Facial Expression Recognition using Robust attention-based CNN

***Abstract*—Facial Expression Recognition (FER), a crucial aspect of social interaction and social communication, is the process of understanding human emotions using facial expressions. Recognizing emotions based on still facial images is a complex task that presents challenges such as understanding similarities between expressed emotions and dealing with variations in pose and illumination. This study proposes a novel approach to FER that aims to automatically analyze sentiment information and recognize emotions using visual data. To address FER challenges, the proposed approach incorporates a residual block with an attention mechanism followed by the MLP classifier with activation functions into the model. The approach uses VGG19 as a backbone for feature extraction and residual block with an attention mechanism to remember the importance of extracted features and utilizes a multi-layer perceptron as a classifier. The proposed architecture is evaluated using well-known datasets of CK+, RAF-DB, and FER2013. The experimental results indicate that the approach achieves 99.13%, 88.89%, and 89.06% classification accuracy on CK+, RAF-DB, and FER2013 datasets, respectively. The results show that the proposed approach out- performs state-of-the-art on CK+ and RAF-DB and performs comparably to the state-of-the-art on FER2013 datasets, attenuating the feasibility of the proposed approach in recognition of facial expressions.**

***Index Terms*—Facial Expression Recognition (FER), VGG19, Attention mechanism, Feature extraction.**

I. Introduction

Facial Expression Recognition (FER) is an increasingly popular research topic in computer vision due to its wide range of applications, including psychology, medicine, security, and digital entertainment [1], [2]. FER complexity is associated with incorporation of a collection of cognitive and sensory processes, involving visual perception, memory, attention, and emotional processing. Facial Expression Recognition task involves two primary steps of feature extraction from facial imagery and mapping of the extracted features to human emotions. The primary objective of FER is to analyze the relationship between emotions and facial expressions such as happiness, sadness, and disgust [3], [4].

Conventionally, FER studies aimed to extract complete features, including semantic and localization features, from facial images [5]. FER remains challenging despite these efforts due to complexities associated with identifying features from images with varying illumination conditions [6], the advert similarities between extracted features from different facial expressions [3], [5] and dealing with imbalanced class distribution in existing facial emotion datasets [7].

Uneven distribution of emotion classes in FER datasets is an additional complexity associated with FER. Imbalance distributions results in biased learning in favor of the classes with higher number of samples, resulting in increased overall accuracy of the prediction model which mostly represent the majority class, albeit the true classification performance tends to be lower [7], [8]. Majority of the publicly available datasets, e.g., extended FER (FER2013) and extended Cohn-Kanade dataset (CK+) [9], suffer from uneven class distribution. Pseudo-sampling, sub-sampling, and image augmentation are of common mechanisms used to increase the number of training samples in favor of less-populated classes which tend to increase the performance of the pre-trained models [10].

The existing advert similarities between features representing various emotions increases the difficulty of discriminating facial expressions [3]. To address this problem, more sophisticated features are identified through increasing the use of convolutional neural networks (CNNs) with deeper architectures or creating new convolutional architecture, i.e. ResNet [11], to extract deeper and better-discriminating features. Increasing the number of layers leads to improved feature extraction in many similar cases with caveat associated to the model training process caused by increasing the model's depths which increases the chance of getting over-fitted [12] and an the chance of vanishing gradient [13] while training the model.

This study proposes a new architecture to improve FER classification performance in multiple datasets. The proposed model prevents longer training time by avoiding sample augmentation. The study aims to implement and design a recognition model that can accurately and automatically recognize multiple expressions in various images. Generally, the process of facial expression recognition consists of the following steps: i) pre-processing of the facial expression data; ii) feature extraction of facial expressions; and iii) classification of facial expressions. A quasi-segmentation structure is used to increase the depth of the model to extract more sophisticated features. To avoid the problem of over-fitting and vanishing gradient descent, "skip connection" is implemented with attention layers [14]. This study utilizes FER, CK+, and RAF-DB [15] datasets to evaluate the proposed model. The model benefits from combining the base feature extraction model and segmentation model for feature extraction techniques. A set of dense layers is evaluated to see the effect of using different combinations of layers with various activation functions on the FER.

A. Study Contribution

The main contributions of the proposed model are listed as follows:

1) **Novel Approach:** The proposed model introduces a novel approach to solve the FER challenge by introducing the customized residual block with an attention

mechanism. This approach can effectively mitigate the over-fitting effect on the validation set.

