarly, Nukathati *et al*. [18] proposed a deep learning model of emotion recognition based on video data that concurrently processes audio and visual information, emphasizing the significance of learning and optimization methods of multimodal features in enhancing the recognition rate. These papers also highlight the importance of architectures that trade-off between feature richness, generalization, and computational efficiency in affective computing tasks.

Critical perspective: Deep backbones are generally more accurate but can be computationally expensive and do not necessarily focus on localized discriminative features (e.g., eyes, eyebrows, and mouth) which are crucial to strong FER in the presence of occlusions or background clutter. Additionally, the performance improvements achieved on curated data sets might not be reliably transferred to noisy and imbalanced in-the-wild data sets like FER-2013, which highlights the importance of architectures that enhance spatial selectivity without sacrificing efficiency. Taken together, these studies show the progression from handcrafted feature fusion to transfer-learning-based deep models; however, they also indicate that robustness, data efficiency, and generalization on noisy in-the-wild benchmarks remain unresolved.

2.2. ATTENTION-BASED FER METHODS

Attention mechanisms have risen as a conceptually sound approach to counteracting the homogenous spatial processing of conventional CNNs by highlighting the salient areas on the face and downplaying the salient areas of insignificant information in the background. The end-to-end attentional convolutional network suggested by Minaee *et al*. [19] focused on the informative aspects of a face, selectively attending to

them and enhancing its resistance to pose and partial occlusions, as well as offering visual information about the importance of evidence of a particular class. Multi-region and multi-resolution strategies are also used as complementary directions to form a fusion of global and local facial information; such as, [20] use landmark-guided region extraction with multi-resolution and extensive augmentation to enhance generalization with strong performance on datasets including SFEW, CK +, and KDEF. In addition to architectural adjustments, learning goals have been put forward to overcome the ambiguity and label mismatch. Zhang *et al*. proposed Relative Uncertainty Learning (RUL), a framework that estimates uncertainty using feature-mixup, resulting in increased resistance to ambiguous expressions and noisy labels [21].

Critical perspective: Although they are effective, much of the attention-driven or uncertainty-aware FER systems are reliant on more comprehensive backbones or pipelines of multi-branch training, and may be problematic when deployed in resource-constrained or latency-sensitive environments. Also, the measure of attention is frequently determined mainly by overall gains in accuracy of its aggregate measurement, with little systematic research determining whether the attended areas correspond to neurophysiologically significant expression signals or whether they account for certain modes of confusion between closely related emotions. The constraints inspire lightweight attention modules to enhance spatial channel selectivity at the cost of maintaining compactness.

2.3. LIGHTWEIGHT AND EDGE-DEPLOYMENT FER METHODS

With the increasing popularity of FER applications in practice, much has been done on lightweight architectures and edge deployment. The authors of Ab Wahab *et al*. [22] suggested the use of an EfficientNet-Lite-based method with a KNN classifier to allow FER to run on an embedded device (Raspberry Pi), showing it is possible to achieve real-time inference and competitive FER-2013 accuracy. The use of hybrid optimization-learning models has been investigated to minimize calculations and maximize discrimination; Sikkandar and Thiyagarajan [23] combined a better cat swarm optimization with deep CNN feature extraction and an ensemble classification to achieve better results in terms of accuracy and efficiency on multiple datasets. Application-oriented FER systems also exhibit scaled domain limitations, e.g., Pabba and Kumar [24] built a FER-based engagement monitoring system in classroom settings using a CNN trained on its own dataset with an augmentation of public datasets, which could estimate the group level on the fly.

Beyond FER specifically, the broader visual understanding literature has increasingly emphasized lightweight, generalizable, and deployment-oriented model design. In their survey of multimodal human action recognition, Bichi and Abdullahi pointed to a shift in the field-wide perspective of benchmark-driven accuracy to computationally efficient, scalable, and deployment-ready systems. Though this view is not about FER but about HAR, it is directly applicable to emotion recognition in real-life contexts, where the efficiency of inference, subject-to-subject generalization, and performance in realistic conditions are also significant. Likewise, Chaudhari *et al*. explored the concept of 1D CNNs to detect depression based on visual cues on the DAIC-WOZ dataset and demonstrated that the dropout and learning rate could be tuned to achieve significant improvements in generalization. Although the task is not similar to categorical facial expression recognition, the work is useful in showing that affect-related visual analysis can be effectively served by compact architectures and constrained optimization decisions, and not necessarily architectural complexity.

Critical perspective: Lightweight techniques typically have a trade-off between size and precision, especially on in-the-wild data where label noise, occlusion, and low resolution negatively affect performance. Additionally, edge-oriented literature tends to focus on runtime feasibility but has little to no interpretability analysis, which is crucial to debug failures and instill confidence in actual deployments.

2.4. INTERPRETABILITY AND EXPLAINABLE FER

With the continued use of FER systems in socially sensitive and high-risk areas, the issue of interpretability has become a major concern. The unresolved issues of surveys are demographic variability, dataset imbalance, label noise, and the absence of a clear evaluation process [25]. Post-hoc explainability techniques offer a more feasible avenue to understand deep models by localizing decision-making evidence; widely used methods include Class Activation Mapping (CAM) [26] and Gradient-weighted Class Activation Mapping (Grad-CAM), which produce class-specific heatmaps to indicate influential regions.

Recent efforts have broadened explainability in FER beyond heatmap visualization to more diagnostic and semantic goals. Punuri *et al*. [27]. Introduced a ranking-based FER architecture with a VGG16 transfer-learning backbone and Layer-wise Relevance Propagation (LRP) to identify contributory pixels and predict emotion-intensity ranks on several datasets (JAFFE, CK+, KDEF, and AffectNet) with annotator-supplied ground truth to label intensity. Mohana *et al*. [28] introduced XAI-DSCSA, a deep semi-supervised convolutional sparse autoencoder designed to reduce reliance on large, labeled datasets by learning latent representations from unlabeled samples, followed by supervised fine-tuning; they further employed Grad-CAM and image-LIME to interpret prediction outcomes on CK+ and JAFFE. Ji *et al*. [29]. Presented an interpretable hybrid framework combining ResNet18 feature extraction with an SVM classifier and used CAM and t-SNE to visualize class

separability and decision evidence, reporting results on FER2013, JAFFE, and CK+.

Critical perspective: Although these studies underscore the value of explainability for trust and analysis, many are validated primarily on controlled or higher-quality datasets (e.g., CK+, JAFFE) and/or rely on large pre-trained backbones or multi-stage pipelines (e.g., transfer learning with VGG16, semi-supervised auto-encoding, or CNN–SVM hybrids). Consequently, their applicability to low-resolution, noisy, and class-imbalanced in-the-wild benchmarks such as FER-2013 remains less explored. This motivates the development of compact, end-to-end FER models that integrate attention for robust representation learning and employ explainability to analyze both correct and failure cases under realistic conditions.

2.5. POSITIONING OF THIS WORK

Collectively, prior work has advanced FER through deep CNN backbones and transfer learning [30, 31], attention and multi-region feature learning [19, 32], uncertainty-aware training [21], edge-oriented efficiency strategies [19, 20], and emerging explainable FER frameworks [27, 33]. Nevertheless, there remains a need for approaches that jointly provide: (i) strong performance on noisy and class-imbalanced in-the-wild datasets such as FER-2013, (ii) compact architectures suitable for scalable deployment, and (iii) transparent decision-making supported by established explanation techniques. In order to meet these requirements, we introduce FER-CBAMNet, a lightweight residual CNN with Convolutional Block Attention Modules (CBAM) [6] and complemented with Grad-CAM-based interpretability [7], which allows us to not only achieve better recognition but also to study the model behavior systematically on FER-2013.

