Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-020-04811-0

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

ML-DCNNet: Multi-level Deep Convolutional Neural Network for Facial Expression Recognition and Intensity Estimation

Muhammad Aamir[1] · Tariq Ali[2] · Ahmad Shaf[1] · Muhammad Irfan[2] · Muhammad Qaiser Saleem[3]

Received: 14 April 2020 / Accepted: 17 July 2020

Abstract
The human face has a great accumulation and a diversity of facial expressions. It explores the feelings of a person and can be used to judge the emotional intents of the person to a certain level. By using facial detection and recognition systems, varieties of applications are working in computer vision, surveillance system, security, authentication, or verification of a person and home automation system based on digital image processing with the help of the Internet of Things. The state of the art in these applications is to detect expressions with their intensity level. It is an attention-grabbing problem due to the complex nature of facial features, which is associated with emotions. For that purpose, it is essential to develop an innovative deep learning model to detect and estimate the facial expression intensity level. To do this, a multi-level deep convolutional neural network is proposed to recognize facial expression and their intensity level. At the first level, Expression-Net classifies face expressions, and at the second level, Intensity-Net estimates the intensity of the facial expression. Evaluation of the proposed model for facial expression recognition and intensity estimation is carried out by using the extended Cohn–Kanade and Japanese Female Facial Expression datasets. The proposed method shows an outstanding performance in terms of accuracy of 98.8% and 97.7% for both the datasets as compared to state-of-the-art techniques.

Keywords CNN · Facial expressions recognition · Facial expression intensity estimation ML-DCNNet · Deep learning · Computer vision

1 Introduction

At present, the communication which occurs only in one direction is considered as an active one. However, the computer may listen to human dialogues through the usage of exclusive audio and speech recognition equipment. Letting computers to recognize humanoid emotions through visualization will connect the gap between computer and human communication, which will make the computer very active, smart, and user-friendly. The facial expressions related to humans are a reliable source of nonverbal communication. To analyze and recognize expressions of the face using computers so that they will understand whether users feel some type of emotions like happy, bored, etc., or not. Many applications are working based on digital image processing with the help of the Internet of Things (IoT) like home automation system, monitoring of different industrial parameters, detection and positioning of different objects, and security systems [1]. It would be phenomenal experimentation and a real-world development in the computer-vision system

✉ Tariq Ali
tariqhsp@gmail.com

Muhammad Aamir
muhammadaamir@cuisahiwal.edu.pk

Ahmad Shaf
ahmadshaf@cuisahiwal.edu.pk

Muhammad Irfan
irfan16.uetian@gmail.com

Muhammad Qaiser Saleem
muhammad.qaiser.saleem@gmail.com

[1] Computer Science Department, COMSATS University Islamabad, Sahiwal Campus, Sahiwal, Pakistan

[2] College of Engineering, Electrical Engineering Department, Najran University, Najran 61441, Kingdom of Saudi Arabia

[3] College of Computer Science and Information Technology, Al Baha University, Al Baha, Kingdom of Saudi Arabia

Published online: 31 July 2020

which can spontaneously identify a variety of humanoid face expressions and intensity estimation.

The intensity factor of facial expressions has a significant aspect for providing a sense of facial expressions, although all facial expression recognition focuses on six basic emotions, namely happiness, anger, fear, sadness, surprise, and disgust [2]. Even though these facial expressions categories are sufficient for facial expression recognition, it makes it difficult to tell how strong the expressions are. Facial expressions recognition with the intensity estimation is currently a growing field of research with the development of new technologies and computing applications. It begins in the mid-nineteenth century when the first algorithm was developed by Kanade [3] in which symmetrical structures were implemented for the facial recognition of a person.

