Emotion Classification using Facial Expression

***Abstract*—The task of facial emotion recognition has gained a lot of interest in recent times with the increased use of devices with high computing capacity in day-to-day lives. With cameras getting embedded in almost all electric devices in use today, It becomes easy to gain access to image data. This makes the use of image data helpful especially in cases of sentiment analysis and monitoring. Understanding facial emotions helps in improving the performance of the tasks being performed, especially in the fields of marketing and teaching.**

There were many papers in the past which have introduced the models that can make this task of identifying emotions possible. While some use the traditional image processing techniques to solve the need, few use neural network architectures. This paper proposes an improved version of one such transfer learning technique.

The paper proposes an addition of attention modules to the existing ResNet18 architecture. The attention modules help the model to focus on the specific parts of input. Focusing on specific parts helps prioritize the important features thereby producing better accuracy results.

The improved ResNet18 model with the help of CBAM and SE attention modules achieves FER accuracies of 98.8 percent and 69.23 percent on CK+ and FER2013 test sets, respectively. The obtained results show that the suggested FER system based on the improved model outperforms the Deep TL techniques in terms of both emotion detection accuracy and evaluation metrics.

***Index Terms*—Facial Emotional analysis, machine learning, deep learning, convolutional neural network, deep belief network, artificial intelligence, attention modules, transfer learning**

I. Introduction

The field of computer vision has undergone remarkable progress in recent years, driven by the development of sophisticated applications that interpret and extract meaningful insights from images and video data. The origins of automated facial expression analysis trace back to the late 1970s, with pioneering research by Suwa and Motoi, who introduced techniques to track the motion of facial points across image sequences. This foundational work laid the groundwork for modern facial emotion recognition (FER) systems, which utilize advanced learning approaches such as deep learning to classify emotional states based on facial expressions.

The human face, composed of approximately 40 muscles, can produce thousands of unique expressions, offering valuable insights into an individual's cognitive and emotional state. FER systems aim to decode this nonverbal communication by analyzing facial cues, enabling applications in diverse domains such as social robotics, healthcare, marketing, and advanced human-computer interactions. These systems commonly classify emotions into seven core categories: Surprise, Fear, Anger, Disgust, Sadness, Happy, and Neutrality. Each emotion manifests through distinct facial cues, making them well-suited for automated analysis.

A. Motivation and Research Problem

While FER has achieved significant advancements through deep learning, several challenges remain unresolved: Traditional FER techniques, relying on feature engineering methods like Local Binary Patterns (LBP) and Principal Component Analysis (PCA), lack robustness in complex, real-world scenarios. Deep learning architectures such as ResNet-18 and EfficientNet offer significant promise, but existing implementations often prioritize either accuracy or computational efficiency, rarely achieving both.The presence of variability in datasets, including occlusion, lighting, and pose differences, further complicates accurate emotion recognition. Thus, there is a need for a unified approach that combines the strengths of state-of-the-art architectures while addressing these limitations

B. Research Question

This research seeks to address the following question: *How can ResNet-18 architecture be optimized for facial emotion recognition to achieve a balance between high accuracy and computational efficiency across diverse datasets?*

C. Contributions of the Proposed Study

This study makes the following contributions:

- Comparative Analysis: A detailed evaluation of ResNet-18 and EfficientNet architectures for FER tasks, highlighting their strengths and weaknesses.
- Model Enhancements: Integration of attention mechanisms into ResNet-18 to enhance feature extraction, improved scalability and real-time application.
- Dataset Evaluation: Comprehensive testing on benchmark datasets, including FER2013, CK+, to ensure the proposed models generalize across varying conditions.

D. Beyond State-of-the-Art

The proposed research advances beyond state-of-the-art by:

- Accuracy and Efficiency Synergy: Unlike existing methods that focus on either high accuracy or low computational cost, this research emphasizes achieving both.
- Attention Mechanisms: Incorporating attention modules into ResNet-18 enhances its ability to focus on critical facial regions, such as the eyes and mouth, which are essential for emotion detection.

- Cross-Dataset Generalization: By evaluating the models on diverse datasets (FER2013 for real-world settings, CK+ for controlled environments), the research ensures robust performance across a range of conditions.

E. Proposed Approach for Enhanced FER

The study outlines a systematic methodology for improving FER:

- Model Evaluation: Testing various pre-trained architectures using transfer learning to identify the optimal framework for FER.
- Architecture Refinement: Enhancing ResNet-18 with attention mechanisms for FER-specific tasks.
- Comprehensive Validation: Conducting extensive experiments on benchmark datasets, using metrics such as accuracy, precision, recall, and F1-score to evaluate performance

The contributions and innovations presented in this study pave the way for developing an accurate, efficient, and generalizable FER system suitable for diverse real-world applications.

The human face, comprising approximately 40 muscles, is capable of producing thousands of expressions, each conveying unique insights into an individual's cognitive and emotional state. FER systems aim to decode this nonverbal communication by analyzing facial cues. These systems find applications across various domains, including social robotics, healthcare, marketing, and advanced human-computer interactions.