3. PROPOSED APPROACH

The proposed FER-CBAMNet framework proposes a small but effective convolutional neural network that can improve the facial emotion recognition performance by selective attention and lightweight design. As shown in Fig. 1, the pipeline starts with the input facial images of the FER-2013 dataset, then the preprocessing phase that standardizes the image resolution, normalization, and augmentation to enhance generalization. The network uses a Batch Normalization-based initial convolutional layer and ReLU activation at the low-level features level, and then uses a series of four residual blocks with Convolutional Block Attention Modules (CBAM). These attention modules dynamically adjust the responses of spatial and channel-wise features and enable the model to concentrate on the discriminative components of faces, e.g., the eyes, mouth, and eyebrows, and disregard irrelevant background features. The feature maps are then extracted, pooled with the Global Average Pooling, flattened, and regularized using a Dropout (p = 0.3) layer, and finally classified using a fully connected layer with softmax activation. It is a hierarchical model that trades efficiency and interpretability and produces correct emotion predictions in seven categories, such as angry, disgusted, fear, happy, neutral, sad, and surprise.

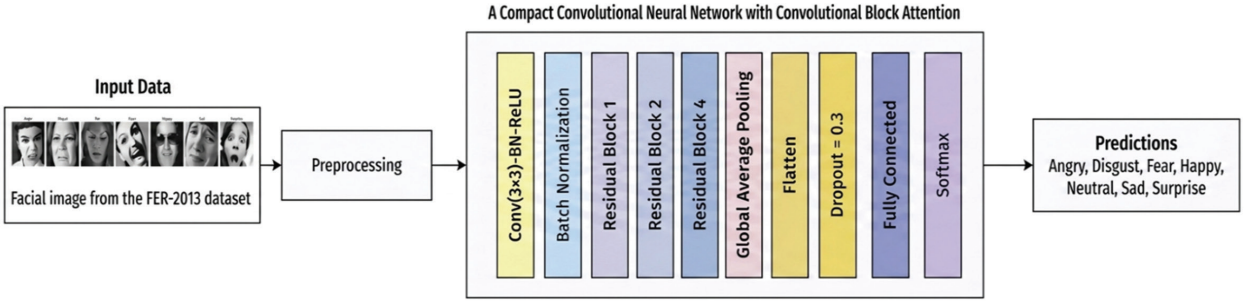

Fig. 1. The Proposed Approach

3.1. DATASET

The dataset that will be used in this paper is the FER-2013 dataset, which has become a standard in FER tasks. The data is 48x48 pixels grayscale facials, each image depicting a single and only human face that has been automatically centered and cropped to give uniform centering and scale. The images are marked with one of seven categories of emotion: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral, which are associated with discrete affective states. FER-2013 data set is composed of 48x48 grayscale facial pictures that are labeled with one of seven emotion classes, namely Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral.

The FER-2013 dataset contains 28,709 training images and two predefined evaluation subsets: PublicTest (3,589 images) and PrivateTest (3,589 images). In the original benchmark protocol, the PublicTest subset is typically used for validation during model development, while the PrivateTest subset is reserved for final performance evaluation.

In this work, we adopt a modified evaluation protocol. Specifically, the PublicTest and PrivateTest subsets are merged to form a unified test set of 7,178 images. This design choice is motivated by the need to obtain a

more statistically stable and comprehensive evaluation across all emotion categories, particularly given the class imbalance and variability inherent in FER-2013.

Out of this training data, we developed a validation subset with 3,445 images (12% of the training data) that is stratified to maintain the distribution of classes. This guarantees that a good model selection process is performed reliably, but no overlap of training, validation and test data is allowed, thus avoiding data leakage.

Fig. 2 shows the representative samples of the FER-2013 dataset; it represents the seven classes of emotions, which are angry, disgust, fear, happy, neutral, sad, and surprise. The size of every image is a 48x48 grayscale crop, which is brought to a uniform size and positioning of faces using automated alignment. The figure emphasizes how the facial expressions are diverse in terms of pose, light, and strength, that the dataset has variability by nature, and is complex in the real world.

Fig. 2. A representative sample from the FER-2013 dataset

Fig. 3 depicts the distribution of classes of the FER-2013 training set, where the distribution between the seven emotion categories is evidently biased. The happy category has the highest number of samples, with over 7,000 images, and the state of disgust has been greatly underrepresented, with less than 1,000 samples. Other categories, like angry, fear, neutral, and sad, are moderate in their representation, and the proportion of surprised is relatively low. This distorted distribution can result in bias within model training, as well as disadvantages for minority classes in generalization.

Fig. 3. Class distribution of the modified FER-2013

3.2. DATA PREPROCESSING

A crucial step in the proposed framework is the data preprocessing pipeline, which standardizes input images, reduces noise, and enhances model generalization. The FER-2013 data is a collection of 48x48-pixel grayscale images, each of which represents a single human face with one of seven emotion labels. To ensure compatibility with convolutional backbones pretrained on ImageNet, all grayscale images $I_g(x, y)$ were converted to three-channel tensors by replicating the grayscale intensity across the red, green, and blue channels:

$$I_{rgb}(x, y, c)=I_g(x,y),\ c\in\{1,2,3\} \quad (1)$$

where $I_g(x, y)$ denotes the grayscale intensity at the spatial location (x, y), $I_{rgb}(x, y, c)$is the value in the RGB image at channel c, and cindexes the red, green, and blue channels, respectively. This transformation preserves the original luminance while matching the expected input dimensionality of pretrained models.

Then all the images were resized to 64x64 pixels by bilinear interpolation. This minor upsampling trades off computational efficiency and spatial granularity, enabling the model to represent finer-grained facial features, such as eyebrow shape, squinting of the eyes, or lips contracting, which are important to distinguish subtle expressions.

To enhance resilience to real-world variability, a full set of stochastic augmentations was used in training:

- Random horizontal flip with probability 0.5
- Random rotation within ±15°
- Brightness and contrast adjustments (±15%)
- Additive Gaussian noise:

$$I'=I+\epsilon,\epsilon\sim N(0,\sigma^2) \quad (2)$$

where I is the original image, I' is the noisy image, $\epsilon$is Gaussian noise sampled from a normal distribution with zero mean and variance σ^2, and σ is set to 0.03.

Random erasing: 0.25 probability, small rectangular areas of the image are masked with zeros to induce the model to utilize distributed contextual information instead of being dependent on a single discriminative region.

These extensions model pose, lighting, partial occlusion, and camera angle variations, which are typical in natural scenes, but are not well represented in controlled datasets.

All pixel intensities were brought to zero mean and unit variance after augmentation:

$$x'=(x-\mu)/\sigma \quad (3)$$

In which x is the original pixel intensity, x is the normalized value, x is the mean of the input distribution, and x is the standard deviation. Normalization stabilizes the gradient flow and enhances optimization dynamics.

Lastly, to guarantee reproducibility and prevent data leakage, a stratified sampling strategy was used to partition the dataset into training, validation, and test sets. In accordance with the modified evaluation protocol described in Section 3.1:

- Official Train partition: 28,709
- Validation (12% sampled from Train): 3,445
- Training used for fitting: 25,264
- Test (PublicTest+PrivateTest): 7,178

Stratification preserves the distribution of classes within subsets, so that the evaluation is a true reflection of the learning statistics.

It is important to emphasize that this partitioning strategy follows a customized protocol, rather than the original FER-2013 benchmark configuration. This design ensures strict separation between training, validation, and testing data, while providing a stable and reproducible evaluation framework.