2) **Improving feature extraction:** The model is designed to improve the quality of FER by extracting distinguishable features from the faces while expressing different emotions. This can lead to better performance in identifying emotions from facial expressions

The paper's outline is as follows: Section II represents previous related works. Section III reports on the datasets used in this study. Section IV introduces the main goal of the proposed model and its inner-workings. Section V reports the results achieved with the model on the datasets used in the study and Section V summarizes the experiments and results followed with discussion and conclusion in sections VI & VII.

II. Related work

Facial expression recognition (FER) received a significant attention in the computer vision community and a range of approaches are proposed various facets of this problem, ranging from traditional feature extraction (i.e., using combinations of geometric-based and texture-based features [16] combined with classification models such as Support Vector Machine (SVM) [17], [18], K-Nearest Neighbour (KNN) [19]), deep learning models (Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Attention mechanisms), and ensemble techniques.

Development of faster Graphical Processor Unit (GPU), provided the opportunity to identify deep features to represent various emotional facial expressions [20], specially in the cases where GPU empowered convolutional neural network (CNN) models are used for FER detection [21]. Stacking different convolutional layers on top of each other helps to find more sophisticated information from the facial features. These convolutional layers can be combined/ensembled with other models to develop hybrid models [22] or two or more CNN models can compete against each other to generate a generative adversarial network (GANs) for generating or discriminating pictures [23].

A. Application of Deep Learning models in FER

Main challenges associated with FER include over-fitting and improving the quality of expressive features [24]. Li and Deng [24] addressed the over-fitting problem through applying GAN for generating additional facial images. Yang et al. [25] introduced a variant of GAN for facial expression recognition in order to produce new pictures to augment the input dataset for FER. Instead of using pixel-level or feature-level differences for facial expression classification, the proposed model learned the deposition (or residue) that remained in the intermediate layers of the generative model [25]. Authors evaluated their model on the extended Cohn-Kanade (CK+) and Oulu-CASIA databases and reported 97.30% and 88% accuracy, respectively.

Mollahosseini et al. [26] developed a new parallel CNN model for FER reporting 81.7% accuracy in detecting six common facial expressions. Georgescu et al. [27] proposed a feature extraction network based on the famous VGG network. Three variants have been used for VGG with the name of VGG-13, VGG-f, and VGG-face networks as feature extraction network [28]. The study also used Supported Vector Machine (SVM) [29] for classification and achieved an imbalance classification accuracy of 87.76% for the FER+ dataset while demonstrating that developing a new model without a proper pre-processing technique negatively impacts classification performance.

Kumari et al. [30] developed a contrast-limited adaptive histogram equalization to improve the quality of pictures before importing them to the deep neural network. Authors used a four-layer CNN to classify emotions and evaluated the performances of the proposed structure using CK+ dataset, reporting 94.9% classification accuracy.

Tong et al. [31] proposed a new CNN model that focused on extracting local discriminating features from pictures to recognize emotions. Tong et al. used an overlapping block to extract local features from faces. The correlation between extracted features was calculated and developed with an adaptive weighting method to emphasize the importance of extracted features. Authors evaluated the feasibility of the proposed model using RAF-DB dataset and reported 89.863% accuracy.

B. Integration of Attention mechanism in FER problem

One of the mechanisms considered for improving quality of extracted features and improving FER classification performance is the use of attention layers. Attention layer in facial expression recognition provides the opportunity to improve network performance by focusing on the most relevant regions of the image. In this process, different sections of an image are selectively weighted which in turn allow the model to identify salient facial features including eyes, mouth, nose, and so on, that are more expressive of various facial expressions.

Farzaneh et al. [32] proposed deep attentive center loss as a mechanism for extracting discriminative features. The study utilized a loss function to integrate the attention mechanism to estimate attention weights correlated with feature importance using the intermediate spatial feature maps in CNN. Authors evaluated their proposed methods on the RAF-DB and AffectNet datasets and reported 80.44% and 65.20% accuracy, respectively.

Ma et al. [10] proposed a novel model to translate facial images into sequences of visual words using a convolutional visual transformer to deal with FER Authors used an attention selective fusion model to generate a feature map from input images. The authors proposed a model based on relationships between extracted visual words and the global self-attention to evaluate the model on RAF-DB, FERPlus, and AffectNet datasets, and they reported 88.14%, 88.81%, and 61.85%, respectively.