From an era, the researchers are working to understand these expressions and their intensity level under the umbrella of computer vision, which is known as machine learning. It initially parses the given data, secondly learns from that data, and finally makes a decision to do some predictions or classification. Further, the method of learning is classified into three types: supervised, unsupervised, and reinforcement learning. These algorithms consider only the simplified data to make predictions but do not handle the complex structures of data, e.g., audio, video, and images. The deep learning techniques are introduced to address these complex situations, which is a subcategory of machine learning [4].

To cope with the raised problem, many deep learning techniques have been proposed for facial expression recognition and intensity estimation. One of them is the convolutional neural network (CNN). Currently, CNN is used in several fields like virtual reality authentication, security, psychology, medical, and human–computer interaction (HCI) [5, 6]. CNN consists of different layers. The first convolutional layer (convent layer) is used for facial feature extraction, which is the main building block of CNN [7]. This layer consists of different filters to extract facial expressions. After every convolutional layer, a nonlinearity layer is placed. This layer handles the nonlinear values of the convent and also for input. In CNN, there are different pooling operations, but the most popular one is max pooling, in which a layer is applied between the successive convolutional layers of CNN. This layer controls the number of parameters and overfitting in the network. The flattening layer prepares data for the classical neural network. This layer uses data passed from the pooling layer or convolutional layer and packed in the matrixes into arrays. After that, these values are an input to the neural network. The final layer is fully connected, and this layer does the actual classification. This layer takes input from the flattening process and feeds and forwards it through the neural network. The working architecture of CNN is shown in Fig. 1.

In the data preprocessing phase, CNN used only the class-wise label data which does not have an intensity label for the learning process of the classifier. At that time, the facial expression intensity is calculated directly as discussed in regional volumetric difference [8], principal component analysis, 2D eigenspace, and visual movement to find the gesture of the face (reference). These approaches performed well at their predefined limits (detect facial expressions), but they could not succeed to determine the intensity level of facial expressions [9].

To do this, a multi-level deep convolutional neural network (ML-DCNNet) is proposed that works in different layers as shown in Fig. 2. ML-DCNNet works in two levels. The first level is Expression-Net in which CNN classifies the facial expressions, and the second level is Intensity-Net, which estimates the intensity of the expressions recognized by the Expression-Net. The intensity level of facial expressions is divided into three phases: onset, offset, and apex, where onset describes the beginning of the expression, offset belongs to medium value and apex defines the peak value of the intensity of expressions. Its architecture consists of different layers like the image input layer, the convolutional layer, the max polling layer, the relu layer, the batch normalization layer, fully connected layer (FC), softmax layer, and classification layer.

2 Related Work

With the advancement in tools and technologies, various solutions have been proposed to extract and classify the facial features by using different facial expression techniques. A novel framework has been used to extract facial features and to predict the intensity of facial expressions (FE) [10–12]. It uses the two methods for classification; one is voted weight scheme and hidden Markov model and a change point detector method to analyze the FE and their intensity estimation based on extended Cohn–Kanade (CK+) and (BHU) datasets. Another modified three-dimensional deep CNN was proposed by Hasani et al. [2]. This network consists of the Inception-ResNet layer with three dimensions which tells about spatial and temporal relationships between facial images. It uses facial landmarks points as an input which makes facial components as a significant indication for facial expression recognition. The first "Facial Expression Recognition and Analysis (FERA)" competition was made in 2011 in which some standards were made for facial expression analysis. The second FERA challenge was made in 2015 in which a system having automatic features to analyze FE and intensities estimation was proposed in [13]. The third FERA 2017 challenge extends FERA 2015 which considered different camera views for auto-analysis of FE for intensities estimation [14]. Lee and Xu [15] focused