3.3. PROPOSED MODEL

The proposed FER-CBAMNet is a residual convolutional neural network with Convolutional Block Attention Modules (CBAM), which is capable of dynamically highlighting salient emotional features in the facial regions at a low cost. This architecture is a balance between representational richness, interpretability and efficiency, and is therefore appropriate to real-time facial emotion recognition (FER) on resource-constrained hardware.

The input is $X \in R^{3\times64\times64}$, a grayscale facial image copied into three channels to comply with standard convolutional backbones. The network begins with a stem convolutional block for early feature extraction:

$$X_1 = \phi(BN(W_s X)) \tag{4}$$

where W_{sis} a 3×3 convolutional kernel with 32 output channels, $BN(\cdot)$ represents batch normalization, $\phi(\cdot)$ is the ReLU activation, and X_1 is the output capturing primitive visual features such as edges and gradients.

The network proceeds through four consecutive residual stages, each increasing the number of feature channels (32→32, 32→64, 64→128, 128→256). Each BasicBlock comprises two convolutional layers, batch normalization, ReLU activations, and a shortcut connection to stabilize gradient flow:

$$Y_t = \phi(BN(W_2\, \phi(BN(W_1 X_t)))) + S(X_t) \tag{5}$$

where X_t is the input to the t-th residual block, W_1 and W_2 are the convolution kernels, and $S(\cdot)$ is either an identity mapping or a 1×1 projection when downsampling (stride=2 for Stages 2–4). Spatial resolution decreases progressively from 64×64 to 8×8.

Each residual block is augmented with a CBAM to improve attention to emotionally discriminative facial regions such as the eyes, eyebrows, and mouth.

The forward propagation can be summarized as:

$$X_{Stem} \rightarrow X_1 \rightarrow \text{Stage 1–4 (CBAM)}\ X_4 \rightarrow \text{GAP} \rightarrow z \rightarrow \text{Dropout} \rightarrow \hat{y} \tag{6}$$

where X_4 is the output of the final residual stage, GAP denotes global average pooling, $z \in R^{256}$ is the embedding vector, and $\hat{y}$ is the seven-class output.

CBAM sequentially applies channel and spatial attention to enhance discriminative feature learning:

$$f_{avg} = GAP(F), f_{max} = GMP(F) \tag{7}$$

where $F \in R^{C\times H\times W}$ is the input feature map. The descriptors are passed through a shared MLP to produce the channel attention mask:

$$m_c = \sigma(MLP(f_{avg}) + MLP(f_{max})) \tag{8}$$

$$F_c = m_c \odot \mathrm{F} \tag{9}$$

Spatial attention then highlights discriminative regions:

$$m_s = \sigma(f^{7\times7}([AvgPool_c(F_c); MaxPool_c(F_c)])) \tag{10}$$

where F_{sis} the refined feature map emphasizing emotionally salient facial areas, even under occlusion or varying illumination.

After the final residual stage, GAP reduces each feature map to a single scalar, forming a 256-dimensional embedding vector z:

$$z_k = \frac{1}{H\times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_k(i,j) \tag{11}$$

This embedding captures salient facial features emphasized by CBAM and is regularized with dropout (p=0.3). The fully connected layer then maps zto the seven-class output:

$$\hat{y} = \text{Softmax}(W_c z + b_c), \hat{y} \in (0,1)^7 \tag{12}$$

These embeddings are interpretable and can be visualized using Grad-CAM to highlight facial regions influencing the model's prediction.

The network is trained end-to-end using a weighted Focal Loss (γ=1.5, label smoothing=0.05) to handle class imbalance. Class weights are computed from the training data. The AdamW optimizer (lr=3×10^{-4}, weight decay=1×10^{-4}) is used with a CosineAnnealingLR scheduler over 25 epochs. Data augmentation includes random horizontal flips, small rotations (±15°), brightness/contrast adjustment, Gaussian noise, and random erasing to improve robustness to real-world variations. Mixed-precision training is applied using GradScaler on a GPU for efficient convergence.

The total number of trainable parameters is approximately 1.239 million (M)

The FER-CBAMNet architecture is detailed in Table 1.

Table 1. Model Architecture

Component	Description
Input	Grayscale images copied to 3 channels, resized to 64×64 pixels
Stem Conv	Conv2d: 3→32 channels, kernel 3×3, stride 1, padding 1; BatchNorm2d; ReLU
Residual Blocks (BasicBlock)	4 stages: (32→32, 32→64, 64→128, 128→256 channels), each with 2 Conv2d + BatchNorm + ReLU + identity/shortcut, stride=2 for downsampling in stages 2–4
CBAM Modules	Channel and spatial attention applied sequentially at each block; reduction ratio r=16; kernel_size=7 for spatial attention
Pooling & Embedding	Global Average Pooling (GAP) → 256-dimensional embedding; Dropout p=0.3
Classifier	Fully connected layer (256→7 classes) with Softmax
Loss Function	Weighted Focal Loss (γ=1.5, label smoothing=0.05)
Optimizer	AdamW (lr=3e-4, weight decay=1e-4)
Scheduler	CosineAnnealingLR
Training Procedure	Mixed-precision with GradScaler (CUDA), class-weighted loss, extensive data augmentation

4. EXPERIMENTS AND RESULTS

To evaluate the performance of the proposed facial emotion recognition (FER) model comprehensively, several quantitative performance indicators were used. These measures are all aimed at evaluating the overall recognition rate, as well as the ability of the model to label minority classes correctly and achieve balanced predictive behavior across all emotional categories. The evaluation measures are calculated based on the confusion matrix of the model predictions and ground-truth labels [34, 35].

Let TP_c, TN_c, FP_c, and FN_c represent the numbers of true positives, true negatives, false positives, and false negatives for a certain class c, respectively. The evaluation measures are stated below.

Accuracy is a measure of the percentage of samples that are properly classified among all instances in the test set:

$$\text{Accuracy} = \frac{\sum_{c=1}^{K} TP_c}{\sum_{c=1}^{K}(TP_c+FP_c+FN_c)} \tag{13}$$

where TP_c, FP_c, and FN_c denote the true positives, false positives, and false negatives for class c, respectively, and K is the total number of emotion categories. While accuracy provides a general indication of model performance, it may be biased in datasets with class imbalance.

Precision evaluates the reliability of positive predictions by measuring the fraction of correctly identified samples among all samples predicted as belonging to the class c:

$$\text{Precision}_c = \frac{TP_c}{TP_c+FP_c} \tag{14}$$

where $Precision_c$ denotes the precision for the class c, TP_c denotes the number of true positives for the class c, and FP_c denotes the number of false positives for the class c. A high precision value indicates that the model makes few false positive predictions for that class.

Recall measures the model's ability to correctly detect all instances of a class c:

$$\text{Recall}_c = \frac{TP_c}{TP_c+FN_c} \tag{15}$$

where $Recall_c$ denotes the recall for class c, TP_c denotes the number of true positives for the class c, and FN_c denotes the number of false negatives for the class c. When the recall is high, this means that the model has been able to capture the majority of the relevant cases, and this is important when it is costly to miss some of the emotions.

The F1-score is the harmonic mean of precision and recall, giving a balanced assessment when both measures are crucial:

$$\text{F1}_c = 2 \times \frac{\text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{16}$$

where $F1_c$ denotes the F1-score for class c, $Precision_c$ denotes the precision for the class c, and $Recall_c$ denotes the recall for class c. The F1-score is effective in partially overcoming the potential for misleading interpretation of precision or recall measures, especially in unbalanced emotion datasets such as FER-2013.