Zhi et al. [33], introduced a new model that incorporates spatial and channel attention layers to the original residual block of the ResNet 18 model. The model is comprised of 16 one-layer convolutional layers with an activation function. The proposed channel Attention layer comprises two paths with

an average and max-pooling layer followed by a Multi-Layer Perceptron (MLP). Finally, these two paths are combined by inner product to specify the most important features of FER. Zhi et al. [33] evaluated their proposed model using CK+ and achieved 89.52% classification accuracy in recognizing Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise.

To show the prospective of recent works on FER, a brief summary of literature studies with corresponding FER classification accuracies achieved using a collection of benchmark FER datasets are presented in Table II. As shown in the Table, majority of leading studies in FER domain considered hybrid architectures and emphasized on extracting more discriminative features. As evident from the results presented in Table II, a collection of datasets are considered by the FER community while there is still a need to gather new facial expression data that are more expressive of various facial emotions. The variation in reported performance with different datasets, ranging from 75% to 98% FER classification accuracy, further indicates the necessity of conducting FER studies and identifying mechanisms for improving the deep neural network architectures in order to identify deep facial expression features with better emotion discrimination.

Motivated by previous works, this study focuses on using a new backbone model for feature extraction and using attention layers to emphasize on the importance of extracted features aiming to improve the facial emotion classification.

III. Dataset

There are various open-source datasets available that are considered with FER community. Due to the versatility of these datasets, this study uses FER2013, CK+, and RAF-DB datasets to evaluate the proposed model. A detailed explanation of all these datasets is discussed below. Sample facial images from these datasets are shown in Figure 1.

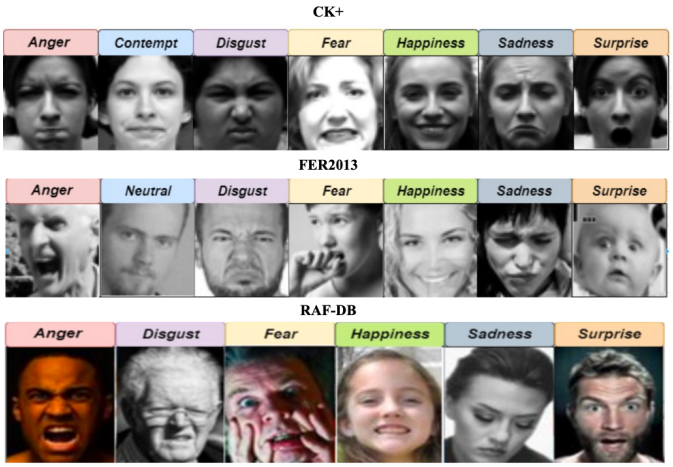

Fig. 1: A sample of different facial expressions from each dataset that are used in this study

A. Cohn-Kanade (CK+)

The extended version of Cohn-Kanade (CK+) contains 593 video sequences from 123 subjects aged 18 to 50. This dataset was collected in 2019. These videos are collected by 30 frames per second strategy, and the resolution of collected pictures is 40×490 or 640×480 pixels. CK+ dataset contains seven emotional categories of Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise [34]. This dataset does not provide a specific train, validation, and test set.

B. FER2013

Face Expression Recognition 2013 (FER2013) contains grayscale images with 48×48 pixels size. Faces in the FER2013 are centered in the middle, and seven facial expression categories such as Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral are available in the dataset. This dataset consists of 28,709 instances for training and 3,589 instances for evaluations [24].

C. Real-world Affective Faces Database (RAF-DB)

Real-world Affective Faces Database (RAF-DB), collected in 2017, consists of 29672 real-world images with two sets of simple and compound emotions [35]. The first set of labels contains simple emotions such as Fearful, Happy, Surprised, Sad, Angry, Disgusted, and Natural. Complex labeled emotions in the dataset is consisted of 12 labels such as Fearfully Surprised, Angrily Surprised, and so on. In current study, 15,339 instances of all images belong to the basic emotions, among which 12,271 samples are placed in training set and 3,068 samples are placed in testing set.

IV. Proposed Model

This paper proposes a deep learning model that utilizes a Convolutional layer for feature extraction. The architecture has three main components: Backbone, Emphasizer, and Classifier.