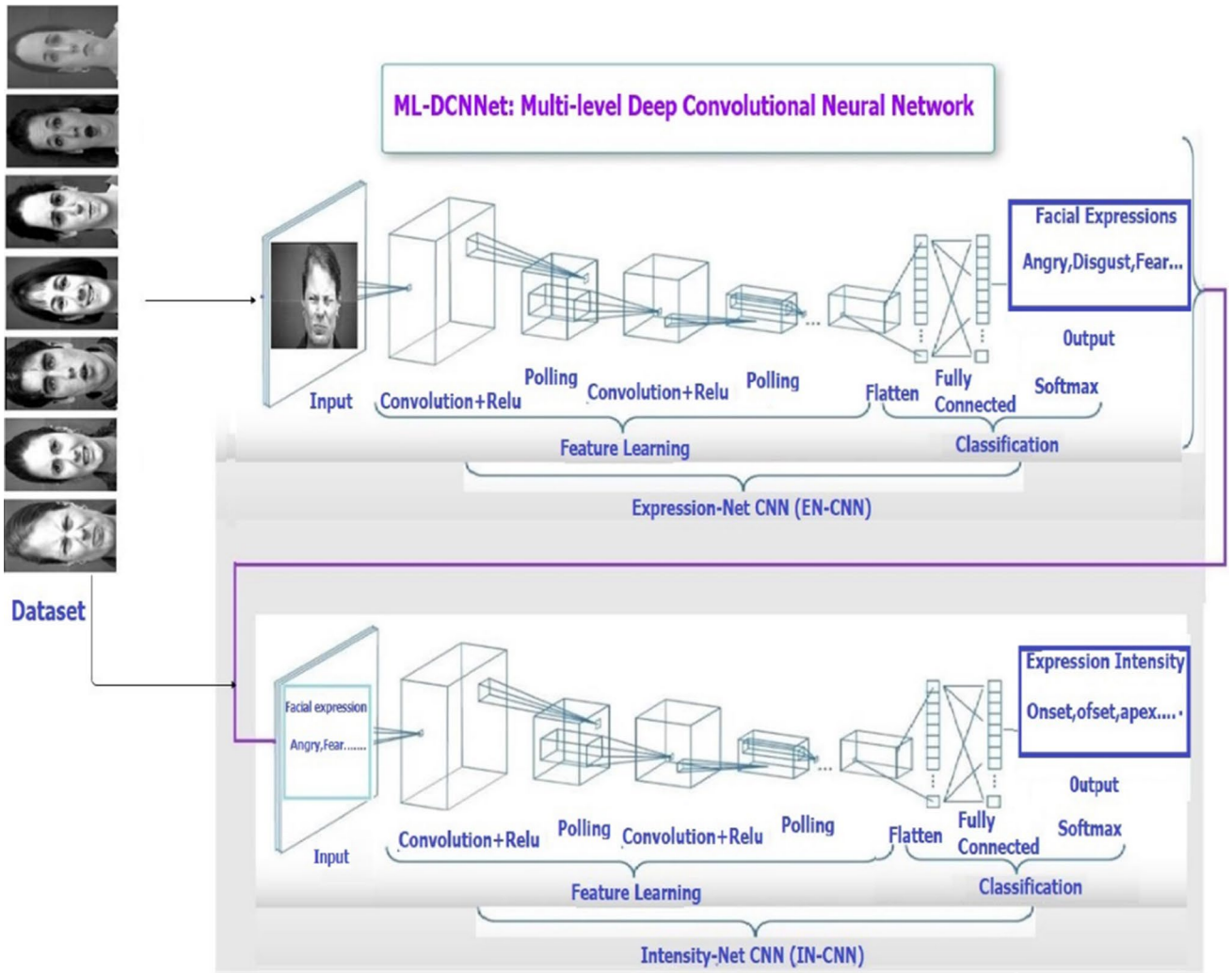

Fig. 1 Multi-level deep convolutional architecture

on a "real-time estimation of facial expression intensity." It uses isometric features mapping, cascading neural networks, and SVM. It evaluates the facial expression intensities of expressions like happy, angry, and sadness in "real time." Quan et al. [16] proposed a model name "K-order emotional intensity model (K-EIM)." It uses SVM and k-means clustering to evaluate the results on Cohn–Kanade (CK) dataset. A framework on a method to estimate the isolated rank system for the intensity of expressions based on a single image was proposed by Chang et al. [17]. This framework uses the scattering transform technique, support vector machine, Gabor wavelet, and appearance model (AAM) to evaluate the results on extended Cohn–Kanade (CK+) dataset.