F. Key Emotions and Their Characteristics

FER commonly classifies emotions into seven core categories: Surprise, Fear, Anger, Disgust, Sadness, Joy, and Neutrality. Each of these emotions manifests through distinct facial cues:

- Surprise: Characterized by raised eyebrows and potential forehead wrinkles, often a precursor to either joy or sadness.
- Fear: Marked by wide-open eyes, dilated pupils, and drawn-in eyebrows, reflecting a reaction to perceived threats.
- Anger: Associated with muscle tension, flushed skin, and increased blood pressure, conveying strong emotional arousal.
- Disgust: Highlighted by a wrinkled nose and raised upper lip, often in response to sensory stimuli like taste or smell.
- Sadness: Identified by lowered eyebrows and drooping lips, often tied to loss or pain.
- Happiness: Evident in upwardly curved lips and visible teeth, signaling positive emotions.
- Neutrality: A baseline state devoid of pronounced emotional expression.

These classifications underpin FER systems, enabling automated identification of emotional states based on observed facial patterns.

II. Literature Review

FER has greatly improved with the generation of deep learning models, providing even greater accuracy and reliability in emotion recognition systems. Among these developments is the utilization of EfficientCNN, encompassing EfficientNet for emotion classification and pre-trained on ImageNet. The same model has achieved greater performance on benchmarks such as FER-2013, RAF-DB, and CK+, while, at the same time, showing the good generalization property of the model to actual applications like emotional health tracking and surveillance systems [1]. The use of soft labels in FER has also been investigated by use of soft labels, derived from pre-trained CNN outputs, and integrated within ensemble learning. This solution involved a wider classifier family with a subsequent increase in performance on FER-2013, SFEW, and RAF databases, suggesting an extremely important role of soft labels in improving the accuracy of FER systems [2].

FERConvNet represents the biggest innovation in low-resolution FER. It allows for the combination of CNN and the use of denoising techniques whereby Gaussian, Bilateral, and Non-Local Means filters work in synergy to increase recognition accuracy on low-resolution facial expressions. It outperforms architectures like VGG16 and EfficientNetB7, especially while subjected to the FER-2013 dataset. This was evidenced by the setup of a customized dataset-the LRFE-to carry on experiments on its development, showing the challenges introduced by low-resolution images in particular applications like surveillance, along with how denoising methods ameliorate these problems [3]. Another significant direction within this area relates to the development of facial landmark detection methods to extract features from the face so that emotion classification can be accurately performed under different conditions [4].

FER might foster facial recognition technologies in completely a different manner in support of security systems throughout the process. Through the social insight provided by these measures, the researchers were of the opinion that FER could very well improve the efficiency of security systems with regard to scanning public places for potential threats through emotive and behavioral analysis [5]. Furthermore, cross-domain emotion recognition has motivated the use of techniques based on domain adaptation and combining CNNs that enhanced FER models transferability across varied datasets and environmental conditions. It has thereby led to greater versatility in FER model application and somewhat greater usefulness in more scenarios [6].

Another promising trend is multimodal emotion recognition, which considers facial expressions as well as vocal inputs together. By incorporating voice analysis besides facial recognition, these systems improved emotion classification especially in noisy environments and, thus, enhanced the accuracy of FER systems [7]. Central to interactive applications has been the pursuit of real-time emotion detection systems. Optimization of CNNs towards high accuracy and yet computationally efficient enables FER to work in dynamic, real-time environ-

ments [8]. Yet another innovative approach combines CNNs with LSTM in the temporal detection of FER, with this hybrid architecture being nimble to capture those time-varying forms of changes on facial expressions. This holds very significantly for applications like video conferencing and virtual avatars. Because of its supposed learning curve for the dependencies in the video sequences, it invented itself furthest away from conventional CNN models while improving evolving abilities for emotion detection through facial expressions [9].

Haar cascades, in these days with the use of neural networks for emotion classification, are good at solving problems like changes in lighting and posing. Hence, it can be seen that FER could enhance the interactive learning scenarios [10].

This concept has put forward few-shot learning for the case when expression recognition must start from a few examples per class with almost no labeled data. This approach handles the challenge of little labeled data, particularly for rare emotional expressions or underrepresented demographic classes [11]. 3D face models have also been propounded to boost FER performance against poor conditions like head pose variations and occlusions. This way, these models capture more details in the features involved in emotion classification through three-dimensional reconstruction of the face, and thus could be observed as more accurate in classifying emotion than conventional two-dimensional systems [12].