A. Backbone

The proposed model uses VGG19 [36] as the backbone to extract initial features. This model is pre-trained with the ImageNet dataset [37], which has 1.2 million high-resolution images in 1000 different classes. As VGG19 comprises 16 convolutional and three dense layers, it is well-tuned to extract facial features from the faces [38]. The backbone produces the first entry of the activation map by convolving input images through its filters. The activation map is generated by repeating this process for every element of the input image. The output volume of the convolutional layer is generated by stacking the activation maps of every filter along the depth dimension. All the neurons in an activation map also share parameters. Due to the local connectivity of the convolutional layer, the network is forced to learn filters that have the maximum response to a local input region.

B. Emphasizer (Attention Layer)

In addition to using VGG19 as a backbone feature extractor, the proposed model is comprised of a customized residual block based on the four convolutional layers. The main goal of a residual block is to solve the deterioration issue that arises from too-deep deep neural networks. The complexity of deep network training leads to the degradation issue. It is challenging for previous layers to learn meaningful representations

throughout the backpropagation process because gradients frequently shrink as they pass through several layers. The term ”vanishing gradient” is frequently used to describe this issue. The vanishing gradient issue can be solved with the use of a residual block, which also makes deep network training effective. Between each layer, a batch normalization layer is placed to preserve the extracted feature map's original average and variance. The proposed customized residual block is based on four convolutional layers followed by batch normalization allows the model to learn the optimal scale and mean of each input layer and to avoid vanishing gradient descent during the process of training the model [39]. Dropout is used to prevent the model from getting over-fitted. The architecture of the proposed residual block is shown in Figure 2.

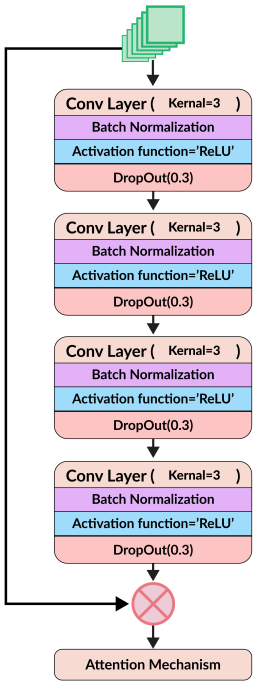

Fig. 2: Architecture of the Proposed Residual Block

As the number of convolutional layers in the model increases, the width and height of the extracted features decrease while the number of extracted feature maps increases. This results in extraction of various types of information from this long-sequence of convolutional layers. Through increasing sequence length, the DL models forget to use important extracted features during the earliest classification stage. To magnify and remember important extracted information, an attention mechanism is used [40]. The attention layer converts extracted features into a vector and calculates the alignment factor to emphasize how many extracted features need to be considered to create an output. The process of using the attention mechanism is shown in Figure 3.

As shown in Figure 3, the model returns two types of features from the customized residual block; extracted features

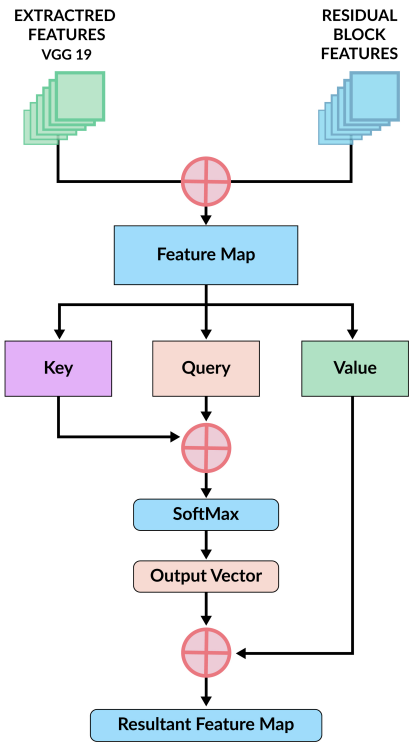

Fig. 3: Overview of the Attention Mechanism in the Architecture

by the convolutional layers and generated features using the backbone architecture model (VGG19) that increase the width and height of the convoluted features. To calculate the alignment factor, an element-wise multiplication is performed. Finally, an attention mechanism followed by a residual block is added to consolidate the attention signal.

C. Classifier

A Multi-Layer Perceptron (MLP) is used for extracting features and recognizing emotions. The model uses an MLP with 1-3 dense layers before the last softmax layer to ensure the best performance is achieved for identifying emotions. The classifier is versatile and can expand or shrink based on achieved accuracy. Figure 4 illustrates the complete architecture of the proposed model.