Macro-averaging is implemented to address label imbalance. This computes the unweighted average across all classes:

$$\begin{aligned} \text{Macro_F1} &= \tfrac{1}{K}\textstyle\sum_{c=1}^{K} \text{F1}_c\,, \\ \text{Macro_Precision} &= \tfrac{1}{K}\textstyle\sum_{c=1}^{K} \text{Precision}_c \\ \text{Macro_Recall} &= \dfrac{1}{K}\displaystyle\sum_{c=1}^{K} \text{Recall}_c \end{aligned} \tag{17}$$

where K denotes the total number of classes, F1c denotes the F1-score of class c, $Precision_c$ denotes the precision of the class c, and $Recall_c$ denotes the recall of class c. Macro-averaged metrics treat all classes equally, preventing most emotions from dominating the overall performance evaluation.

4.1. RESULTS

To evaluate the effectiveness of the proposed FER-CBAMNet model, the experiment was conducted on the FER-2013 dataset to identify the effectiveness of the model in identifying seven different categories of emotions, namely, angry, disgust, fear, happy, neutral, sad, and surprise. In this section, the quantitative data concerning the results of the test set are provided with the help of the interpretative treatment, which hints at the discriminative power of the model and its strength, and the comparative advantage.

As illustrated in Table 2, the model performs highly and consistently in all 10 of the runs, and the accuracy rates are between 96.18 per cent and 97.20 per cent, which demonstrates that the model has a high over-

all classification ability. The scores of Macro-Precision, Macro-Recall, and Macro-F1 are all approximately 0.96, indicating that the model treats the various classes equally without a lot of bias towards a particular class. The slight upward difference between the Weighted-F1 and the Macro-F1 shows that there is a slight imbalance of the classes, however, which the model handles quite effectively. In general, the fact that there is not much fluctuation between runs reveals the strength and reliability of the model.

The learning curves in Fig. 4 illustrate the training dynamics of the FER CBAMNet model over 50 epochs. The loss curves show a steady decrease in both training and validation loss, with training loss consistently lower than validation loss. This implies that the model continues to minimize errors and still generalizes to unseen data. Notably, the validation loss does not deviate or level off too soon, indicating that the model is not overfitting or underfitting severely.

Table 2. Overall Performance and Stability of the Classification Model across Multiple Runs

Run	Accuracy (%)	Macro-Prec	Macro-Rec	Macro-F1	Weighted-F1
1	96.72	0.9637	0.9638	0.9637	0.9649
2	96.25	0.9605	0.9614	0.9609	0.9635
3	96.23	0.9609	0.9630	0.9619	0.9627
4	96.18	0.9636	0.9656	0.9646	0.9676
5	96.28	0.9675	0.9617	0.9645	0.9676
6	96.87	0.9645	0.9619	0.9631	0.9668
7	96.54	0.9613	0.9612	0.9612	0.9638
8	96.68	0.9649	0.9671	0.9659	0.9693
9	96.86	0.9628	0.9615	0.9621	0.9643
10	97.20	0.9615	0.9639	0.9627	0.9644

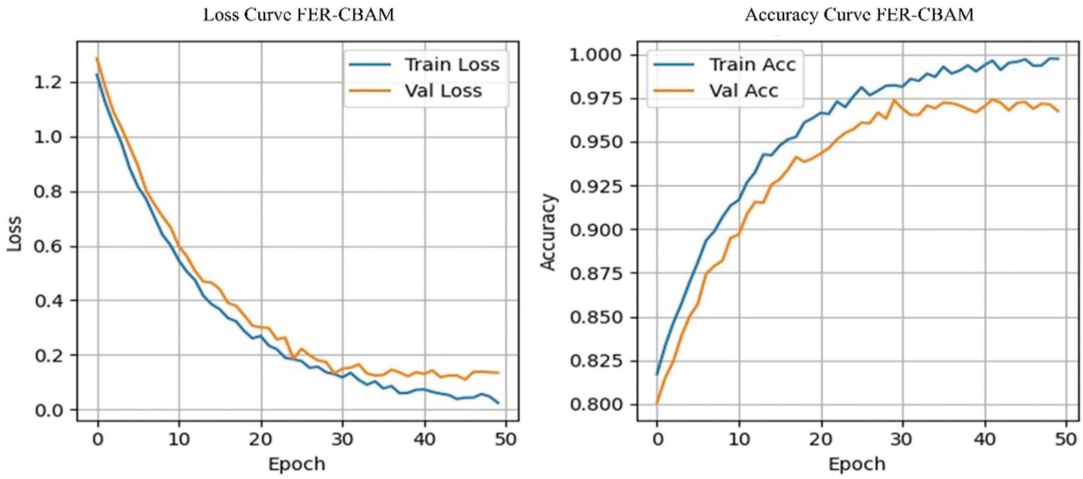

Fig. 4. Training and validation loss/accuracy curves

The trend of the accuracy curves is complementary: the training and validation accuracy are growing steadily with the number of epochs, approaching high values (>0.95). Training accuracy remains slightly higher than validation accuracy, which is expected due to the generalization gap. The close alignment of the two curves demonstrates that the model learns effectively without memorizing the training set.

Table 3 represents the performance statistics of the model averaged between all the runs and demonstrates that the results are good and very consistent. The overall predictive performance of the model is supported by the mean accuracy of 96.58%, although the standard deviation is low (0.34), demonstrating that the variability among runs is minimal. Equally, the Macro-Precision, Macro-Recall, and Macro-F1 scores are almost close to the same means (0.963) and extremely low standard deviations, which indicates a leveled and consistent behavior across classes. The score of the Weighted-F1 is a bit better, indicating that the frequency of classes is well managed without unfairness. Generally, the fact that the range between minimum and maximum values is small indicates that the model is sound and dependable, having a consistent performance.

Table 3. Statistical Summary of Model Performance and Robustness across Runs

Metric	Mean	Std	Min	Max
Accuracy (%)	96.58	0.34	96.18	97.20
Macro-Precision	0.9631	0.0022	0.9605	0.9675
Macro-Recall	0.9631	0.0020	0.9612	0.9671
Macro-F1	0.9631	0.0016	0.9609	0.9659
Weighted-F1	0.9655	0.0022	0.9627	0.9693

The results of Table 4 provide the class-wise performance of the proposed framework in terms of accuracy, recall, and the F1-score. Every single macro-averaged precision, recall, and F1-score is at 0.963 with minimal standard deviations, which means that the classification performance is very stable and well-balanced in all emotion categories. Some of the classes include Happy (97.4%), Surprise (96.8%), and Fear (96.6%), and it implies that

the model can learn better those expressions that have strong visual cues and unique facial muscle movements. Meanwhile, F1-scores of the other expressions, Angry, Sad, and Neutral, are also relatively high (around 96%), which means that the model can be adapted to more subtle or ambiguous expressions. Although Disgust is typically underrepresented, it still attains a solid F1-score of 95.2%, highlighting the effectiveness of the proposed framework in handling class imbalance and maintaining reliable performance across all emotion classes.