Attention mechanisms working with CNN have been very conducive to FER, in recognition of the fact that this is how the model pays attention and focus on the most yielding areas of the face, mostly like the eyes and the mouth, which are of critical relevance to interpretation through emotions. This pursuer enhances classification accuracy as well as interpretability for the model, in that such a model indicates the areas of the face that give maximum reward to the recognition of emotions [13]. The next school of thought is focused on the credence to create a FER system, so to say, that works in an interactive environment. The very reasoning with dynamic communities could involve the retaliation of shades, occlusions, and poses-well, these have stimulated the interest of many researchers who have made amalgamation models in order to create the frontiers between deep learning and the traditional computer vision techniques for extreme FER models to euro towards such complex scenescapes [14]. Eventually, the perspective of FER regarding applications in mental health looks very convincing. Continuous recording of emotions aids one in monitoring and following up emotional states over time, thus affording early detection and management of conditions like depression and anxiety. Hence it can be safely concluded that FER systems will have to put custom healthcare solutions in place [15].

III. Experiments

A. Datasets

Creating a robust system for deep learning-based facial emotion recognition (FER) necessitates a substantial volume of labeled training data. These datasets must encompass diverse variations in facial features and environmental factors such as pose, lighting, and occlusions to ensure the model's adaptability to real-world scenarios. This section highlights two widely used datasets—CK+ and FER2013—employed during the training and evaluation phases of the proposed FER system.

1) CK+ Dataset: The Extended Cohn-Kanade (CK+) dataset is a widely adopted, laboratory-controlled benchmark in FER research. It extends the original CK dataset by addressing the limitations of incomplete and imprecise emotion labeling. CK+ contains image sequences that transition from a neutral state to peak emotional expressions, making it ideal for analyzing subtle facial emotion dynamics. This dataset includes approximately 3,150 annotated images, with emotions distributed across seven categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. The improved ResNet18 model achieved a remarkable recognition accuracy of 98 percent on this dataset, demonstrating its ability to classify facial emotions with high precision and reliability.

2) FER2013 Dataset: FER2013 is a large-scale dataset collected for emotion recognition tasks under real-world, unconstrained conditions. Released as part of the ICML 2013 competition on representation learning, it consists of 48×48 grayscale images of faces, each pre-aligned to occupy a consistent position within the frame. The dataset comprises 28,709 training samples, 3,589 validation samples, and 3,589 test samples, each categorized into seven emotion classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. The improved ResNet18 model, when trained on FER2013, achieved an accuracy of 69.7 percent, reflecting its capability to handle diverse and complex real-world scenario.

Both datasets were instrumental in fine-tuning and validating the proposed FER system. The ResNet18 model with attention mechanisms, demonstrated superior performance across various evaluation metrics, including precision, recall, and F1-score.

IV. Proposed Methodology

The methodology proposed in this paper is an extension of the existing resnet18 model, a popular transfer learning model used for multiclass classifications. The changes include the addition of attention blocks in layer4 of the model. As can be seen in the Fig. 1, the Convolution Block Attention Module (CBAM) and Squeeze and Excitation (SE) blocks were added in layer4 to add attention to focus on important features to enhance the accuracy. The adaptive average pooling was done before feeding the features to the fully connected layer.

The model proposed has an initial convolution 2d block with kernel size of 7 x 7, followed by batch normalization layer and an in-place relu activation function followed by max pool layer of kernel size 3 and stride of 2. Out of the 4 layers of the resnet18 only the last layer has a change which is the addition of CBAM and SE blocks. Each layer has a pair of blocks where each block has a set of convolutions, batch normalization, relu activation followed by another set of convolutions and batch normalization.