V. Experimental Results

The proposed model in this study uses a Deep Learning ensemble that combines feature extraction with an attention mechanism for emotion recognition. The model is trained using an Adam optimizer with Adamax and a batch normalization of 128. The training is carried out for 1000 epochs, and early stopping techniques are employed to save the best-trained model based on accuracy and interrupt the training procedure with no progress after 100 epochs to prevent over-fitting.

The proposed deep learning model for facial expression recognition utilizes a convolutional layer for feature extraction, followed by an attention mechanism to enhance the importance

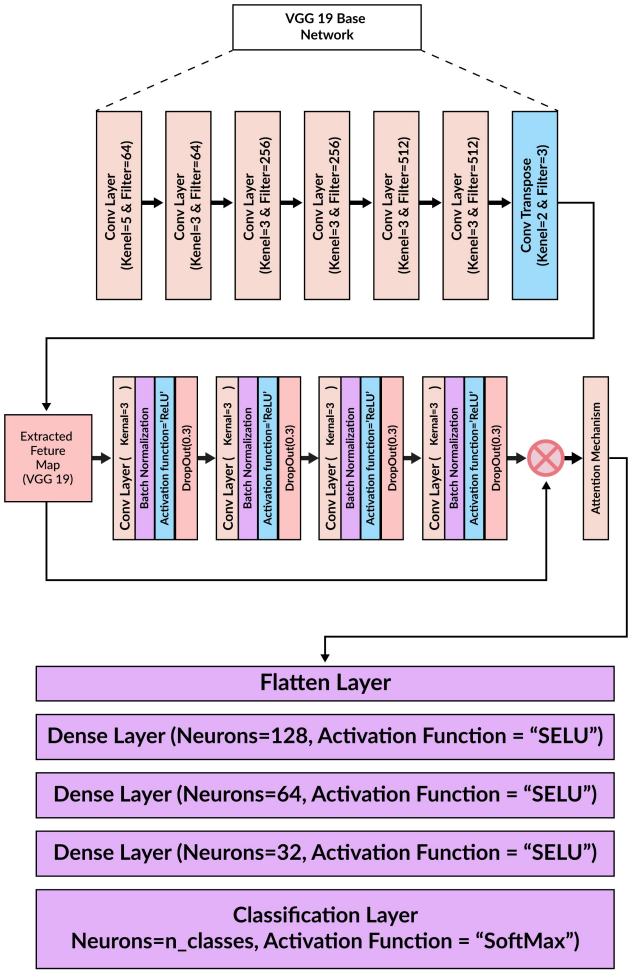

Fig. 4: The proposed Attention-Based CNN Model

of extracted features, and finally, an MLP for classification. The model was trained using an Adam optimizer with early stopping techniques to avoid over-fitting. The model achieved high accuracy in recognizing emotions such as Anger, Fear, and Happiness, which are characterized by distinguishable features on the lips or eyebrows. However, recognizing neutral emotions can be more challenging as there are no clear signs from facial features. The attention mechanism is used to help the model remember important information for emotion recognition. The input dataset is augmented using techniques such as Random Horizontal Flip, Random Vertical Flip, and normalization. The performance achieved on CK+, FER, and RAF-DB datasets are promising, as shown in Table I. The accuracy and loss on the training and validation set during the training of the model are presented in Figures 5 and 6 respectively.

While a degree of performance variation is observed across datasets used in the study, the results are within the range reported by literature on these datasets (see Table II). It is noticeable that among all datasets, Natural and Sad facial emotions in RAF-DB datset achieved the lowest classification performances of 71.3% and 79.1% respectively while Neutral emotion classification achieved 99.9% accuracy in FER2013 dataset. It is noteworthy that Sad emotion also showed a low performance of 81.3% classification accuracy in FER2013 albeit this emotion achieved 99.5% classification accuracy in CK+ dataset, further attenuating the performance differences across these benchmark datasets, possibly reflecting the variation the way such emotional facial expression is performed from one dataset to the next causing a degree of inconsistency in emotion prediction performances across datasets.

