An Improved Facial Expression Recognition using CNN-BiLSTM with Attention Mechanism

Samanthisvaran Jayaraman[1], Anand Mahendran[2]
School of Computer Science and Engineering, Reserach Scholar, Vellore Institute of Technology, Tamilandu, India[1]
School of Computer Science and Engineering, Professor Grade1, Vellore Institute of Technology, Tamilandu, India[2]

Abstract*—In the recent years, Facial Expression Recognition is one of the hot research topics among the researchers and experts in the field of Computer Vision and Human Computer Interaction. Traditional deep learning models have found it difficult to process images that has occlusion, illumination and pose dimensional properties, and also imbalances of various datasets has led to large distinction in recognition rates, slow speed of convergence and low accuracy. In this paper, we propose a hybrid Convolution Neural Networks-Bidirectional Long Short Term Memory along with point multiplication attention mechanism and Linear Discriminant analysis is incorporated to tackle aforementioned non-frontal image properties with the help of Median Filter and Global Contrast Normalization in data preprocessing. Following this, DenseNet and Softmax is used for reconstruction of images by enhancing feature maps with essential information for classifying the images in the undertaken input datasets i.e. FER2013 and CK+. The proposed model is compared with other traditional models such as CNN-LSTM, DSCNN-LSTM, CNN-BiLSTM and ACNN-LSTM in terms of accuracy, precision, recall and F1 score. The proposed network model achieved highest accuracy in classifying the facial images on FER2013 dataset with 95.12% accuracy which is 3.1% higher than CNN-LSTM, 2.7% higher than DSCNN-LSTM, 2% higher than CNN-BiLSTM and 3.7% higher than ACNN-LSTM network models, and the proposed model has achieved 98.98% of accuracy with CK+ in classifying the images which is 5.1% higher than CNN-LSTM, 5.7% higher than DSCNN-LSTM, 3.3% higher than CNN-BiLSTM and 6.9% higher than ACNN-LSTM network models in facial expression recognition.

Keywords*—Facial expression recognition; occlusion; attention mechanism; convolution neural networks; bidirectional long short time memory

I. Introduction

In general, facial expression contains a vital non-verbal communication of a human that includes eye contact, hand gestures and etcetera. Basically, these factors convey their emotion, inner thoughts, intention and mental states. This has increased the interest among the scientists, researchers and academicians on studying about human emotions and expressions [1].Machine learning approaches play important role in various kind of research works such as security, emotion detection, natural disaster management, data protection and monitoring [2], [3]. Human emotions play an important role in both psychology and computer vision in which emotion is classified into categorical and dimensional respectively. In case of categorical model, emotions are treated as happy, sad, neutral, anger, fear, surprise and more whereas in dimensional model emotion is treated as valence and arousal. Facial Expression Recognition (FER) is part of computer vision with a lot of practical applications and the number of studies on FER has been increasing over the last two or more decades [4], [5].

In the past, Convolutional Neural Networks (CNN) was very much successful in addressing this issue and performed well in extracting features from the given input images. But, CNN was suffering with issues like vanishing gradient and decreasing the accuracy of deep networks [6]. Residual Neural Networks (ResNet) was in introduced in 2015 by He and Zhang et al. and which helped them to add residual learning to CNN to tackle the issues faced by CNN [7]. A dataset can have a variety of facial images with different poses, contrast, brightness, dimensions, age and some of the images might have unclear properties where some part of the face could be hidden or occluded. For any given facial image, human's expression can be identified with its unblocked regions. When some part of the facial image is blocked, the expression has to be determined based on systematic part or other regions that are highly visible or clear [8], [9]. Despite, the contributions and the accuracy performance delivered by various deep learning approaches for facial expression recognition, the classification accuracy could still be improved with additional methods and mechanisms [10].

To overcome these challenges, the concept of Attention Mechanism (AM) was proposed in computer vision and thereafter used in natural language processing as well. The main objective of having attention mechanism is to obtain richer information from input facial images by paying more attention to the key parts of image features [11]. The problem of information redundancy and the loss of key features of the input images can be prevented while using Attention Mechanism (AM) along with deep neural networks [12].

In deep learning, attention mechanism is not only important, but are ubiquitous, integral part and necessary element in neural machine learning techniques. The main function of this mechanism is to optimize the issue of learning the desired target by non-uniformly weight the contributions of input feature vectors [13].

Currently, the attention mechanism has a crucial role in determining human perception and is applied successfully in various fields of deep learning such as machine translation, image generation and some other fields as well. There are few many researches done on expression recognition using attention mechanism [14]–[16].

This research work is aimed to predict the emotions of a driver using CNN-BiLSTM approach. Section 2 discuss about the existing works related to the topic of the study and section 3 introduces authors proposed approach and algorithms used

in this study. Section 4 presents the experimentation results of this study and results are compared with other works which is followed by discussion and justification of this study. Finally, the conclusion of this study is presented and future work of the authors also presented.

II. Related Work

In [17], Identity Aware CNN(IA-CNN) model was proposed in which the importance is given to identity and expression, sensitive contrastive loses. This was considered to reduce the variations in learning the information related to identity and expression. In [18], end to end architecture was proposed along with attention model. In [10], a novel Region Attention Network (RAN) was proposed which was robust with real world pose of images and occlusion variations in such images. This approach was effective in capturing important facial regions for occlusion and pose variant to deliver expected results of Facial Expression Recognition.