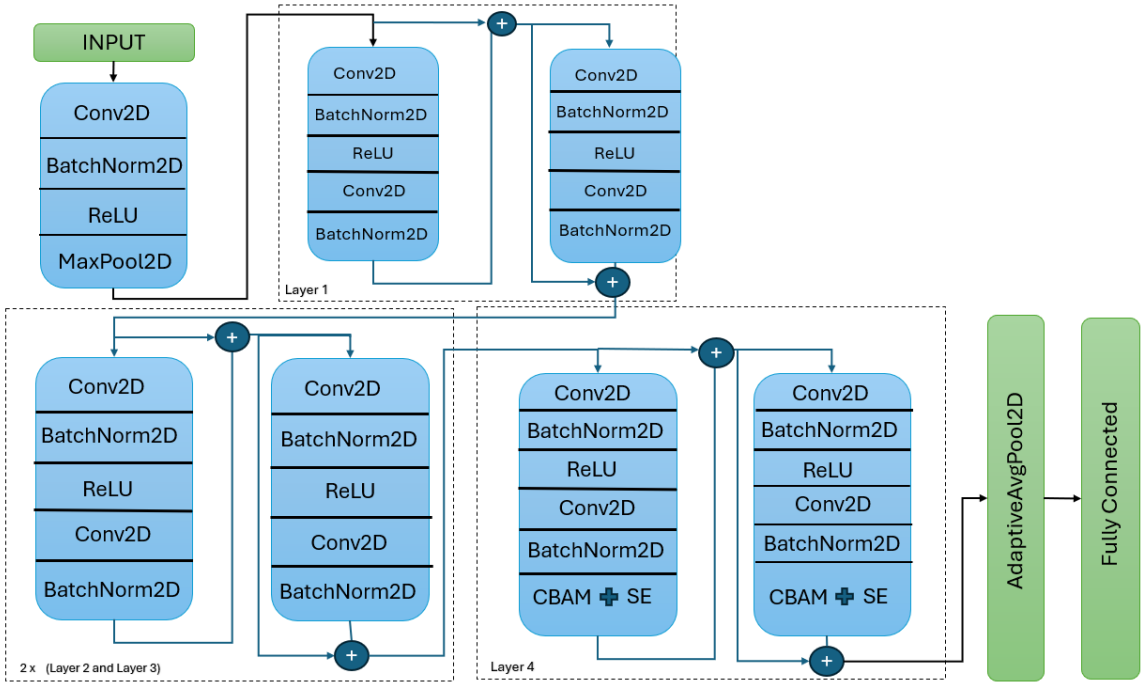

Fig. 1. Proposed ResNet18 model with CBAM and SE attention blocks

The usage of an ensemble model where we ResNet18 and EfficientNet models were ensembled to get better accuracy results was also done. As the EfficientNet contains SE block internally only CBAM will be used for it. For clear understanding of each component involved, each component is discussed below.

A. *ResNet18*

ResNet-18 is a member of the Residual Networks family of deep-learning architectures that have greatly changed approaches to image recognition tasks. Kaiming He and colleagues introduced ResNet-18 in 2015 to address a common issue in very deep neural networks: the vanishing gradient problem. Residual connections that allow the network to learn an identity mapping characterize the ResNet concept. This technique lets the gradient flow more easily through the network and ensures that it does not vanish in deeper layers.

ResNet-18 is, therefore, relatively shallowed with respect to deeper ResNet architectures like ResNet-50 and ResNet-101, although it performs extraordinarily well across various image-classification datasets. The architecture consists of several parts but begins with a 7x7 convolution with 64 filters with a stride of 2, followed by batch normalization and ReLU activation. This initial layer allows the first-level feature extraction from the input image.

Upon that, a 3x3 max-pooling layer is applied with a width of 2 to reduce the spatial dimension of the feature map. The heart of the ResNet-18 model comprises the two-layered residual blocks with 3x3 convolution layers with batch normalization and ReLU activation in between each convolution. These are assembled into four layers with two basic elementary blocks in each layer. Further into the network, the number of filters increases, allowing the model to progressively capture more complex features. The residual connection in each block gives the output of the previous block to the output of the present block, permitting the network to learn residual mappings during things, rather than trying to learn the entire mapping.

Layers are organized as follows: the first layer consists of two basic blocks equal to 64 filters, the second layer contains two basic blocks equal to 128 filters, the third layer consists of two basic blocks equal to 256 filters, and the fourth layer contains two basic blocks equal to 512 filters. An adaptive average pooling layer, which reduces the spatial dimensions of the feature maps, follows at the end of the network and a fully connected layer to give the final classification output.

B. *Squeeze and Excitation Block*

The Squeeze-and-Excitation Block is a novel yet compact attention architecture, introduced in the paper titled ”Squeeze-and-Excitation Networks,” authored by Jie Hu and others, in 2018. The purpose of this block lies in enhancing channel-wise feature representation, giving importance to, or recalibrating the importance of, various channels according to their feature map. The SE block is conceptualized to establish connection across channels such that one channel could emphasize a salient feature while suppressing non-valuable features.

The block operates in the following steps: it squeezes and then excites. First, during the squeezing step, global average pooling is done across the spatial dimension (height and width) of the feature map and put together into a 1D vector. This yields global feature information that portrays the whole

spatial extent of the feature map. The next step is excitation which is where most of the learning is done. The squeezed vector is passed through two fully connected layers interlinked with a ReLU activation function, giving rise to a set of channel-wise attention weights characterizing the important features within each channel. These attention weights are subsequently employed for the recalibration process, where the feature map is adjusted through channel-wise multiplication.

Such recalibration corresponds well in amplifying the important channels and inhibiting the trivial channels permitting heightened concentration of the model on the salient features. Computed very efficiently, the SE block can embed itself into already existent architectures such as ResNet-18 without introducing significant overhead, like achieving improved performance for the networks without an additional requirement on their resources.

C. Convolution Block Attention Module

An attention mechanism is quite as promising as CBAM has enhanced feature extraction in CNNs. Originally devised by Sanghyun Woo et al. in 2018, this module applies an attention mechanism to channel and spatial axes of a feature map. The fundamental idea behind CBAM is to focus on attending to the most relevant channels and spatial regimes in the feature map for better decision-making during forward passes.

CBAM creates channel attention and spatial attention in two basic steps: channel attention followed by spatial attention. Channel attention first aggregates the spatial information regarding the input feature map using both global average pooling and global max pooling. In the second phase, the pooled feature maps act as input to a shared fully connected layer to generate a set of weights for channel attention. Therefore, the use of the channel attention weights is a static mask that creates a spatial attention map to enhance important regions and reduce the relevance of the others through channel-wise multiplication for feature reconstruction.