TABLE I: Result of 5 folds cross validation of seven classes

Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Avg.	F1
CK+								
Anger	Contempt	Disgust	Fear	Happy	Sad	Surprise	Avg.	F1
99.6%	99.6%	99.7%	99.6%	96.2%	99.5%	99.6%	99.13%	99.32%
FER2013								
Anger	Neutral	Disgust	Fear	Happy	Sad	Surprise	Avg.	F1
99.7%	99.9%	99.2%	99.8%	99.6%	81.3%	83.9%	89.06%	88.89%
RAF-DB								
Fear	Natural	Disgust	Anger	Happy	Sad	Surprise	Avg.	F1
95.1%	71.3%	98.3%	80.3%	99.8%	79.1%	99.6%	88.8%	84.45%

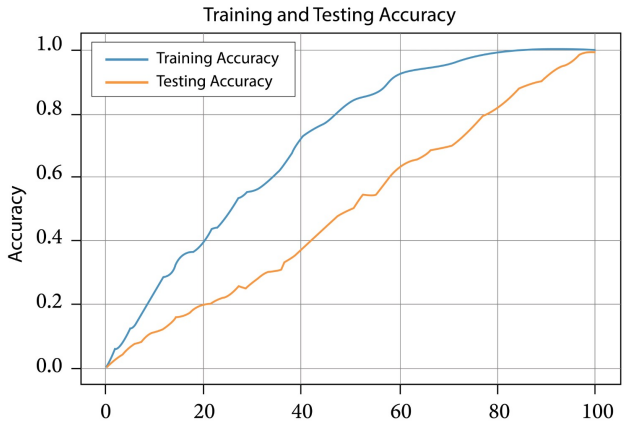

Fig. 5: Training and Testing Accuracy Graph

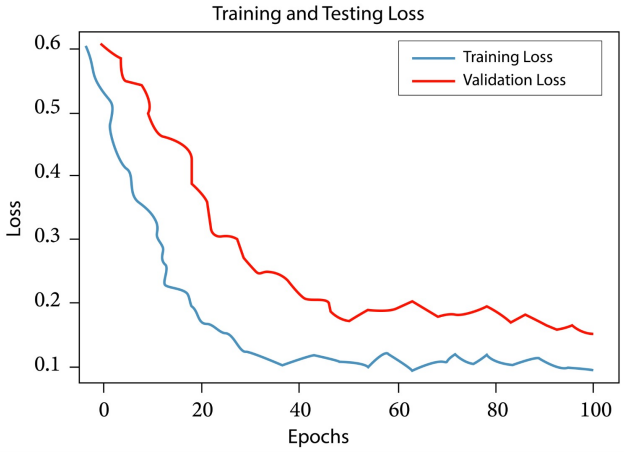

Fig. 6: Training and Testing Loss Graph

Normally real-world environments are far different from the simulated environment. The trained model usually faced

noise and lighting issues which decrease the performance of the model. different types of noises were added to the testing samples of each dataset and the noised samples were used to evaluate the performance of the proposed model in a noisy environment. Gaussian noise, blurring, brightening, and darkening effects were applied to the test samples with different intensity values to evaluate the robustness of the proposed model. The main purpose of this experiment is to evaluate how the proposed model will perform in restricted and noisy environments. The accuracy scores of the proposed model on different types of noisy images are presented in Table III. The proposed model also showed minimum 93.45%, 83.59% and 84.45% accuracy scores for CK+, FER, and RAF-DB dataset.

TABLE II: Comparison of the proposed approach with recent models on different datasets

Author	**Model**	**Acc.(%)**	**Dataset**
Ghazouani [41]	Genetic CNN	98.0	CK+
Wang et al [42]	OAENet	98.13	CK+
Li et al [43]	Auto-FERNet	98.89	CK+
Zhi et al [33]	CNN	89.52	CK+
Kumari et al [30]	Shallow CNN	94.9	CK+
Proposed work	**CNN+Attention**	**99.13**	**CK+**
Tong et al [31]	CNN with Overlapping Blocks	89.86	RAF-DB
She et al [44]	Auxiliary CNN with multi-branch	87.76	RAF-DB
Proposed work	**CNN+Attention**	**88.89**	**RAF-DB**
Kim et al. [45]	Multi-task cascade neural network	75	FER2013
Pramerdorfer et al. [46]	CNN (Inception)	88.26	FER2013
Proposed work	**CNN+Attention**	**89.06**	**FER2013**