Table 4. Test Performance Metrics for the FER-CBAMNet Model

Class	Precision	Recall	F1-Score
Angry	0.966 +/- 0.009	0.961 +/- 0.006	0.963 +/- 0.004
Disgust	0.949 +/- 0.008	0.956 +/- 0.008	0.952 +/- 0.004
Fear	0.960 +/- 0.011	0.972 +/- 0.006	0.966 +/- 0.005
Happy	0.979 +/- 0.009	0.970 +/- 0.009	0.974 +/- 0.006
Neutral	0.954 +/- 0.005	0.964 +/- 0.007	0.959 +/- 0.005
Sad	0.959 +/- 0.007	0.957 +/- 0.005	0.958 +/- 0.005
Surprise	0.975 +/- 0.008	0.961 +/- 0.009	0.968 +/- 0.006
Macro Avg	0.963 +/- 0.002	0.963 +/- 0.002	0.963 +/- 0.002

Fig. 5 depicts the confusion matrix of the FER-CBAMNet model on a test sample. There is an excellent diagonal dominance in the matrix, which proves the high classification reliability in all categories. Minor misclassifications are mainly between similar expressions that are visually similar, like fear vs. surprise and angry vs. disgust, which have common facial muscle activation. As an example, only 34 out of all samples of the 'angry' were wrongly recognized as 'disgust', and 36 out of all samples of 'fear' were recognized as 'happy'.

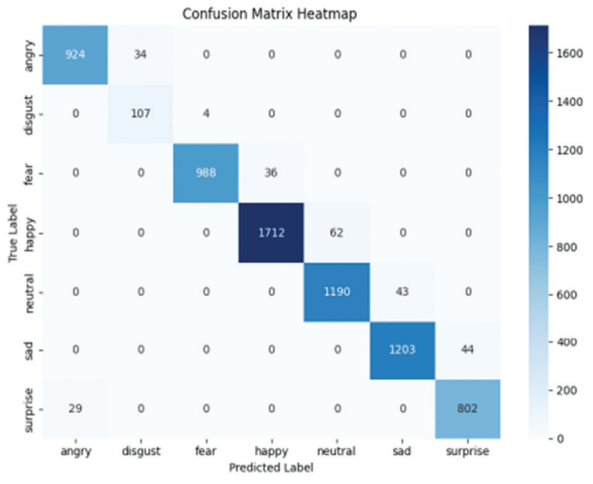

Fig. 5. Confusion matrix for the FER-CBAMNet model

Table 5 presents an ablation analysis that isolates the contribution of the Convolutional Block Attention Module (CBAM) by comparing FER CBAMNet against an identical lightweight CNN backbone in which CBAM blocks are removed, while keeping the same data split, preprocessing/augmentation pipeline, and training configuration. Incorporating CBAM yields a substantial improvement across all metrics: accuracy increases from 92.28% to 96.58% (+4.30 percentage points), and Macro F1 improves from 0.9217 to 0.9631 (+0.0414). The corresponding gains in macro precision (0.9220 → 0.9631) and macro recall (0.9215 → 0.9631) indicate that the improvement is not driven only by majority classes but reflects more balanced recognition across all seven expressions under the class imbalance of FER 2013. The increase in Weighted F1 (0.9235 → 0.9655) further supports the effectiveness of attention-based feature refinement. Overall, these results empirically validate CBAM as a key component of FER CBAMNet and explain a large portion of the performance gap over the CNN-only baseline.

The confusion matrix also shows the high classification performance with most of the predictions being clustered around the diagonal. The direct computation of the accuracy based on the confusion matrix is 96.49%, which is in line with the average accuracy of 96.58% obtained when the computation was done across several independent runs. The slight difference is attributed to the round-off and run-to-run error, since the confusion matrix represents one representative run, whereas the reported accuracy represents the average performance across 10 runs.

Table 5. Ablation study: CNN vs. CNN + CBAMNet on FER-2013

Metric	CNN + CBAM	CNN
Accuracy (%)	0.9658	0.9228
Macro Precision	0.9631	0.9220
Macro Recall	0.9631	0.9215
Macro F1	0.9631	0.9217
Weighted F1	0.9655	0.9235

The visualizations of the Grad-CAM using the proposed FER-CBAMNet model of various samples of the angry emotion are shown in Fig. 6. The activation heatmaps demonstrate the spatial areas that are the most influential in the decision-making process of the network. As noted, the model is always concerned with semantically relevant parts of the faces, specifically the mouth, eyebrows, and eyes, that are physiologically linked to the intensity of the emotion in tankerful expressions.

Properly categorized cases demonstrate a high concentration of activities in and around these regions, which means that the attention modules are successful in attracting the network to process discriminative cues and inhibit irrelevant background features. In contrast, misclassified samples (e.g., angry misinterpreted as sad or disgust) exhibit diffused or misplaced activations, suggesting overlapping local features between these closely related emotions. All in all, these GradCAM maps reveal that the attention-augmented architecture is not only more accurate at classification, but also at predicting the models in a better way that would allow the network to be more explainable, so that the predictions of the network are based on visual evidence that can be interpreted by a human.

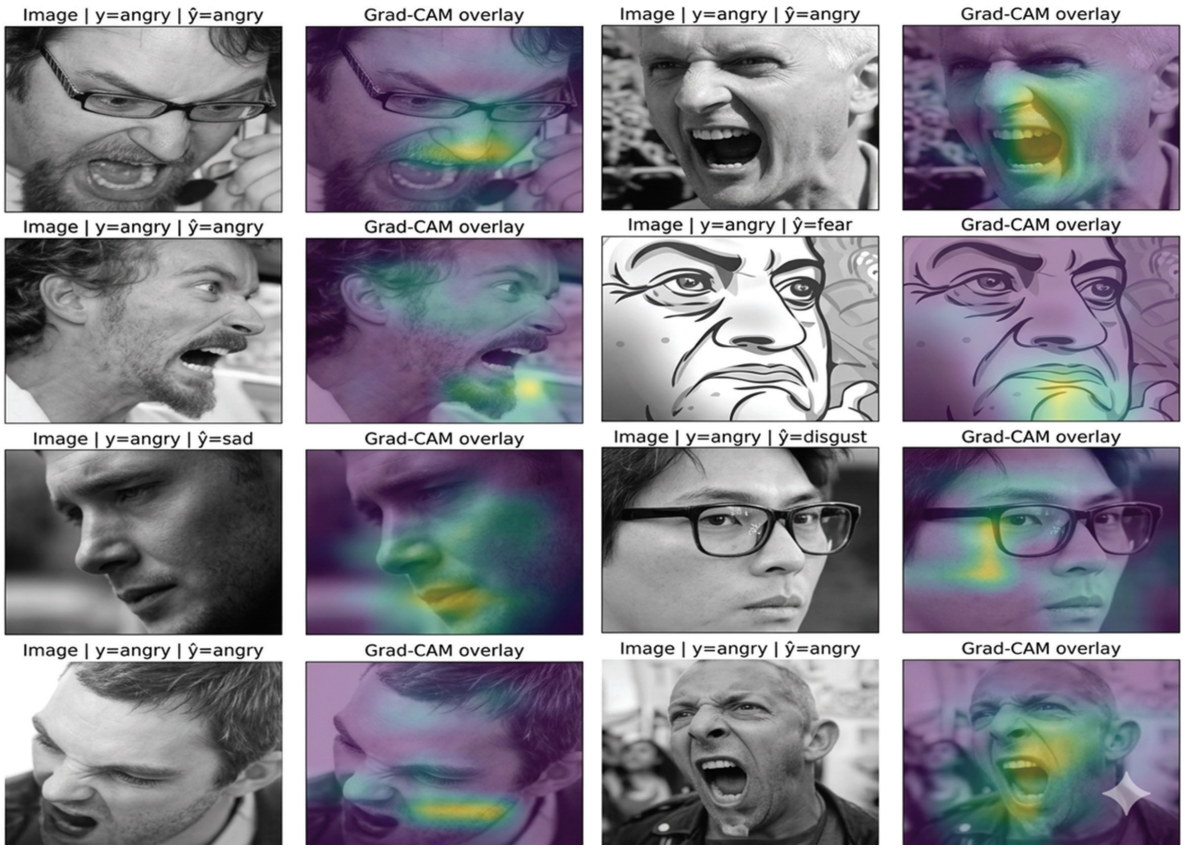

Fig. 6. Grad-CAM visualizations

Properly categorized cases demonstrate a high concentration of activities in and around these regions, which means that the attention modules are successful in attracting the network to process discriminative cues and inhibit irrelevant background features.