Computer vision-based techniques have been developed day by day providing capabilities to recognize facial expressions and intensity estimation. We will briefly analyze the strong points and weaknesses of certain main models. A geometric wireframe of the face model of three dimensions has been used by Aizawa, Choi, Essa, Terzopoulos for facial expression recognition, synthesis and used two ways to identify FE. The first technique is to classify the five emotions (smiling face, surprise expression, eyebrow raised, angry, and disgust) using counting 36 highest muscle actuations based on the dot-product similarity in association with the standardized training section, but all other temporal segments are ignored. The average recognition accuracy rate was 97.8% with six different subjects having different image sequences for training and testing purposes. The second technique uses the temporal segment mapping for 2D gray-scale image sequences. The time changing is a significant deliberation that increases the rate for recognition accuracy since the temporal segment mapping calculates the association among testing and standardized segments of image sequences. As compared to the use of the complicated 3D symmetrical approaches, Himer and Kaiser implemented a method for automatic recognition of facial movements by

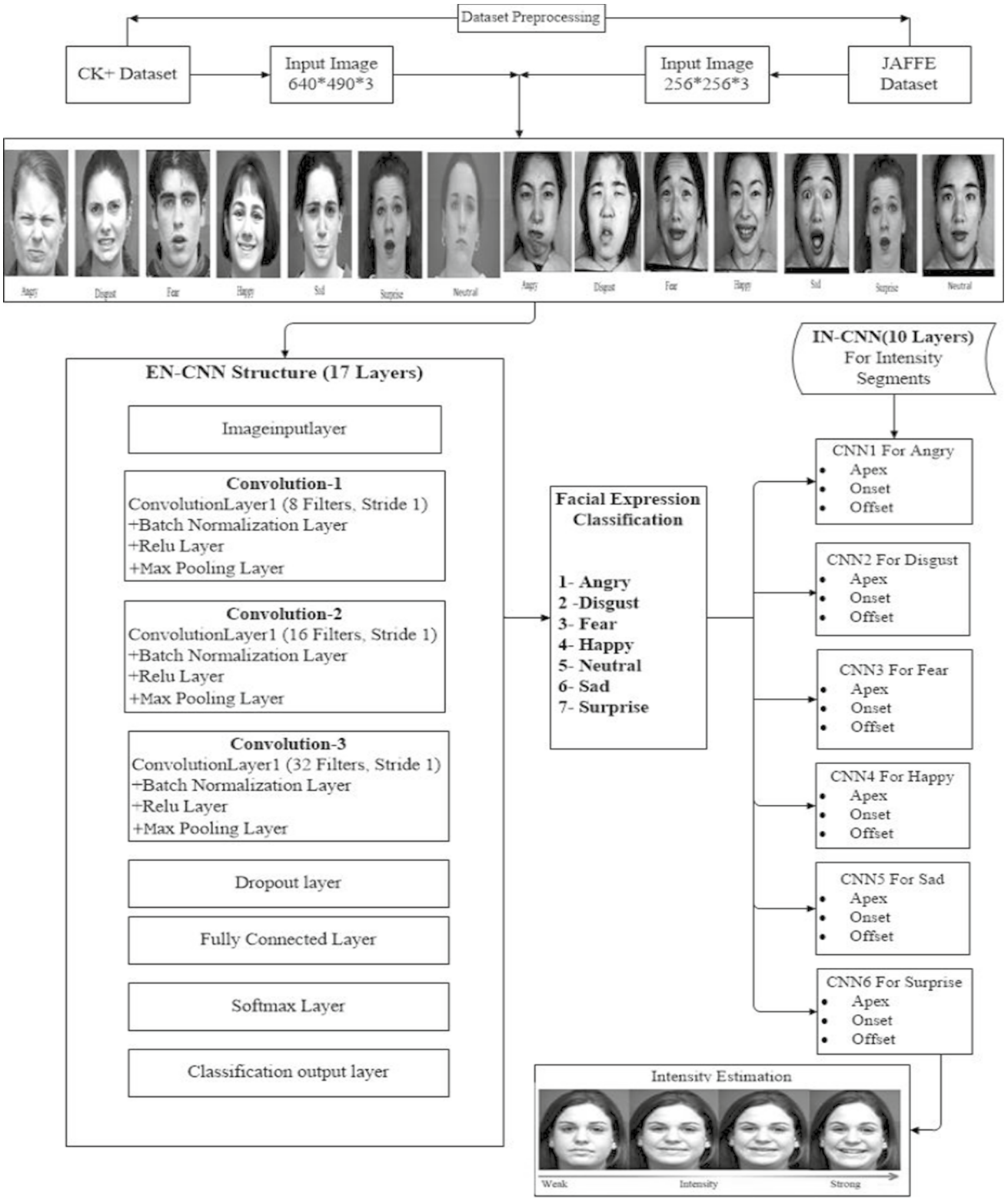

Fig. 2 Working process representation of ML-DCNNet

tracing the locations of involved points on the facial surface to give the impression in a series of images. Meanwhile, the structure of points could be distorted owing to skin tissue movements during FE. It is challenging to find correctly the consistent dominant locations for distorted points and therefore disturb the tracing precision. In research work by Mase [18], clues of humanoid skin tissues were calculated instead of counting the facial expressions.

The facial muscle sections were labeled as most important by manually marking important regions. The visual movement was figured to mine between 12 and 44 facial features, in amalgamation with feature positions that were understood as suitable AUs. Mase's methodology depends comprehensively on correctly finding the physically nominated muscles section; movement instructions within each separate area is meant to display the movement path of the selected section. On the other hand, once the particular part matches the flat, dull surface of the face, the visual movement prediction would