In [19],a Region Aware Subnet (RASnet) that locates expression related regions using binary masks and those critical regions are identified with coarse-to-fine granularity levels and Expression Recognition subnet (ERSnet). The study has used Multiple Attention mechanism learns discriminative features of an input image and this MA block consists of hybrid attention branch with many sub branches where region specific attention is performed by each sub branch. In [20], the authors have proposed Distract Your Attention Network (DAN) which consists of three components: Feature Clustering Network (FCN) Multi-head Cross Attention Network (MAN) and Attention Fusion Network (AFN). To maximize class separability FCN extracts robust features using large margin learning. MAN builds attention maps on critical regions by simultaneously attending multiple facial areas through instantiating multiple attention heads. AFN fuses the created these attentions maps and converts it into a comprehensive one.

In [19], the authors have proposed a Multiple Attention Network with three components, namely, Multi-Branch Stack Residual Network (MRN) which deploys attention heads on critical facial regions to generate attention maps, Transitional Attention Network (TAN) which learns objectives to maximize class separability and Appropriate Cascade Structure (ACS) which determines the appropriate construction method for the model. In [21], the authors have suggested that attention mechanism help to focus on more useful features, therefore, they have proposed an end to end network with AM for automatic facial expression recognition. For their experimentation purposes, the study considered its own data set of 35 subjects from different peoples between the age group of 20 to 25. In total, the study had 26950 images that includes both RGB and depth images. Local Binary Pattern (LBP) was adopted for feature extraction and the results were compared with JAFFE, CK+, FER2013 and Oulu-CASIA datasets.

In [22], the authors have proposed an RCLnet to recognize wild facial expression which has high occlusion or illumination using attention mechanism and LBP feature fusion method. The proposed model had two branches: ResNet-CBAM, the residual attention branch and local binary feature extraction branch (RCL-Net). The study has performed validation on for different datasets: FER2013, FERPLUS, CK+ and RAF-DB datasets. In [9], authors proposed Attention based CNN (ACNN) that could recognize information from occlusion region of facial images and focus on un-occluded regions of the same image. In the end, the model combines multiple representations (each weighted via gate unit) from facial regions of interest. The study has also used two types of CNN, namely, patch based ACNN and global-local-based ACNN where the experimentation results proved that their proposed model has improved recognition accuracy for both occluded and non-occluded facial images.

In [23], the authors have proposed an Enhanced CNN with attention mechanism to recognize occluded facial images of RAF-DB dataset and their experimentation results has achieved 86.2% of accuracy with patch based ECNN-AM and Global Gated Unit (GG-U) which automatically weighs global facial representations. In [24], the authors have developed Deep CNN along with Binary Attention Mechanism (BAM) that is trained with original pixel data characteristics. Data preparation was done using Histogram of Oriented Gradients (HOG), dropout and batch normalization along with L2 regularization was employed to minimize the over fitting issue. The proposed model has used FER2013 dataset to extract and examine the performance of their approach with various metrics.

In [25], a Symmetric Speed up Robust Features (SURF) framework was used to identify the hidden part of images by critically locating a horizontal symmetric area and heterogeneous soft partitioning assigned weights for each part of the input image recognition while training. The weighted image was given as input to the trained network model to detect facial expression recognition and the experimentation was performed Cohn-Kanade (CK+) and FER2013 datasets. The results have showed 7% to 8% improvement compared to other works. In [26]–[28],], AlexNet based Deep CNN was used to tune the outputs obtained in three steps: the first two steps are in training stages where frontal images of FER2013 dataset used and the third stage included non-frontal image poses of the same data set. The experimentation was conducted using VT-KFER and 300W databases where the results have outplayed other systems in expression recognition.

In [29], two layer based CNN-LSTM mechanism was proposed which extracts rich information from important regions of FER2013 and CK+ datasets. This approach has outperformed some other methods like CNN-ALSTM, ACNN-ALSTM and patch based ACNN. In [30], CNN-BiLSTM model was proposed and was experimented on CK+ dataset and to prevent over fitting data augmentation was incorporated. This approach was compared with CNN and CNN-LSTM models in terms of accuracy, and the proposed approach returned improved results.

III. Proposed Methodology

Based on the observations from traditional Convolution Neural Networks (CNN) and its performance on Facial Expression Recognition (FER), it is found that existing models based on CNN fails to extract rich and useful information from key parts of occluded, variety pose and blurred input images. In our previous work, we have proposed CNN-LSTM based hybrid modelto extract rich information from frontal images along with point multiplication attention mechanism to correctly identify the expression with improved accuracy.In

this paper, we propose CNN-BiLSTM based hybrid model with point multiplication attention mechanism, and some other methods for the betterment of recognizing facial expressions with improved accuracy. The proposed model consists of four important components and the same is represented in Fig. 3. The components are, namely, CNN, Bi-LSTM, Attention layer and reconstruction and classification layer.

A. Data Preprocessing

CK+ dataset is a better organized dataset with quality images since the quantity of images are very small when compared with FER2013 dataset. Since FER2013 dataset consists of large number of images, data preprocessing is very important to obtain quality images with rich feature information. In our first work, we only considered 7074 images from FER2013 dataset which has both quality and resolution; here, we consider the whole dataset. Thus, preprocessing is performed by resizing images into 128*128 pixels, median filters are used for noise removal and normalization is done using Global Contrast Normalization (GCN).