In contrast, misclassified samples (e.g., angry misinterpreted as sad or disgust) exhibit diffused or misplaced activations, suggesting overlapping local features between these closely related emotions. All in all, these GradCAM maps reveal that the attention-augmented architecture is not only more accurate at classification, but also at predicting the models in a better way that would allow the network to be more explainable, so that the predictions of the network are based on visual evidence that can be interpreted by a human.

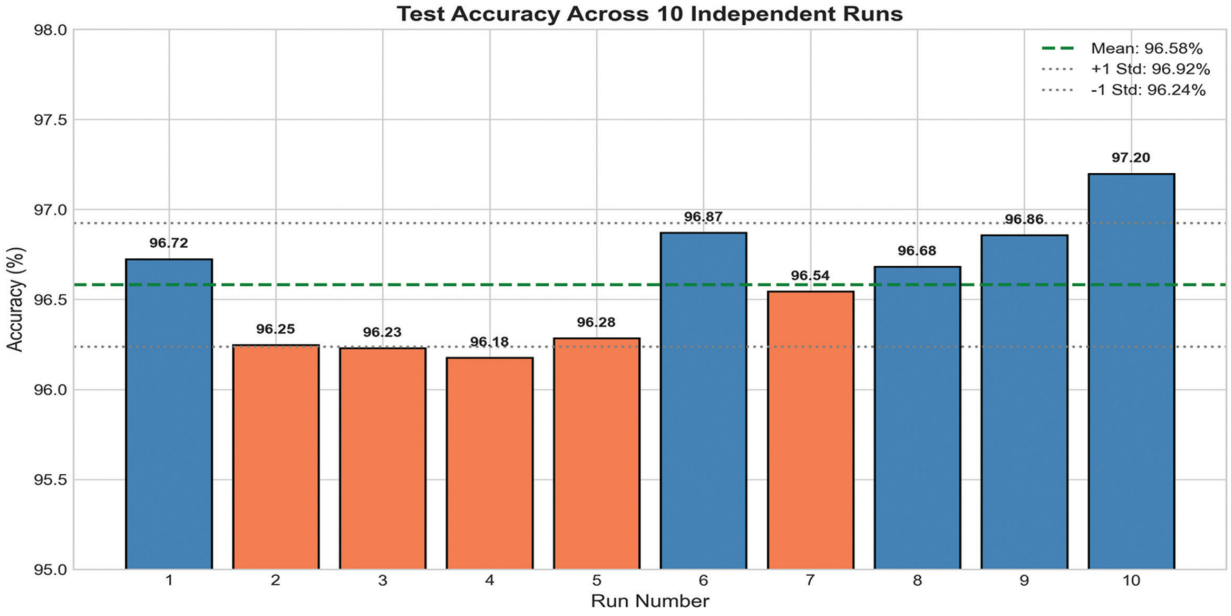

Fig. 7. Bar chart showing test accuracy for each run. Blue bars indicate runs above the mean, coral bars indicate runs below the mean

Fig. 7 shows the consistency of the proposed model in 10 independent runs, which also indicates the test accuracy. The average accuracy is 96.58% as shown by the green straight line, and the dotted line shows the one standard deviation range above and below the line. The bulk runs are near the mean, with some of the runs (in blue) outlying the mean, which are reflective of good and consistent performance. The low variability between minimum and maximum accuracy is verified by the low spread between the minimum and maximum accuracies of the model, and it proves the fact that the model does not suffer a significant decline in performance during a certain run.

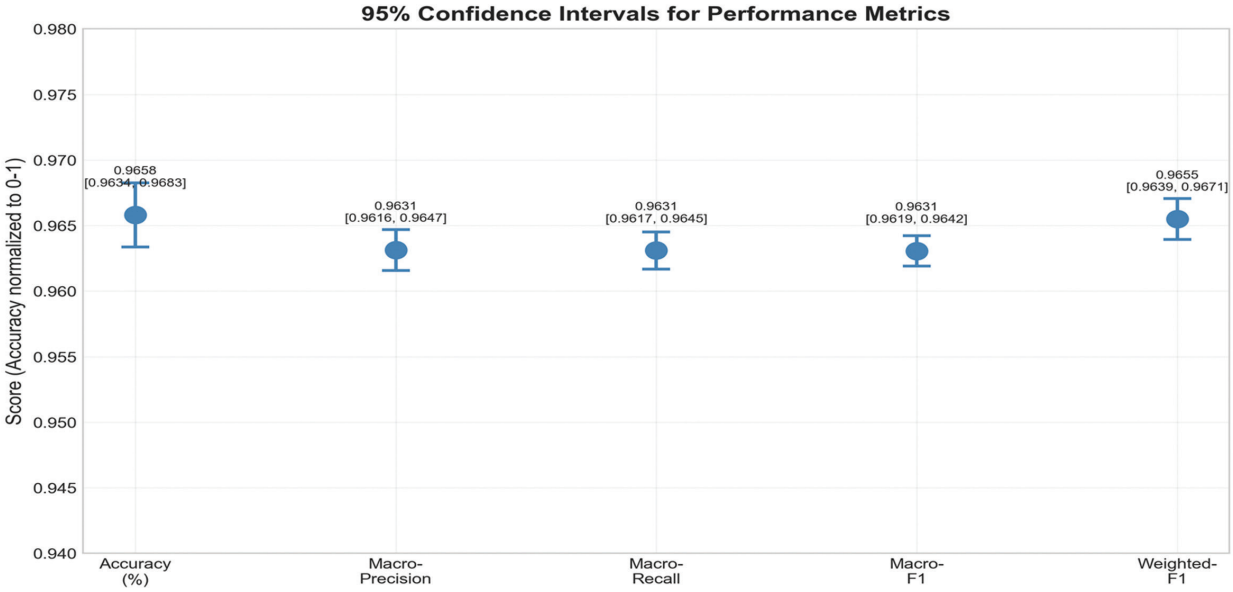

Fig. 8. 95% confidence intervals for all performance metrics

Fig. 8 presents the 95% confidence intervals for all evaluation metrics, including accuracy, macro-precision, macro-recall, macro-F1, and weighted-F1. The small confidence intervals imply that the model is not very variable and that the model performance is statistically reliable across the runs. The focus of all metrics is on high mean values (around 0.96-0.97), which proves the high and stable predictive power of the model. The close proximity and narrowness of the intervals also indicate that the model has balanced and stable performance with regard to various evaluation criteria, which supports the soundness of the suggested framework.