While constructing the attention weighting, however, the interesting part of this attention module focuses on where the most important spatial regions reside to carry out the intended task. The module does first combine the features from global average pooling and global max pooling along the channel axis. This will generate a very compact feature set across all channels, which is finally passed through a 7x7 convolutional layer coupled with the sigmoid activation function to yield adequate attention weights. The feature map is then readjusted using element-wise multiplication based on the attention weights calculated above.

By utilizing channel-wise and spatial-wise attention, CBAM lends further weights to the model, which outputs more descriptive and enhanced features across the intercepting space of the image and in turn enhances network performance. The CBAM module is highly flexible and could be easily embedded into a different architecture such as ResNet-18 to push its ability to learn sophisticated features further.

D. Adding CBAM and SE attention to Models

This results in the application of attention mechanisms, such as inserting CBAM (Convolutional Block Attention Module) and SE (Squeeze-and-Excitation) blocks to ResNet18 and EfficientNet, thereby improving significantly due to strong weights being applied to more important features and downgrades for other less important features. As ResNet18 is a slightly simplified version of ResNet, it inherently benefits from its deep residual connections, while the addition of CBAM allows the model to use channel as well as spatial attention for feature maps to better describe. Within their SE block, the channel-wise feature responses are adaptive and allow the model to find important channels and suppress less informative channels. Compound scaling strategy with attention modules benefits EfficientNet effectively. Together, CBAM and SE allow EfficientNet to handle the complex features of input data more efficiently, giving better predictions with fewer parameters. These attention blocks contribute to a model's concentration on important patterns, especially during tricky tasks like facial emotion recognition or image classification. Therefore, it can be stated that models ResNet18 and EfficientNet with constructed CBAM and SE attention blocks are gaining higher accuracy as compared to the models constructed without attention, which shows that the role of attention is very significant in improving the feature representation while reducing a regular fit to help the model generalize across many datasets.

V. Experimental Section

This section details the experimental setup, datasets utilized, training methodology and results for evaluating the effectiveness of ResNet18 architecture enhanced with attention mechanisms, in facial emotion recognition (FER). The experiments were done using NVIDIA GPU P 100 accelerator and max disc space of 57 GB. The memory associated with CPU and GPU respectively are 29 and 16 GB.

A. Datasets and Preprocessing

Two benchmark dataset CK+ and FER2013 were employed in this study to ensure the robustness of the proposed models under varying conditions. These datasets represent controlled and unconstrained environments, respectively.

1) CK+ Dataset: The Extended Cohn-Kanade (CK+) dataset is a widely-used resource in FER research, offering sequences of facial expressions transitioning from a neutral state to a peak emotional expression. It contains 593 image sequences with annotations for seven emotion categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. *Key Characteristics :* Laboratory-controlled environment. High quality annotations , making it suitable for model benchmarking. *Preprocessing:* Images were aligned using facial landmarks to ensure uniformity. Normalized pixel values to a range of [0, 1]. Data augmentation techniques such as flipping, rotation, and cropping were applied to increase variability and reduce overfitting.

2) FER2013 Dataset: The FER2013 dataset is a large-scale, unconstrained dataset released during the ICML 2013 competition on representation learning. It consists of 35,887 grayscale images (48x48 pixels), divided into seven emotion categories.

Key Characteristics: Real-world scenarios with significant noise, occlusion, and pose variations. Diverse facial expressions captured under various lighting and environmental conditions.

Preprocessing : Standardized pixel intensity values. Applied data augmentation to address class imbalances and improve generalization. Partitioned into training (28,709 images), validation (3,589 images), and testing (3,589 images) sets.

B. Model Architectures and Enhancements

The study utilized ResNet18 architecture, enhanced with attention mechanisms to improve their feature extraction capabilities.

1) ResNet18 with Attention: ResNet18 is an 18-layer residual network designed to address vanishing gradient issues through skip connections. Its lightweight architecture and robust feature extraction make it suitable for FER tasks.ResNet18 with attention modules has given great accuracy to CK+ dataset. While FER2013 also found considerable improvement in metrics using attention.

Enhancements: Incorporated Squeeze-and-Excitation (SE) Blocks to dynamically adjust channel-wise feature importance, focusing on critical facial regions. Added spatial attention modules to prioritize facial landmarks such as the eyes and mouth, which are essential for emotion recognition. Fine-tuned on CK+ and FER2013 datasets using transfer learning from pre-trained ImageNet weights.

C. Results on Datasets

1) CK+ Dataset Results: ResNet18 with attention achieved 98.8% accuracy, demonstrating exceptional performance in this controlled environment. The integration of attention mechanisms significantly improved the model's ability to focus on subtle facial changes, particularly for emotions such as fear and surprise, which are often confused with other categories.