Generally, a dataset contains images with different sizes and with varying pixels. Hence, we resize the dataset images into 128*128 pixels to ensure uniformity of images in the dataset under study. Resizing of images include both enlarging the size and reducing the size of an image through cropping. To overcome the uncertainty or variations present in the image such as brightness, color and etc., unwanted noise is removed using Median Filters (MF). Median filter is a non linear operation and commonly used to remove 'salt and pepper' noise and it removes the noise and preserve edges simultaneously. To deal with poor contrast image feature, GCN is used to normalize the image to ensure uniform intensity with improved visualization of an image. The basic operation of GCN is to subtract each pixel value of an image with mean value and then divides it with standard deviation. The equation is derived as follows,

$$X'_{i,j,k} = s\frac{X_{i,j,k} - \bar{X}}{\max\left\{\in, \sqrt{\lambda + \frac{1}{3\mathrm{rc}}\sum_{\mathrm{i=1}}^{\mathrm{r}}\sum_{\mathrm{j=1}}^{\mathrm{c}}\sum_{\mathrm{k=1}}^{3}\left(X_{i,j,k-\bar{X}}\right)^2}\right.} \tag{1}$$

In equation (1), X represents the image and i,j,k represents row, column and color depth of the image X and $\bar{X}$ represents the mean intensity of entire image.

B. Training of CNN

Considering the quantity of images we deal with the datasets undertaken for this study, training those images is a challenging task. Though K-Nearest Neighbor (KNN) and Siamese are commonly used approaches, we have considered a new center loss function from [31] for the enhancement of discriminative power of the deeply learned features. For the deep features of each image classes, the center is learnt while training the dataset. During training, the distance between the deep features and its class centers are minimized and updated simultaneously. The update is done by using mini-batch since updating centers of every class for each iteration is inefficient and practically impossible.

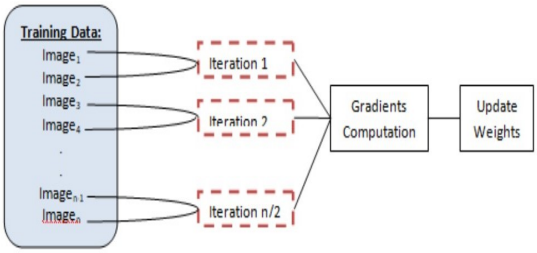

Fig. 1. Mini-batch processing of dataset images.

When a researcher deal with large amount of images, the model should ensure that the batch should be less than the original dataset, yet effective in making batches. Iterations are required to minimize the redundancy, thus computational complexity can be reduced. Importantly, batches can operate in even numbers i.e. from 2 to 2^n . Fig. 1 represents the working principle of mini-batch processing with the number of iterations required in order to assign weights for landmarks of the face with point multiplication mechanism.

Thus, for each iteration the center is computed only for corresponding classes (not all centers are updated) and large perturbations are avoided b adopting scalar factor $'\alpha'$ to control the learning rate of the centers. This can be computed as follows,

$$\frac{\partial L_C}{\partial x_i} = x_i - C_{y_i} \tag{2}$$

$$\Delta C_j = \frac{\sum_{i=1}^{m}\delta\left(y_i = j\right)\cdot\left(C_j - x_i\right)}{1 + \sum_{i=1}^{m}\delta\left(y_i = j\right)} \tag{3}$$

where δ(condition)=1 if the condition is satisfied and δ(condition)=0 if not, importantly the value of α is restricted between [0,1]. For discriminative feature learning purpose, this study has considered 'joint supervision' of softmax loss and center loss $L = L_S + \lambda L_C$ to train CNN model. The equation is given as,

$$L = -\sum_{i=1}^{m}\log\frac{e^{W_{y_i}^T x_{i+by_i}}}{\sum_{j=1}^{n} e^{W_j^T x_{i+b}}} + \frac{\lambda}{2}\sum_{i=1}^{m}\left\|x_i - C_{y_i}\right\|_2^2 \tag{4}$$

To balance both center loss and softmax loss, $'\lambda'$ scalar function is used for the better training of CNN.

C. Network Architecture of Proposed Approach

In this part of the paper, we propose hybrid CNN-BiLSTM model with point multiplication Attention Mechanism along with Linear Discriminant Analysis method for efficient facial expression recognition with improved accuracy and other matrices. The proposed approach consists of following components: Data preprocessing, CNN with Feature Extraction, BiLSTM with Attention Mechanism followed by dimensionality reduction using LDA, the reconstruction module with

DenseNet. Finally, the classification module uses Softmax which categorizes the classes of images with respective expression. Fig. 2 represents the network architecture of our proposed work.

The Network is composed of seven layer convolution neural network with four convolution layer and three down-sampling layers. The parameters of each CNN layer is set with following convolution kernels: (1*1, 32), (5*5, 32), (3*3, 32) and (3*3, 64) respectively. In the convolution layer setup is in (N*N, K) where N*N represents number of convolutions performed and K (32, 64) represents the number of feature maps created. This layer of the network model aims to extract abstract features from the local region of the facial expression images, and generates and serializes the feature vector which is fed to BiLSTM network as an input. The layer begins with C1 that performs point-by-point convolution on the input image with 1*1 convolution kernel. This not only helps to improve the feature representation ability but also beneficial to increase the non-linear representation of the input as well. Since, 1*1 convolutions have few parameters the network calculation complexity can be reduced too. The pooling layer employs the method of maximum-pooling to perform further extraction on the input which identifies and returns strongest features. Thus, the computational complexity is reduced along with the resolution of the feature map using local aggregation function.

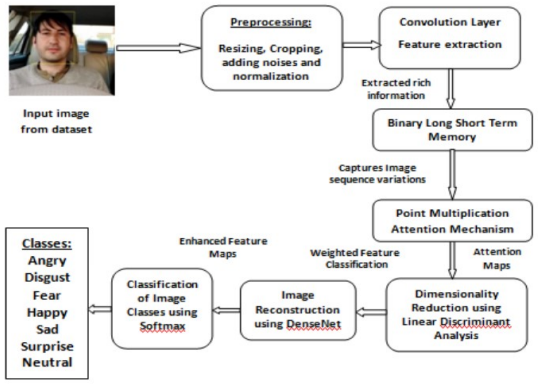

Fig. 2. Architecture of proposed CNN with attention mechanism based LSTM.