be variable, important to trace the fault. Particular muscles section might be challenging to find physically as they are negligible and extremely moveable. The work in [14, 19] used this moderate-level demonstration to categorize the six basic FE as well as eyes blinking. The degree of recognition was 88 percent among 32 subjects with 46 image series. There is an extension of work by Yacoob and Davis's [19] based on the parallel mid-level demonstration to recognize facial expressions for smile and surprise expressions via an ANN with "radial basis function (RBF)." The recognition degree attained was 88 percent with 32 subjects. In his work, Yacoob implemented a native parameterized technique of the image gesture to distinguish and identify the non-rigid to rigid head motion. The defined complex recognition methodology was related to that of Yacoob and Davis's method which is based on the mid-level index of the gesture track of every facial feature area. The moderate-level illustration was forecast, still by taking the variation of the gesture restriction valuation and threshold importance. The sign restrictions thresholding would find out some delicate gestures. Moreover, various limitations were used for diverse motion constraints in the research. These methods set some threshold for sign constraints, as a result, compact consistency and accuracy of the recognition. In the work of recognition of six basic FE, the mean rate of recognition was 92% with 40 subjects in 70 image series.

The dissertation implemented the three groups of ANN to identify six elementary FE, miscellaneous FE, and the intensity for each facial expression, respectively. Further work in Japan has implemented methodologies the same as of Kobayashi and Hara depending upon the dislocation of physically nominated facial feature facts. Others applied fuzzy logic [20], Ralescu and Hartani [21], and chaos joint with ANN to identify six elementary FE. Further in the work, Ekman implemented three methods to mine information on higher facial expressions and used ANN for recognition. To deal with the interval deforming challenge, researchers anticipated to physically preference six image sequences from every image series to form a new order for further dealing out: unbiased expression for the first sequence, low-value-scale expressions for the second sequence, average degree for the third and fourth sequence, and peak-level expressions for the remaining last two sequences.