2) FER2013 Dataset Results: Using ResNet18 with CBAM and SE attentions alone, the model achieved 69.23% accuracy, highlighting the challenges posed by real-world noise and variability.

D. Training Methodology

The experimental design was set to effectively train the modified ResNet18 model enhanced by CBAM and SE blocks for good feature representation, with PyTorch's Data Loader taking care of data loading to provide efficiency and enable parallel processing in batches. The training operation involved batch size 32 shuffled to speed up the learning process but on test time was left unshuffled such that the evaluation results were like this only, and architecture was using custom fully connected for emotion classification stemmed from class labels in this dataset.

The training procedure consisted of 100 epochs with a CrossEntropyLoss function for multi-class classification with the Adam optimizer and a learning rate starting at 0.001. To ensure a smooth decay in learning rate, a CosineAnnealingLR scheduler is used. In addition, a 30% dropout in the last layers was used to ward off overfitting. The whole model along with data was transferred into GPU, and torch-no-grad was used during the test to handle the memory appropriately.

E. Accuracy and Loss Graphs

For FER2013, the accuracy achieved was using the improved ResNet18 with CBAM and SE attention modules with an accuracy of 69.23%. Similarly, for the CK+ dataset, the best accuracy and metrics were achieved using ResNet18 with CBAM and SE attention with an accuracy of 98.88%. The Confusion matrices can also be seen for both.

F. Ablation Study

The experiments held on ResNet18 against FER2013 and CK+ datasets helped us to analyze the improvements in accuracy with addition of blocks incrementally. The Table I shows the accuracies achieved for FER2013 and CK+ datasets with each version of ResNet18 experimented with different attention modules.

It can be observed that ResNet18 on FER2013, without any attention modules was able to produce an accuracy of 65.3%. With the addition of SE block alone the accuracy has jumped to 68.22%. With the addition of CBAM block alone to ResNet18 the accuracy has increased to 68.61%. When both the CBAM and SE were added, the accuracy of 69.23% was achieved.

And for CK+, The initial accuracy when only ResNet18 was used is 97.75%. The addition of attention blocks has increased the accuracy to 98.88%.

TABLE I
TEST ACCURACY OF MODELS ON FER2013 AND CK+ DATASETS

Model	***Test Accuracy on FER2013***	***Test Accuracy on CK+***
ResNet18	65.3%	97.75%
ResNet18 with CBAM	68.61%	98.88%
ResNet18 with SE	68.22%	98.88%
ResNet18 with CBAM, SE	69.23%	98.88%

G. Other Performance Metrics

The other performance metrics that we are interested in are precision, recall and f1-score. The results were calculated to the best models of each dataset we have found.

For FER2013, using ResNet18 with CBAM and SE attention blocks.Table II has the class values of metrics under study.

It can be found that sad has comparatively lesser precision whereas happy shows the highest precision. In fact, happy class has the highest precision, recall and f1-scores compared to all the other classes.