D. Bidirectional Long Short Term Memory

The core idea of LSTM is the state with few linear interactions and it also uses internal mechanism refereed as 'gate' to regulate the flow of information by determining what data to retain and what data to discard using forward and backward layer. Fig. 3 represents the Bi-LSTM model used in this paper that consists of a forward LSTM and backward LSTM network layers. Bi-LSTM helps in getting time series information of difference images as it deals with images that are taken at different periods and different angles. When such information is fully considered, obtaining time series feature vectors is trustworthy and the same will be forwarded to the attention mechanism module of our proposed approach.

This model assumes six shared weights $w1$ to $w6$ that are calculated in the forward layer from time 1 to t and the output h_t is obtained and saved. In the same way, reverse process was performed to obtain h'_t in the backward layer and then the final output O_t is obtained by combining both forward and backward layer outputs. The following equations represent the whole process.

$$h_t = (w_1 x_t + w_2 h_{t-1}) \tag{5}$$

$$h'_t = \left(w_3 x_t + w_5 h'_{t+1}\right) \tag{6}$$

$$O_t = g\left(w_4 h_t + w_6 h'_t\right) \tag{7}$$

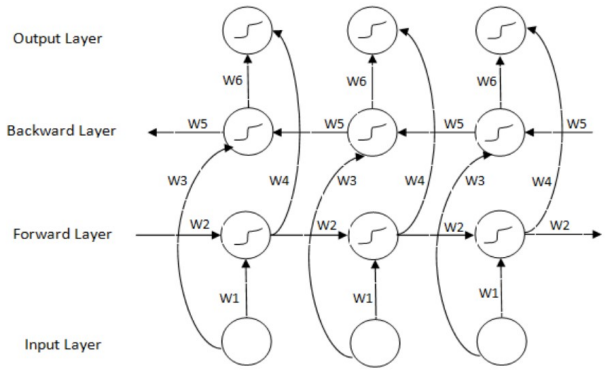

Fig. 3. Bi-LSTM Structure of proposed approach.

By extracting sequence features from the images of undertaken datasets (FER2013 and CK+), the context relationship of the sequence is automatically generated which not only helps in increasing the amount of information to the network model, but also helps in improving the accuracy of facial image expression recognition.

E. Attention Layer

In general, attention layer increases the weight of useful features that are identified in both frontal and non-frontal images of the dataset (especially occluded, complex and blurred images). Upon identifying important features, this layer encourages the network to focus more on such vital features that in return help the network model to recognize and classify the facial expression images more accurately [21] Mainly, the concept of attention layer is inspired by the special attention paid by humans when required, but in computer vision, it represents a weighted mean function. This layer takes three parameters as inputs: the query, the values and the keys. Normalization of attention vector dV is completed using softmax activation function, yet attention mechanism equation is a hyper-parameter [32], [33]. The relevant equations and algorithm is presented in algorithm 1.This attention module focuses more on the useful features and increases its weights

to enhance the efficiency of the model to recognize different expressions present on input images.

Algorithm 1: Point Multiplication Attention Mechanism

INPUT: Image sequence variations from BiLSTM
OUTPUT: Attention Vector Maps

BEGIN
FOR each hidden image sequence variations $L = \{L_1, L_2, \ldots\ldots\ldots L_{N-1}, L_N\}$ at moment n
Initiate random weight matrix W and Value matrix V
IF $W \leq W_6$ then
Compute learned key matrix $K_L = \tanh VW^a$
//where V = value matrx and W^a = weight factor

ENDIF
WHILE $(L! = 0)$
Find current key matrix $K_c = ||qC_v||$
// where q = query and C_v = current key value
Compute normalized weight vector $d = \mathrm{softmax}\left(qK^T\right)$
Compute Attention Vector $a = d * V$
END WHILE
END FOR
END

F. Dimensionality Reduction with Classification

The objective of this layer is to classify the weighted features fused by the attention layer. This is achieved by reducing the dimensionality of those features into number of expression categories considered in this study i.e. 7 (anger, disgust, fear, happy, surprise, sad and neutral). Dimensionality reduction is the process of transforming high dimensional data features into lower dimensions which will still retain the rich and essential features of original data. These high dimensional data needs to be reduced as the model's performance might gradually decrease as the number of features increases [34] . Thus, our proposed model utilizes Linear Discriminant Analysis (LDA) as it achieves both dimensionality reduction and classification to optimize the distinction between various classes (7 classes) within the dataset under study i.e. FER2013 and CK+ datasets using linear combination of features. The dimensionality reduction using LDA is as follows:

Step 1: In original space, for different kind of facial images calculate average sample values where total number is denoted as $C.x_{ij}$ represents the j^{th} objects of the i^{th} class sample

$$S_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij} \cdot x_{ij} \in R^d, \text{ where } i = 1, 2, 3 \ldots .C. \quad (8)$$

$$m = \sum_{i=1}^{C} p^i m^i \quad (9)$$

Step 2: For each class, calculate covariance matrix

$$C_i = \frac{1}{N_i}\sum_{j=1}^{N_i} (x_{ij} - m_i) \cdot (x_{ij} - m_i)^T \quad (10)$$

Step 3: Calculate scatter matrices within and between classes

$$S_B = \sum_{i=1}^{C} p_i \left(m_{i-} m\right) \cdot \left(m_{i-} m\right)^T \quad (11)$$

$$S_W = \sum_{i=1}^{c} C_i \quad (12)$$

Step 4: To get projection vectors, compute Eigen vector of matrix $S_W^{-1} S_B$ to obtain reduced data.