VI. Discussion

The study introduced a novel deep-learning model that addresses the challenging task of facial expression recognition. The proposed model combined feature extraction using a backbone with an attention mechanism to enhance the recognition of important extracted features. The model also utilizes an MLP for the classification of the extracted features. The performance of the model is evaluated using three datasets: CK+, FER, and RAF-DB. To compare the results of the proposed model with similar studies, the authors reviewed some articles that utilized the same datasets with different DL model architectures. The results presented in Table I show that the proposed model outperformed recent works on the FER2013 dataset and had comparable results for CK+ and RAF-DB. The model demonstrated promising results for recognizing emotions such as happiness, fear, and disgust, while neutral and sad emotions proved to be more challenging. Moreover, the authors evaluate the performance of the proposed model under different environmental conditions. For this, several types of noises were added to test the robustness of the model which produced significant results (Table III). Despite the promising results, the authors recognize the need to work with multiple datasets collected under different conditions and to explore various feature extraction and ensemble techniques to further improve the model.

TABLE III: Robustness Evaluation of the Proposed Model.

Noise	Value	Average Accuracy Score		
		CK+	FER-DB	RAF 2013
Gaussian	(0.005)	97.76	87.73	85.26
Gaussian	(0.01)	96.89	86.34	84.37
Salt & Paper	(0.005)	94.34	85.00	86.58
Salt & Paper	(0.01)	92.45	84.89	86.01
Brightening	10%	98.54	86.56	85.28
Brightening	20%	98.05	85.96	84.33
Darkening	10%	97.23	85.12	84.82
Darkening	20%	96.87	84.45	83.59

VII. Conclusion

Facial expression recognition (FER) is one of the most challenging tasks in computer vision due to the uneven distribution of emotions in real-world datasets and the similarity between various facial features in expressing certain emotions. To address these challenges, this study proposed a novel combination of a convolutional neural network (CNN) as a backbone feature extractor and a customized residual block with an attention mechanism to extract more distinguishable features and improve emotion recognition performance. The proposed model is evaluated on three datasets, FER, CK+, and RAF-DB, achieving an accuracy of 89.06%, 99.13%, and 88.89%, respectively. The model is then compared to recent state-of-the-art models, and the experimental results demonstrate better performance on FER 2013 than similar studies. Additionally, the proposed model shows an extraordinary detection rate for recognizing Happy, Fear, and disgust emotions. Moreover, the proposed model robustly performed on the noisy environments that were achieved by manually adding the noise in test samples. Despite the promising results achieved, the authors plan to work on multiple datasets, including those collected in the wild, and explore different feature extraction and ensemble techniques to create an even better model in future.

References

[1] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5562–5570.

[2] R. J. Hassan, A. M. Abdulazeez *et al.*, "Deep learning convolutional neural network for face recognition: A review," *International Journal of Science and Business*, vol. 5, no. 2, pp. 114–127, 2021.

[3] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7660–7669.

[4] F. Zhang, M. Xu, and C. Xu, "Weakly-supervised facial expression recognition in the wild with noisy data," *IEEE Transactions on Multimedia*, 2021.

[5] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5375–5384.

[6] U. Chinta, J. Kalita, and A. Atyabi, "Soft voting classifier for multimodal emotion recognition using deep-learning methods - facial images and eeg," *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, 2023.

[7] U. Chinta and A. Atyabi, "A framework pipeline to address imbalanced class distribution problem in real-world datasets," *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, 2023.

[8] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, "Finding the best classification threshold in imbalanced classification," *Big Data Research*, vol. 5, pp. 2–8, 2016.

[9] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 222–237.

[10] F. Ma, B. Sun, and S. Li, "Robust facial expression recognition with convolutional visual transformers," *arXiv preprint arXiv:2103.16854*, 2021.

[11] R. Wightman, H. Touvron, and H. Jégou, "Resnet strikes back: An improved training procedure in timm," *arXiv preprint arXiv:2110.00476*, 2021.

[12] A. D. Gavrilov, A. Jordache, M. Vasdani, and J. Deng, "Preventing model overfitting and underfitting in convolutional neural networks," *International Journal of Software Science and Computational Intelligence (IJSSCI)*, vol. 10, no. 4, pp. 19–28, 2018.

[13] H. H. Tan and K. H. Lim, "Vanishing gradient mitigation with deep learning neural network optimization," in *2019 7th international conference on smart computing & communications (ICSCC)*. IEEE, 2019, pp. 1–4.

[14] Y. Ji, H. Zhang, and Q. J. Wu, "Salient object detection via multi-scale attention cnn," *Neurocomputing*, vol. 322, pp. 130–140, 2018.