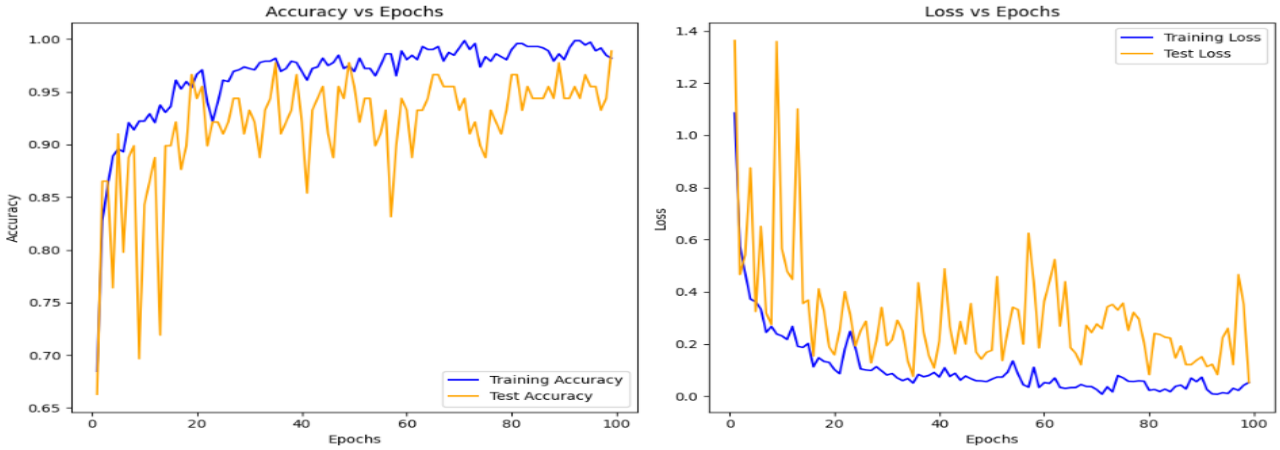

Fig. 2. Accuracy and Loss of ResNet18 with attention vs epochs for CK+

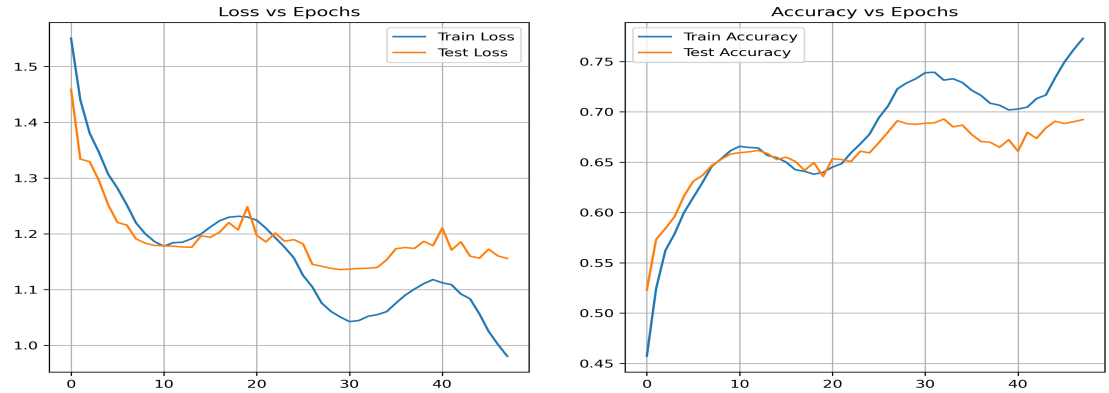

Fig. 3. Accuracy and Loss of ResNet18 with attention vs epochs for FER2013

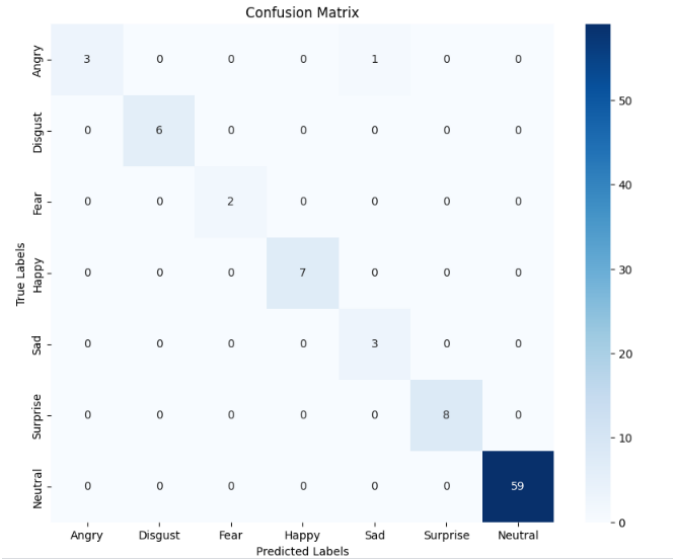

Fig. 4. Confusion matrix for ResNet18 with attention for CK+

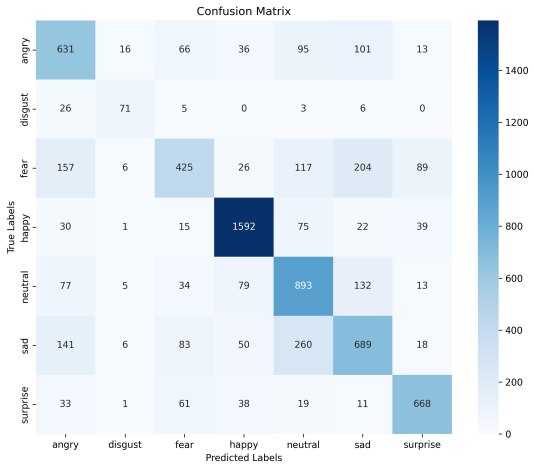

Fig. 5. Confusion matrix for ResNet18 model with attention for FER2013

TABLE II
PERFORMANCE METRICS ON FER2013 WITH RESNET18 WITH CBAM AND SE

Emotions	***Precision***	***Recall***	***F1-Score***
Angry	0.58	0.66	0.61
Disgust	0.67	0.64	0.65
Fear	0.62	0.41	0.50
Happy	0.87	0.90	0.86
Neutral	0.61	0.72	0.66
Sad	0.59	0.55	0.57
Surprise	0.79	0.80	0.80
Weighted Average	0.69	0.69	0.69

TABLE III
PERFORMANCE METRICS ON CK+ WITH RESNET18 WITH CBAM AND SE

Emotions	***Precision***	***Recall***	***F1-Score***
Angry	1.00	1.00	1.00
Disgust	1.00	1.00	1.00
Fear	1.00	1.00	1.00
Happy	1.00	1.00	1.00
Neutral	0.98	1.00	0.99
Sad	1.00	1.00	1.00
Surprise	1.00	0.87	0.93
Weighted Average	0.98	0.98	0.98

The weighted average achieved was 0.69 for all the metrics under study that is precision, recall and f1-score. For the CK+ dataset the model was able to produce 0.98 as the weighted average for all the metrics under study. The performance metrics of each class can be seen in III.