With these rich features, attention map is created as a final output of this module which is fed as an input to the DenseNet for reconstruction purposes. With DenseNet, the reconstruction module producesnew enhanced feature maps with narrow layersand reduced redundancy as an output with an activation function $x_l = H_l\left([x_0, x_1, \ldots\ldots, x_{l-1}]\right)$. Unlike ResNet, DenseNet concatenates the incoming feature maps with the output feature maps instead of summing up them. These new feature maps are forwarded to classification module i.e. to fully connected layers with softmax that performs final classification of facial expression on images. Batch normalization is employed after each layer to overcome over fitting issue and to speed up the convergence of network. Softmax activation function is effective and it transforms any integer or fraction values and transforms them between 0 and 1. Full softmax variant is used in this paper since the study deal with multiple classes (7 classes).

IV. Experimentation Results

For this research study, we have considered two commonly used datasets, namely, FER2013 and Cohn Kanade + which contains images with seven facial expressions. Both datasets are described below and the results based on training and simulations implementation are presented in this section along with expression recognition rates and accuracy. The effectiveness of LSTM parameters is shown and also the impact of each module of the proposed model and its effectiveness are also presented. The results of proposed model are compared with few existing models in terms of accuracy in detecting facial expression of images in FER2013 and CK+ datasets. Matlab2021a was used to implement the proposed approach with windows 10 operating system, Intel i7 processer with 6GB RAM.

A. FER 2013 Dataset

A well-known data science competition platform kaggle created this dataset by searching on Google search engine with image keywords and this dataset consists of 35,887 gray-scale images with the resolution of 48*48 pixels. Though, the nature of the dataset is rich and diverse since all the images are obtained from Internet, the dataset images contains a lot of noise including occlusion, different poses and unclear images. These properties of images in FER2013 imposes lot of challenges for the researchers while recognizing expressions and classifying them [35].

B. Cohn Kanade + Dataset

Initially, the extended CK+ dataset was introduced in 2010 by Patrick Lucey team and the Zara Ambadar team [36] and the dataset consists of 123 subject's facial image and

the expressions were recorded as per requirements. Out of 593 images in CK+, 327 images display 7 different facial emotions. Since the quantity of the dataset is less, this study has considered that 327 images which express 7 emotions. As a first step, invalid background is trimmed on those 327 images and to make the datasets similar (both FER2013 and CK+) the resolution is kept as same as FER2013 dataset. The images are rotated and flipped, their brightness and saturation are adjusted when required.

C. Evaluating Module's Effectiveness

As mentioned in Section 3, the proposed model has four modules. In order to prove the effectiveness of each module and show the performance when all these modules work together, we have kept classification module as it is, and removed other modules i.e. one at each time. In such scenario, the recognition rate is measured and given in Table I.

TABLE I. RECOGNITION RATE COMPARISONS OF MODULES IMPACT ON FER

Architecture	Recognition rate (%) FER 2013	Recognition rate (%) CK+
No feature extraction module	59.12	72.28
No attention Module	71.42	75.31
No Reconstruction module	73.87	79.46
Complete Network	79.56	98.92

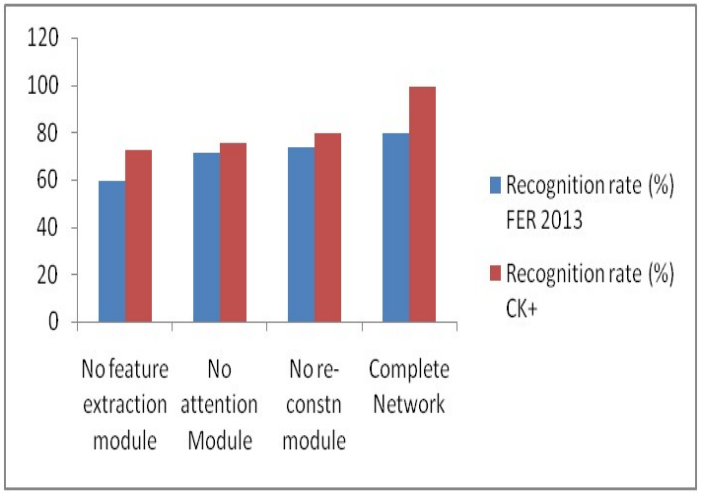

Fig. 4. Recognition rate comparisons of modules impact for FER2013 and CK+ dataset.

From Table I, it is evident that in the absence of feature abstraction layer, the recognition rate for FER2013 and CK+ stands at 59.12% and 72.28respectively. When attention module is removed from the model, the recognition rate for FER2013 and CK+ stands at 71.42% and 75.31% respectively. In case of reconstruction module removal, the recognition rate for FER2013 and CK+ stands at 73.87% and 79.46% respectively. When all these modules are combined together and work as a single network model, the recognition rate goes up for both FER2013 and CK+ datasets at 79.56% and 98.92% respectively. Fig. 4 represents the impact of each modules recognition rate on facial expression images in FER2013 and CK+ datasets respectively.

D. Performance Comparisons

The proposed hybrid model of CNN-BiLSTM aimed to combine the advantages of both the networks along with the attention mechanism to extract both frontal and discriminative features of given input images with improved classification accuracy and recognition rate as well. The results of proposed approach were compared with other hybrid network models such as CNN-LSTM, DSCNN-LSTM, CNN-BiLSTM and ACNN-LSTM network models.