[15] X. Tong, S. Sun, and M. Fu, "Data augmentation and second-order pooling for facial expression recognition," *IEEE Access*, vol. 7, pp. 86 821–86 828, 2019.

[16] R. Kadota, H. Sugano, M. Hiromoto, H. Ochi, R. Miyamoto, and Y. Nakamura, "Hardware architecture for hog feature extraction," in *2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE, 2009, pp. 1330–1333.

[17] V. Cherkassky and Y. Ma, "Practical selection of svm parameters and noise estimation for svm regression," *Neural networks*, vol. 17, no. 1, pp. 113–126, 2004.

[18] H. I. Dino and M. B. Abdulrazzaq, "Facial expression classification based on svm, knn and mlp classifiers," in *2019 International Conference on Advanced Science and Engineering (ICOASE)*. IEEE, 2019, pp. 70–75.

[19] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2003, pp. 986–996.

[20] A. R. Brodtkorb, T. R. Hagen, and M. L. Sætra, "Graphics processing unit (gpu) programming strategies and trends in gpu computing," *Journal of Parallel and Distributed Computing*, vol. 73, no. 1, pp. 4–13, 2013.

[21] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (cnn) in vegetation remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 24–49, 2021.

[22] A. Tavakolian, F. Hajati, A. Rezaee, A. O. Fasakhodi, and S. Uddin, "Fast covid-19 versus h1n1 screening using optimized parallel inception," *Expert Systems with Applications*, p. 117551, 2022.

[23] L. Gonog and Y. Zhou, "A review: generative adversarial networks," in *2019 14th IEEE conference on industrial electronics and applications (ICIEA)*. IEEE, 2019, pp. 505–510.

[24] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, 2020.

[25] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2168–2177.

[26] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[27] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64 827–64 836, 2019.

[28] W. Wang, C. Zhang, J. Tian, X. Wang, J. Ou, J. Zhang, and J. Li, "High-resolution radar target recognition via inception-based vgg (ivgg) networks," *Computational Intelligence and Neuroscience*, vol. 2020, 2020.

[29] G. N. Kouziokas, "Svm kernel based on particle swarm optimized vector and bayesian optimized svm in atmospheric particulate matter forecasting," *Applied Soft Computing*, vol. 93, p. 106410, 2020.

[30] N. Kumari and R. Bhatia, "Efficient facial emotion recognition model using deep convolutional neural network and modified joint trilateral filter," *Soft Computing*, pp. 1–14, 2022.

[31] X. Tong, S. Sun, and M. Fu, "Adaptive weight based on overlapping blocks network for facial expression recognition," *Image and Vision Computing*, vol. 120, p. 104399, 2022.

[32] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2402–2411.

[33] J. Zhi, T. Song, K. Yu, F. Yuan, H. Wang, G. Hu, and H. Yang, "Multi-attention module for dynamic facial emotion recognition," *Information*, vol. 13, no. 5, p. 207, 2022.

[34] H.-S. Lee and B.-Y. Kang, "Continuous emotion estimation of facial expressions on jaffe and ck+ datasets for human–robot interaction," *Intelligent service robotics*, vol. 13, no. 1, pp. 15–27, 2020.

[35] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[38] Z. Tao, X. Zhou, Z. Xu, S. Lin, Y. Hu, and T. Wei, "Finger-vein recognition using bidirectional feature extraction and transfer learning," *Mathematical Problems in Engineering*, vol. 2021, 2021.

[39] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" *Advances in neural information processing systems*, vol. 31, 2018.

[40] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 705–10 714.

[41] H. Ghazouani, "A genetic programming-based feature selection and fusion for facial expression recognition," *Applied Soft Computing*, vol. 103, p. 107173, 2021.

[42] Z. Wang, F. Zeng, S. Liu, and B. Zeng, "Oaenet: Oriented attention ensemble for accurate facial expression recognition," *Pattern Recognition*, vol. 112, p. 107694, 2021.

[43] B. Li and D. Lima, "Facial expression recognition via resnet-50," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 57–64, 2021.

[44] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6248–6257.

[45] J. H. Kim and D. S. Han, "Data augmentation & merging dataset for facial emotion recognition," in *Proceedings of the Symposium of the 1st Korea Artificial Intelligence Conference, Jeju, Korea*, 2020, pp. 12–16.

[46] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: state of the art," *arXiv preprint arXiv:1612.02903*, 2016.