5. DISCUSSION

Table 6 places the proposed FER-CBAMNet in a larger and more representative context of FER-2013 baselines, including traditional CNN pipelines, lightweight deployment-focused systems, and more recent attention/transformer-based systems. Besides earlier CNN versions that report high accuracies on FER-2013 [36, 37], we also add popular reference backbones and more powerful modern comparisons, including VGGNet (single network) [41], a ResNet-18 based model with effective channel attention and contrastive losses (ECAN) [42], a Swin Transformer baseline [43], and a collection of transfer-learning models [44], along with a small CNN-KNN model that is built to be efficient [22]. In these representative families, the reported test accuracies are generally in the range of 71-75% with standard backbones and lightweight/ensemble versions, although more specialized architectures that use attention or graph/transformer blocks can be more performant (e.g., 82.47% with EfficientNetV2 + attention [39] and 90.40% with a multi-scale simplicial transformer + GAT [27]). Using FER-2013 under our modified split protocol, the accuracy of FER-CBAMNet is 96.58%, higher than the mentioned CNN, attention-based, and transformer-based models in Table 6.

Table 6. Comparison with recent FER methods on FER-2013

Ref	Model	Dataset	Accuracy
[17]	CNN + VGG	FER-2013	67.70%
[22]	CNN-KNN	FER-2013	75.26%
[33]	Multi-scale Simplicial Transformer + GAT	FER-2013	90.40%
[36]	CNN-based FER	FER-2013	95.0
[37]	CNN	FER-2013	93.7
[38]	Zoning + VGG-16 + FCNN	FER-2013	65.0
[39]	EfficientNetV2 + Attention	FER-2013	82.47%
[40]	Lightweight Convolutional Neural Networks + Attention-based Vision Mode	FER-2013	93.88 %
[41]	VGGNet (single network)	FER-2013	73.28 %
[42]	ECAN (ResNet-18 + losses contrastives)	FER-2013	73.73%
[43]	Swin Transformer	FER-2013	71.11%
[44]	Ensemble 4 transfer learning models	FER-2013	75.06%
Ours	FER-CBAMNet	FER-2013 (modified split)	96.58

This gain is in line with the main design hypothesis of the suggested framework: the incorporation of CBAM in a lightweight residual backbone improves feature selectivity by focusing on expression-relevant facial areas and inhibiting irrelevant background responses, which reinforces recognition in the conditions of noise and class imbalance of FER-2013. Simultaneously, it is indicated that cross-paper comparisons can still be affected by the difference in the evaluation protocol (e.g., public vs. private test split), preprocessing, and use of pretraining or ensembles; thus, the literature comparison is supplemented by the ablation analysis provided in Table 5, which isolates the contribution of CBAM under the same experimental setup.

6. CONCLUSION

This paper has shown that a succinct and tightly designed deep learning model can be used to obtain state-of-the-art results at facial expression recognition with attention mechanisms directly incorporated into the model. The proposed FER-CBAMNet uses a lightweight residual network with attention modules of CBAM to highlight the most informative parts of the face, including the eyes, eyebrows, and mouth, and suppress unwarranted background details. This focused feature refinement goes a long way in enhancing robustness against typical threats such as variations in illumination, pose variations, and partial occlusions. The comparative analysis of FER-CBAMNet on the FER-2013 dataset shows that the model can achieve high consistency in its results after a few runs, which demonstrates that it is a well-generalized model despite the imbalance of classes. Moreover, Grad-CAM visualizations have meaningful interpretability that proves that predictions of the model are guided by semantically meaningful facial regions and not random patterns. Overall, this paper has shown that the attention-based mechanism is a great approach in using lightweight convolutional architectures to build an accurate, efficient, and understandable system of facial expression recognition, which is why the specified approach is highly suitable in real-life conditions and scenarios with limited resources.

The research in the future will take three tangible directions. To measure the strength of generalization, we will first test it more strictly using cross-dataset transfer and domain-shift conditions to assess how well it can be applied to other datasets. Second, we will expand the framework to multimodal affect recognition by incorporating audio and/or physiological cues with attention-based fusion to more effectively deal with ambiguous expressions. Third, we will streamline the model to be deployed by using model compression methods and reporting latency/throughput on representative edge hardware. The goals of these extensions are to enhance the practical deployability as well as the methodological rigor of small, explainable FER systems to real-world human-centered applications.

DECLARATION OF THE USE OF AI TOOLS

Grammarly was used by the authors to edit the language. The authors reviewed and edited all AI-assisted content and are fully responsible for the integrity and accuracy of this work. No AI tool is mentioned as an author, and AI tools are not mentioned as primary sources.

REFERENCES

[1] R. A. Fayadh, M. K. Wali, A. M. Breesam, "Real-Time Practical Face Recognition System of Distorted Images Using PCA", Proceedings of the 4th International Conference on Current Research in Engineering and Science Applications, Baghdad, Iraq, 20-21 December 2022, pp. 290-295. doi: 10.1109/ICCRESA57091.2022.1035249

[2] A. M. Breesam, M. K. Wail, R. A. Fayadh, "Principle component analysis based face recognition", AIP Conference Proceedings, Vol. 2398, No. 1, 2022. doi: 10.1063/5.0093821

[3] H. J. Hassan, M. I. Shujaa, A. M. Breesam, M. K. Wali, "Review of face-based recognition systems", AIP Conference Proceedings, Vol. 3232, No. 1, 2024. doi: 10.1063/5.0236493

[4] M. S. M. Rahim, A. E. Rad, A. Rehman, A. Altameem, "Extreme facial expressions classification based on reality parameters", 3D Research, Vol. 5, 2014. doi: 10.1007/s13319-014-0022-7

[5] P. Khorrami, T. Paine, T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?", Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7-13 December 2015, pp. 19-27.

[6] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, "CBAM: Convolutional block attention module", Proceedings of the European Conference on Computer Vision, Munich, Germany, 8-14 September 2018, pp. 3-19.

[7] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, "Grad-CAM: Why did you say that?", arXiv1611.07450, 2016.

[8] I. J. Goodfellow *et al*. "Challenges in representation learning: A report on three machine learning contests", Proceedings of the International Conference on Neural Information Processing, Daegu, Korea, 3-7 November 2013, pp. 117-124. doi: 10.1007/978-3-642-42051-1_16

[9] T. Kopalidis, V. Solachidis, N. Vretos, P. Daras, "Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets", Information, Vol. 15, No. 3, 2024, p. 135. doi: 10.3390/info15030135

[10] S. Li, W. Deng, "Deep facial expression recognition: A survey", IEEE Transactions on Affective Computing, Vol. 13, No. 3, 2020, pp. 1195-1215. doi: 10.1109/TAFFC.2020.2981446

[11] S. Shaees, H. Naeem, M. Arslan, M. R. Naeem, S. H. Ali, H. Aldabbas, "Facial emotion recognition using transfer learning", Proceedings of the International Conference on Computing and Information Technology, Tabuk, Saudi Arabia, 9-10 September 2020, pp. 1-5. doi: 10.1109/ICCIT-144147971.2020.9213757

[12] A. Sultana, S. K. Dey, M. A. Rahman, "Facial emotion recognition based on deep transfer learning approach", Multimedia Tools and Applications, Vol. 82, No. 28, 2023, pp. 44175-44189. doi: 10.1007/s11042-023-15570-z

[13] K. Meethongjan, M. Dzulkifli, A. Rehman, A. Altameem, T. Saba, "An intelligent fused approach for face recognition", Journal of Intelligent Systems, Vol. 22, No. 2, 2013, pp. 197-212. doi: 10.1515/jisys-2013-0010

[14] U. Sharif, Z. Mehmood, T. Mahmood, M. A. Javid, A. Rehman, T. Saba, "Scene analysis and search using local features and support vector machine for effective content-based image retrieval", Artificial Intelligence Review, Vol. 52, No. 2, 2019, pp. 901-925. doi: 10.1007/s10462-018-9636-0