H. Comparative Study

The comparative study is shown in IV, comparing the accuracies achieved compared to other papers in the references. CK+ was improved to 98.88% while there is no significant improvement for FER2013.

I. Other Experiments

FERConvNetHDM is a facial expression recognition model that combines de-noising techniques like Gaussian, Bilateral, and Non-Local Means filtering to enhance image quality. These filters remove noise while preserving important details, improving the model's performance on low-resolution images. The method outperforms traditional models like VGG16 and VGG19, achieving higher accuracy on both FER2013 and custom datasets.

TABLE IV
COMPARATIVE STUDY WITH REFERENCE PAPERS

ReferenceNo.	***Test Accuracy on FER2013***	***Test Accuracy on CK+***
[1]	71% (on subset)	-
[2]	73%	-
[4]	-	97%
[14]	73.67%	-
ResNet18 with CBAM,SE attention	69.23%	98.88%

TABLE V
TEST ACCURACY OF MODELS ON FER2013 AND CK+ DATASETS

Model	***Test Accuracy on FER2013***	***Test Accuracy on CK+***
DenseNet	51%	75.82%
DenseNet with CBAM	52.8%	72.53%
DenseNet with CBAM, SE	53%	77.90%
FERConvNetHDM	66.91%	89.67%
FERConvNetHDM with CBAM	69.30%	88%
FERConvNetHDM with CBAM, SE	65.73%	91%
EfficientNet	58.45%	64.13%
EfficientNet with CBAM	61.0%	96.20%
EfficientNet with CBAM, SE	60.0%	94.57%

VI. CONCLUSION

This study presented an effective facial recognition system featuring attention-driven enhancements that could deal with real-life and controlled environments while employing advanced deep-learning architectures. Use of attention mechanism has pushed the models to state-of-the-art performance on benchmark datasets.

A. Key Findings

ResNet18 with attention mechanisms achieved an accuracy rate of 98.88% on the CK+ dataset, showing its good performance within controlled environments. It also achieved reasonably promising performance with an accuracy rate of 69.23% on FER2013.

B. Ablation Studies

Attention increasingly added mechanisms, increasing classifications for both datasets; this shows that attention mechanisms are critical for drawing close the attention of the classifier toward essential facial features in detecting FER. The Weighted Precision, Recall, and F1-scores are highly indicative of a strong classification capability with CK+ dataset doing particularly well in this regard, owing to the controlled conditions.

C. Drawbacks

The model proposed only uses the facial expressions to identify the emotion of an individual, but in reality, many more features might be useful in identifying the emotion. The features can be hand gestures, walking style and many more

D. Future Directions

Further research could explore multimodal FER, combining audio and visual data for improved emotion recognition. Investigation into domain adaptation techniques could enhance the generalization of FER systems across unseen datasets. Real-time deployment and evaluation in dynamic environments could validate the practical applicability of the proposed methods. The task of facial emotion recognition can be improved further using the video data where the temporal pattern is

observed. The results of analyzing temporal patterns using RNNs might give results that can be more interesting

ACKNOWLEDGMENT

We express our sincere gratitude to Prof. Ilaiah Kavati, Assistant Professor at the Department of Computer Science and Engineering, National Institute of Technology Warangal, for his insightful guidance and encouragement throughout the development of this paper.

REFERENCES

[1] "EfficientNet for Human FER using Transfer Learning," ICTACT Journal.
[2] "Facial expression recognition boosted by soft label with a diverse ensemble," ScienceDirect.
[3] "Facial expression recognition for partial occluded and low-resolution images," Signal, Image, and Video Processing, vol. 17, pp. 1234-1247, 2023.
[4] "DTL-I-ResNet18: facial emotion recognition based on deep transfer learning and improved ResNet18," Springer.
[5] "A Brief Review of Facial Emotion Recognition Based on Visual Information," MDPI Sensors
[6] "The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi," IEEE Xplore.
[7] "Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction.," IEEE Xplore.
[8] "Facial emotion recognition using convolutional neural networks (FERC)," Springer.
[9] "Facial Emotion Recognition by Ensemble-DenseNet Networksh," IEEE Xplore.
[10] "Facial emotion recognition using deep learning: review and insights," ScienceDirect.
[11] "The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi," IEEE Xplore.
[12] "A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing," IEEE Xplore.
[13] "A real time face emotion classification and recognition using deep learning model," IOP Conference Series: Materials Science and Engineering
[14] "Usage of ResNet18 with CBAM Attention Mechanisms in Facial Emotion Recognition," IEEE Xplore.
[15] "Facial Emotion Recognition Using Deep Convolutional Neural Network," IEEE Xplore.
[16] CK+ Dataset, "Extended Cohn-Kanade Dataset," Kaggle.