Because of the comparative symmetrical communication of facial image sequences, to acquire the attention of cyclic and plane ranking based upon both the eyes, that was inadequate for bringing into line the facial images correctly since the perpendicular ranking was misplaced with the magnitudes of humanoid faces possibly will vary between subjects. The data implemented in this research for the PCA are the variances in image sequences gained by subtracting the grayscaled values neutral to other sequence values, and that subtraction procedure is not appropriate to look out of the differences among distinct facial sequences. For higher slope parts recognition, this research does not distinguish between a few wrinkles that may be created by FE, whereas others possibly will be a persistent characteristic of the individual's face. Similarly, they projected to wrinkles along several lines where some things may or may not look lined with the same FE. The best accuracy rate was 91% from different 20 proficient subjects having 400 images.

Additionally, different approaches to detection have been implemented to face exploration. Beymer suggested a way to standardize facial images through various subjects, and this explored and combined facial images by adding structural and quality calculations using visual movement and PCA in grayscale values. Bregler and Konig used PCA and HMM for voice recognition. Individual the innovators in face recognition, used symmetry structures of the humanoid face, the size of face features, space among features, to classify an expression. Samaria and Young [22] improved every 2D fixed and separate face image into a connected one-dimension grayscale base values vectors aimed in constant HMM to recognize a humanoid face. These works are ultimately connected to our proposed work and provide valuable references to our work. The comparison of all the techniques with accuracies is shown in Table 1.

3 Methodology

This section describes the details of the overall methodology for the proposed multi-level convolutional neural network. In the proposed research, a multi-level deep convolutional neural network (ML-DCNNet) is implemented for FER along with intensity estimation of facial expression. It consists of different layers. It is implemented in two parts: the first one

Table 1 Comparison with deep learning techniques

Cited	Methodology	Dataset	Accuracy (%)
[12]	Hidden Markov models and weighted voting scheme	CK and BHU	54.07%
[2]	Deep 3D CNN	CK+, MMI, FERA, DISFA	89.50%, 67.50%, 67.74%, 51.35%
[13]	Fera 2015-baseline method	BP4D, SEMAINE	68.1%,50.2%
[14]	Fera 2017-baseline method	BP4D-Spontaneous	56.1%
[15]	SVM, cascading neural network	CK+	80%
[16]	SVM and k-means clustering	CK	88.32%
[17]	Support vector machine, Gabor wavelet, and appearance model (AAM)	CK+	39.8%,40.9%,67.9%

is Expression-Net CNN (EN-CNN) and the second one is Intensity-Net with different layers as explained below.

3.1 Multi-level Deep CNN (ML-DCNNet) Architecture

The existing techniques for facial expression detections have the minimum number of filters and large time complexity. To deal with this issue, multi-level deep neural network-based image classification has been proposed in this article. It is related to an ordinary neural network having a combination of different neurons, learning rates, and other parameters. CNN consists of different layers. The first convolutional layer (convent layer) is used for facial feature extraction which is the main building block of CNN [7]. This layer consists of different filters to extract facial expressions. After every convolutional layer, a nonlinearity layer is placed. This layer handles the nonlinear values of the convent and also for input. In CNN, there are different pooling operations, but the most popular one is max pooling, in which a layer is applied between the successive convolutional layers of CNN. This layer controls the number of parameters and overfitting in the network. The flattening layer prepares data for the classical neural network. This layer uses data passed from the pooling layer or convolutional layer and packed in the matrixes into arrays. After that, these values are an input to the neural network. The final layer is a fully connected layer, and this layer does the actual classification. This layer takes input from the flattening process and feeds and forwards it through the neural network. In this research for the selection of networks, we used seven different CNN approaches as shown in Fig. 2. All the CNNs implemented in two parts and its architecture are explained in Fig. 1.