Accuracy is defined as the ratio of true outcomes including both true positives and true negatives to the total number of cases examined.

$$\text{Accuracy} = \frac{TP + TN}{\text{Total population}} \tag{13}$$

F1 score can be divided into two ways based on temporal and spatial features, generally. Event based and frame based and its respective equations are given below where Represents recall and P represents precision, EP-event based precision, ER-event based recall. Total F1 Score can be computed as follows,

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \tag{14}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

TABLE II. ACCURACY,PRECISION, RECALL AND F1 SCORE FOR FER 2013 DATASET

Methods	Accuracy (%)	Precision (%)	F1 Score (%)	Recall (%)
CNN-LSTM	92.23	92.70	92.48	91.89
DSCNN-LSTM	92.48	93.02	93.24	93.24
CNN-BiLSTM	93.14	93.18	92.98	93.21
ACNN-LSTM	91.43	91.12	91.82	91.04
Proposed Method	95.12	94.68	94.87	95.01

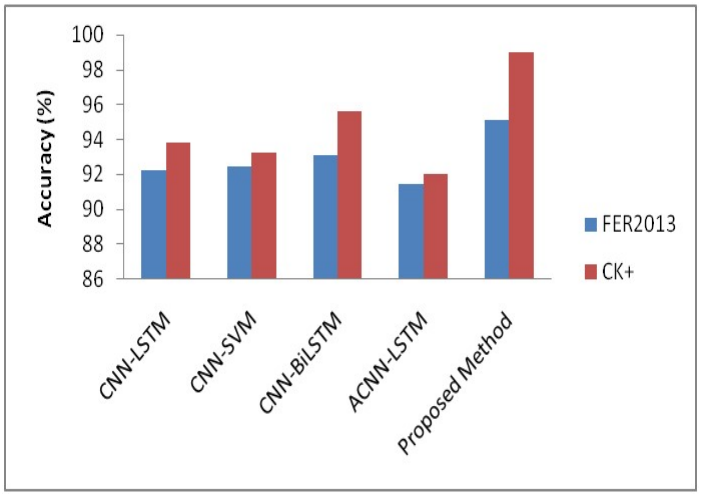

Fig. 5. Performance on accuracy of all methods.

Tables II and III presents different metrics values for various network models and proposed approach for both FER2013 and CK+ datasets. From Table III, it is understood that our proposed approach produced better results than other traditional benchmarking approaches such as CNN-LSTM, DSCNN-LSTM, CNN-BiLSTM and ACNN-LSTM in terms of metrics like accuracy, precision, F1 score and recall. The proposed network model achieved highest accuracy in classifying the facial images on FER2013 dataset with 95.12% accuracy which is 3.1% higher than CNN-LSTM, 2.7% higher than DSCNN-LSTM, 2% higher than CNN-BiLSTM and 3.7% higher than ACNN-LSTM network models in facial expression recognition. With CK+ dataset the proposed model has achieved 98.98% of accuracy in classifying the images which is 5.1% higher than CNN-LSTM, 5.7% higher than DSCNN-LSTM, 3.3% higher than CNN-BiLSTM and 6.9% higher than ACNN-LSTM network models in facial expression recognition. Fig. 5 represents accuracy comparisons for FER2013 and CK+ dataset.

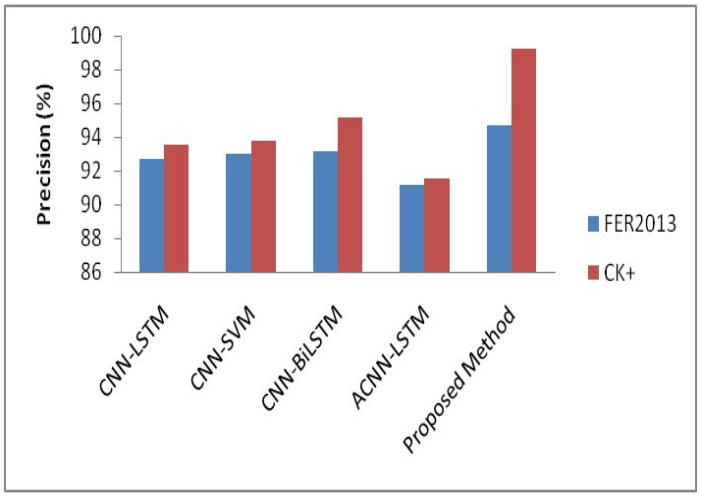

Fig. 6. Performance on precision of all methods.

In terms of Precision, the proposed network model achieved highest percentage of 94.68% on FER2013 dataset which is 1.9% higher than CNN-LSTM, 1.6% higher than DSCNN-LSTM, 1.5% higher than CNN-BiLSTM and 3.5% higher than ACNN-LSTM network models. With CK+ dataset the proposed model has achieved 99.24% which is 5.7% higher than CNN-LSTM, 5.5% higher than DSCNN-LSTM, 4.1% higher than CNN-BiLSTM and 7.7% higher than ACNN-LSTM network models. Fig. 6 represents the precision comparisons for FER2013 and CK+ dataset.

TABLE III. ACCURACY,PRECISION, RECALL AND F1 SCORE FOR FER 2013 DATASET

Methods	Accuracy (%)	Precision (%)	F1 Score (%)	Recall (%)
CNN-LSTM	93.84	93.52	94.52	94.14
CNN-SVM	93.26	93.74	91.89	91.92
CNN-BiLSTM	95.62	95.16	95.34	95.32
ACNN-LSTM	92.02	91.54	91.24	91.56
Proposed Method	98.98	99.24	99.54	98.78

In terms of F1 Score, the proposed network model achieved highest percentage of 94.87% on FER2013 dataset which is 2.4% higher than CNN-LSTM, 1.6% higher than DSCNN-LSTM, 0.9% higher than CNN-BiLSTM and 3.0% higher than ACNN-LSTM network models. With CK+ dataset the proposed model has achieved 99.54% which is 5% higher than CNN-LSTM, 7.6% higher than DSCNN-LSTM, 4.2% higher than CNN-BiLSTM and 8.3% higher than ACNN-LSTM network models. Fig. 7 represents the F1 Score comparisons for FER2013 and CK+ dataset.