[15] G. Meena, K. K. Mohbey, S. Kumar, R. K. Chawda, S. V Gaikwad, "Image-based sentiment analysis using InceptionV3 transfer learning approach", SN Computer Science, Vol. 4, No. 3, 2023, p. 242. doi: 10.1007/s42979-023-01695-3

[16] G. Meena, K. K. Mohbey, K. Lokesh, "FSTL-SA: Few-shot transfer learning for sentiment analysis from facial expressions", Multimedia Tools and Applications, Vol. 84, No. 21, 2025, pp. 24457-24485. doi: 10.1007/s11042-024-20518-y

[17] Y. Bordjiba, H. F. Merouani, "Facial expression recognition via a jointly-learned dual-branch network", International Journal of Electrical and Computer Engineering Systems, Vol. 13, No. 6, 2022, pp. 447-456. doi: 10.32985/ijeces.13.6.4

[18] R. K. Nukathati, U. B. Nagella, A. P. Kumar, "A deep learning framework with optimizations for facial expression and emotion recognition from videos", International Journal of Electrical and Computer Engineering Systems, Vol. 16, No. 3, 2025, pp. 217-229. doi: 10.32985/ijeces.16.3.3

[19] S. Minaee, M. Minaei, A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network", Sensors, Vol. 21, No. 9, 2021, p. 3046. doi: 10.3390/s21093046

[20] M. R. Koujan, "3D Face Modelling, Analysis and Synthesis", University of Exeter, United Kingdom, 2022, PhD thesis.

[21] Y. Zhang, C. Wang, W. Deng, "Relative uncertainty learning for facial expression recognition", Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 17616-17627.

[22] M. N. Ab Wahab, A. Nazir, A. T. Z. Ren, M. H. M. Noor, M. F. Akbar, A. S. A. Mohamed, "Efficientnet-lite and hybrid CNN-KNN implementation for facial expression recognition on raspberry pi", IEEE Access, Vol. 9, 2021, pp. 134065-134080. doi: 10.1109/ACCESS.2021.3113337

[23] H. Sikkandar, R. Thiyagarajan, "Deep learning based facial expression recognition using improved Cat Swarm Optimization", Journal of Ambient Intelligence and Humanized Computing, Vol. 12, No. 2, 2021, pp. 3037-3053. doi: 10.1007/s12652-020-02463-4

[24] C. Pabba, P. Kumar, "An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition", Expert Systems, Vol. 39, No. 1, 2022, p. e12839. doi: 10.1111/exsy.12839.

[25] A. R. Khan, "Facial emotion recognition using conventional machine learning and deep learning methods: current achievements, analysis and remaining challenges", Information, Vol. 13, No. 6, 2022, p. 268. doi: 10.3390/info13060268

[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, "Learning deep features for discriminative localization", Proceedings of the IEEE Conference

on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27-30 June 2016, pp. 2921-2929.

[27] S. B. Punuri, S. K. Kuanar, T. K. Mishra, V. V. R. M. Rao, S. S. Reddy, "Decoding human facial emotions: a ranking approach using explainable AI", IEEE Access, Vol. 12, 2024, pp. 186229-186245. doi: 10.1109/ACCESS.2024.3474012

[28] M. Mohana, P. Subashini, G. Ghinea, "XAI-DSCSA: explainable-AI-based deep semi-supervised convolutional sparse autoencoder for facial expression recognition", Signal, Image Video Process., Vol. 19, No. 5, 2025, p. 394. doi: 10.1007/s11760-025-03984-1

[29] L. Ji, S. Wu, X. Gu, "A facial expression recognition algorithm incorporating SVM and explainable residual neural network", Signal, Image and Video Processing, Vol. 17, No. 8, 2023, pp. 4245-4254. doi: 10.1007/s11760-023-02657-1

[30] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN", Electronics, Vol. 10, No. 9, 2021, p. 1036. doi: 10.3390/electronics10091036

[31] H.-W. Ng, V. D. Nguyen, V. Vonikakis, S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning", Proceedings of the ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9-13 November 2015, pp. 443-449. doi: 10.1145/2818346.2830593

[32] C. Bisogni, A. Castiglione, S. Hossain, F. Narducci, S. Umer, "Impact of deep learning approaches on facial expression recognition in healthcare industries", IEEE Transactions on Industrial Informatics, Vol. 18, No. 8, pp. 5619-5627, 2022. doi: 10.1109/TII.2022.3141400

[33] S. N. Yousafzai *et al*. "A multi-scale simplicial transformer with graph attention for facial emotion recognition", Ain Shams Engineering Journal, Vol. 16, No. 10, p. 103584, 2025. doi: 10.1016/j.asej.2025.103584

[34] S. R. Adnan, A. M. Breesam, A. H. Abdulaal, "Enhancing AD Classification with Deep Learning: A Study of Transfer Learning and Snake Optimization on MRI Data.", Mathematical Modelling of Engineering Problems, Vol. 12, No. 9, 2025. doi: 10.18280/mmep.120925

[35] E. Najjar, N. A. Alkhaykanee, A. M. Breesam, "Classification of Arabic Documents with Five Classifier Models Using Machine Learning", JOIV : International Journal on Informatics Visualization, Vol. 9, No. 1, 2025, pp. 365-370. http://dx.doi.org/10.62527/joiv.9.1.2539

[36] G. Meena, K. K. Mohbey, A. Indian, M. Z. Khan, S. Kumar, "Identifying emotions from facial expressions using a deep convolutional neural network-based approach", Multimedia Tools and Applications, Vol. 83, No. 6, 2024, pp. 15711-15732. doi: 10.1007/s11042-023-16174-3

[37] Ö. Ezerceli, M. T. Eskil, "Convolutional neural network (CNN) algorithm based facial emotion recognition (FER) system for FER-2013 dataset", Proceedings of the International Conference on Electrical, Computer, Communications and Mechatronics Engineering, Maldives, 16-18 November 2022, pp. 1-6. 10.1109/ICECCME55909.2022.9988371

[38] T. Shahzad, K. Iqbal, M. A. Khan, N. Iqbal, "Role of zoning in facial expression using deep learning", IEEE Access, Vol. 11, 2023, pp. 16493-16508. doi: 10.1109/ACCESS.2023.3243850

[39] M. E. Konuk, E. Kılıç, "Efficientfer: Efficientnetv2 based deep learning approach for facial expression recognition", Proceedings of the 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Ankara, Turkiye, 23-24 May 2025, pp. 1-7. doi: 10.1109/ICHORA65333.2025.11017006

[40] G. Mishra, S. Bajpai, D. Saini, R. Jain, D. K. Jain, V. Štruc, "EmoVisioNet: A hybrid network unifying lightweight CNN and attention-based vision model for facial emotion detection", Neurocomputing, Vol. 665, p. 132224, 2025. doi: 10.1016/j.neucom.2025.132224

[41] Y. Khaireddin, Z. Chen, "Facial emotion recognition: State of the art performance on FER2013", arXiv2105.03588, 2021.

[42] J. Zhu, S. Liu, S. Yu, Y. Song, "An extra-contrast affinity network for facial expression recognition in the wild", Electronics, Vol. 11, No. 15, p. 2288, 2022. doi: 10.3390/electronics11152288

[43] M. Bie, H. Xu, Y. Gao, K. Song, X. Che, "Swin-FER: Swin transformer for facial expression recognition", Applied Sciences, Vol. 14, No. 14, 2024, p. 6125. doi: 10.3390/app14146125

[44] C. Białek, A. Matiolański, M. Grega, "An efficient approach to face emotion recognition with convolutional neural networks", Electronics, Vol. 12, No. 12, 2023, p. 2707. doi: 10.3390/electronics12122707