3.2 Expression-Net CNN (EN-CNN)

This section presents CNN that is used to recognize the facial expression class. It is known as Expression-Net (EN-CNN) architecture as shown in Fig. 3. In this architecture, we used a CNN having 17 different layers: an image input layer, convolutional, max pooling, and fully connected softmax and classification layers with learning rate 0.001 and epochs size 60 as shown in Table 2. This network, first of all, takes an image as an input in the image input layer and passes through the different CNN layers. The image passes through all the CNN layers to get its features matrix of the image, and at the end of the network classification layer, an image is returned with the corresponding classification label. This classified image is used as an input in the next Intensity-Net network to estimate the intensity level.

3.3 Intensity-Net CNN (IN-CNN)

In this part of ML-DCNNet architecture, six CNNs are used for the intensity estimation of facial expressions. It is named as Intensity-Net CNN (IN-CNN). This network uses classified images from the EN-CNN network to estimate the intensity of the images. The intensity level of facial expression is divided into three phases: onset, offset, and apex, where onset describes the beginning of the expression, offset belongs to medium value and apex defines the peak value of the intensity of expressions. For each facial expression recognition in the Expression-Net phase, we implemented one CNN algorithm, so a total of six CNN architectures are used in this part. The architecture of IN-CNN is shown in Fig. 4. We used ten layers with a learning rate of 0.001 and 30 epoch size as explained in Table 3.

4 Experimental Results and Analysis

This section describes the overall results of the experiments of facial expression recognition, intensity estimations, and their comparison with other existing techniques.

4.1 Results with CK+

The number of epochs is counted based on validation data for the ML-DCNNet. We used Adam optimizer with