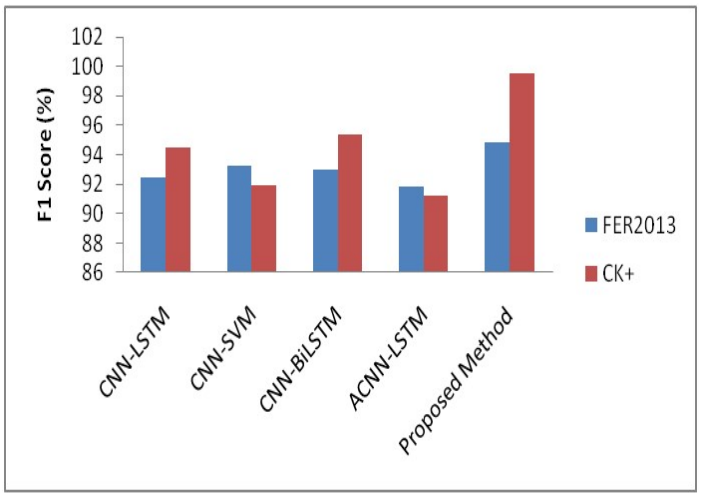

Fig. 7. Performance on F1 score of all methods.

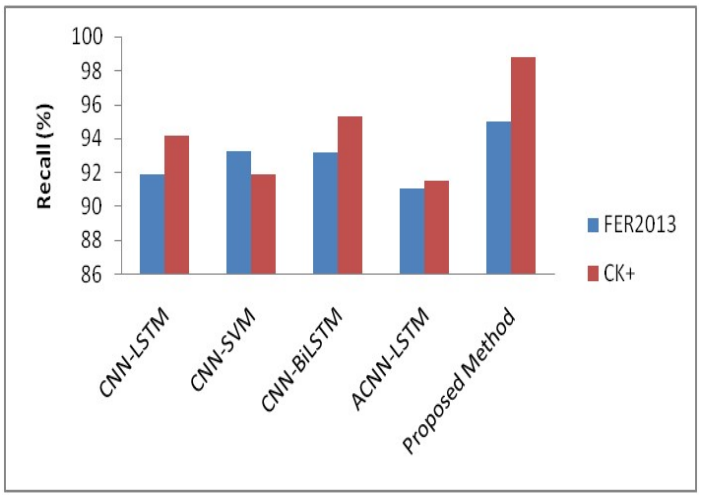

Fig. 8. Performance on recall of all methods.

In terms of recall, the proposed network model achieved highest percentage of 95.01% on FER2013 dataset which is 3.1% higher than CNN-LSTM, 1.8% higher than DSCNN-LSTM, 1.8% higher than CNN-BiLSTM and 4% higher than ACNN-LSTM network models. With CK+ dataset the proposed model has achieved 98.78% which is 4.6% higher than CNN-LSTM, 7.8% higher than DSCNN-LSTM, 3.4% higher than CNN-BiLSTM and 7.2% higher than ACNN-LSTM network models. Fig. 8 represents the recall comparisons for FER2013 and CK+ dataset.

E. Discussion on Findings

Deep Learning based hybrid network models along with CNN has contributed much on classifying Facial Expression Recognition (FER) or emotions of various datasets over a decade.Since the data sets have different images, the results vary according to the image quality. This research work considered two datasets as they are more studied and have images with different scenarios, emotions. Our proposed CNN-BiLSTM based hybrid network model with attention mechanism proved that the accuracy and other matrices of FER can further be improved with right combination of techniques, algorithms and mechanisms. Likewise, we have considered Median Filter for image resizing, GCN for image normalization, CNN for feature extraction, BiLSTM for extracting rich information and discarding unnecessary information, point multiplication attention mechanism for creating attention maps and finally these maps were used to create feature maps that helps in reconstruction. Finally, classification was done using full softmax variant to categorize the emotions of images into seven image expression classes. Our approach was also compared with other benchmarking methods which showed that the proposed network model delivered better results than other models due to the combination of techniques and methods incorporated in or proposed approach.

V. Conclusion and Future Enhancement

This paper has presented a Hybrid CNN-BiLSTM network model with point multiplication attention mechanism for facial expression recognition on dataset images like FER2013 and CK+. Data preprocessing was performed sing Median Filters to resize the image to 128*128 pixels through either enlarging or reducing the original image, followed by image normalization was done using Global Contrast Normalization (GCN). The output obtained from CNN model is forwarded to Bidirectional LSTM where the sequential features of images were extracted using forward and backward layers, and the output is forwarded to point multiplication Attention Mechanism (AM) module. Dimensionality reduction was applied using LDA to the attention map created by the AM to obtain enhanced feature map which can be used in reconstruction module with Full softmax variant to classify the facial expression of images into seven classes. The results were evaluated with other existing network models such as CNN-LSTM, DSCNN-LSTM, CNN-BiLSTM and ACNN-LSTM. The proposed approach has outperformed other models in terms of accuracy, precision, recall and F1 score matrices. For our future work, we can consider some more emotions, datasets and also enhance attention mechanism with additional techniques to improve network model performance further. To achieve better emotion detection of a driver is vital goal to prevent accidents with improved attention mechanism techniques along with deep learning algorithms is our future work.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request Not Applicable

Conflicts of Interest

The authors declare no conflicts of interest

Funding Statement

This study did not receive any funding in any form

Authorship Contribution Statement

Samanthisvaran Jayaraman Writing-Original draft preparation, Conceptualization and Anand Mahendran done Supervision

References

[1] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131 988–132 001, 2020.

[2] S. Dasari and R. Kaluri, "An effective classification of ddos attacks in a distributed network by adopting hierarchical machine learning and hyperparameters optimization techniques," *IEEE Access*, 2024.

[3] M. B. Begum, N. Deepa, M. Uddin, R. Kaluri, M. Abdelhaq, and R. Alsaqour, "An efficient and secure compression technique for data protection using burrows-wheeler transform algorithm," *Heliyon*, vol. 9, no. 6, 2023.

[4] M. Maithri, U. Raghavendra, A. Gudigar, J. Samanth, P. D. Barua, M. Murugappan, Y. Chakole, and U. R. Acharya, "Automated emotion recognition: Current trends and future perspectives," *Computer methods and programs in biomedicine*, vol. 215, p. 106646, 2022.

[5] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[6] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski, "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Information Sciences*, vol. 582, pp. 593–617, 2022.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] M. A. Adil, "Facial emotion detection using convolutional neural networks," 2021.

[9] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018.

[10] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[11] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided cnns," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2281–2293, 2020.

[12] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "Eleatt-rnn: Adding attentiveness to neurons in recurrent neural networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1061–1073, 2019.

[13] S. Yan, *Visual attention mechanism in deep learning and its applications*. The University of Liverpool (United Kingdom), 2018.

[14] J. Daihong, D. Lei, and P. Jin, "Facial expression recognition based on attention mechanism," *Scientific Programming*, vol. 2021, pp. 1–10, 2021.

[15] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," *Advances in neural information processing systems*, vol. 27, 2014.

[16] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.

[17] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 558–565.

[18] P. D. Marrero Fernandez, F. A. Guerrero Pena, T. Ren, and A. Cunha, "Feratt: Facial expression recognition with attention net," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[19] Y. Gan, J. Chen, Z. Yang, and L. Xu, "Multiple attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 7383–7393, 2020.

[20] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, no. 2, p. 199, 2023.

[21] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based cnn for facial expression recognition," *Neurocomputing*, vol. 411, pp. 340–350, 2020.

[22] J. Liao, Y. Lin, T. Ma, S. He, X. Liu, and G. He, "Facial expression recognition methods in the wild based on fusion feature of attention mechanism and lbp," *Sensors*, vol. 23, no. 9, p. 4204, 2023.

[23] K. Prabhu, S. S. Kumar, M. Sivachitra, S. Dineshkumar, and P. Sathiyabama, "Facial expression recognition using enhanced convolution neural network with attention mechanism." *Computer Systems Science & Engineering*, vol. 41, no. 1, 2022.

[24] K. Krishnaveni *et al.*, "A novel framework using binary attention mechanism based deep convolution neural network for face emotion recognition," *Measurement: Sensors*, vol. 30, p. 100881, 2023.

[25] K. Hu, G. Huang, Y. Yang, C.-M. Pun, W.-K. Ling, and L. Cheng, "Rapid facial expression recognition under part occlusion based on symmetric surf and heterogeneous soft partition network," *Multimedia Tools and Applications*, vol. 79, pp. 30 861–30 881, 2020.

[26] P. Arunkumar and S. Kannimuthu, "Mining big data streams using business analytics tools: a bird's eye view on moa and samoa," *International Journal of Business Intelligence and Data Mining*, vol. 17, no. 2, pp. 226–236, 2020.

[27] S. Kannimuthu, K. Bhuvaneshwari, D. Bhanu, A. Vaishnavi, and S. Ahalya, "Performance evaluation of machine learning algorithms for dengue disease prediction," *Journal of Computational and Theoretical Nanoscience*, vol. 16, no. 12, pp. 5105–5110, 2019.

[28] P. A. S. Kannimuthu, "Machine learning based automated driver-behavior prediction for automotive control systems," *Journal of Mechanics of Continua and Mathematical Sciences*, vol. 7, pp. 1–12, 2020.

[29] Y. Ming, H. Qian, L. Guangyuan *et al.*, "Cnn-lstm facial expression recognition method fused with two-layer attention mechanism," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[30] R. Febrian, B. M. Halim, M. Christina, D. Ramdhan, and A. Chowanda, "Facial expression recognition using bidirectional lstm-cnn," *Procedia Computer Science*, vol. 216, pp. 39–47, 2023.

[31] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14*. Springer, 2016, pp. 499–515.

[32] A. D. White, "Deep learning for molecules and materials," *Living journal of computational molecular science*, vol. 3, no. 1, 2022.

[33] Ł. Maziarka, T. Danel, S. Mucha, K. Rataj, J. Tabor, and S. Jastrzebski, "Molecule attention transformer," *arXiv preprint arXiv:2002.08264*, 2020.

[34] Y. Lu, S. Wang, and W. Zhao, "Facial expression recognition based on discrete separable shearlet transform and feature selection," *Algorithms*, vol. 12, no. 1, p. 11, 2018.

[35] H.-D. Nguyen, S. Yeom, G.-S. Lee, H.-J. Yang, I.-S. Na, and S.-H. Kim, "Facial emotion recognition using an ensemble of multi-level convolutional neural networks," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 11, p. 1940015, 2019.

[36] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.