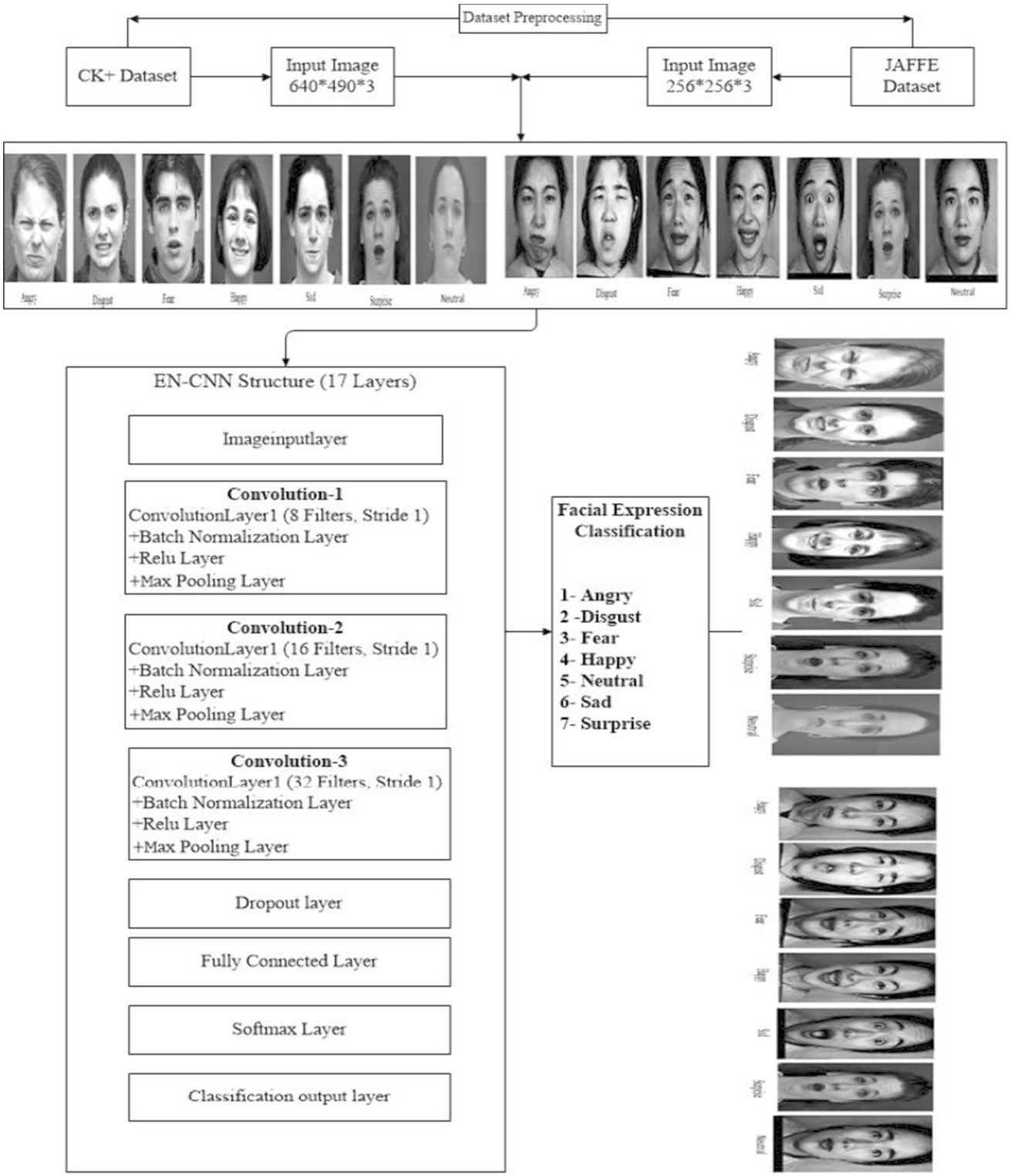

Fig. 3 Layer-wise representation of EN-CNN architecture

learning rate 0.001. After setting all the parameters, we analyze CNN in terms of accuracy for training, testing, and validation dataset. The proposed algorithm was evaluated on extended Cohn–Kanade (CK+) [22] and Japanese Female Facial Expression (JAFFE) [23] datasets. Both the datasets are publically available for research purposes. The description of both the datasets is given in Table 4.

Table 2 Layer-wise description of Expression-Net CNN (EN-CNN)

Sr. no.	Layers description	Parameters
CNN layers description with learning rate = 0.001 and epochs = 60		
1	Image input layer	640 * 490/256 * 256
2	convolution2dLayer	3 * 3, $K = 8$, stride = 1
3	Batch normalization layer	3 * 3, stride = 1
4	Relu layer	3 * 3, stride = 1
5	maxPooling2dLayer	3 * 3, stride = 1
6	convolution2dLayer	3 * 3, $k = 16$, stride = 1
7	Batch normalization layer	3 * 3, stride = 1
8	Relu layer	3 * 3, stride = 1
9	maxPooling2dLayer	3 * 3, stride = 1
10	convolution2dLayer	3 * 3, k = 32, stride = 1
11	Batch normalization layer	3 * 3, stride = 1
12	Relu layer	3 * 3, stride = 1
13	maxPooling2dLayer	3 * 3, stride = 1
14	Dropout layer	Dropout = 0.7
15	Fully connected layer	7 classes
16	Softmax layer	7 classes
17	Classification layer	7 classes

4.2 Dataset Preprocessing

First of all, dataset preprocessing is implemented to get the required structure of data. In this phase, we deal with the image sequences from the various datasets to make a common and suitable format for the training of the network. No modification of lighting was used to the sequence of images as shown in Fig. 5. We decided to resize images with the same size for network training to avoid any ambiguity related to size. The format and dimensions for each image dataset are defined; for CK+ dataset, it mainly contains face frontal poses only with 640 * 490 dimensions and JPG format. A sequence of images is labeled with 1, 2, 3, 4 labels, which defines neutral, onset, offset, and apex levels of intensities, respectively, as shown in Fig. 6. Neutral defines when there is no expression intensity, and that's why in this research, we ignore neutral expression, while onset shows when expression intensity begins to change, and offset demonstrates the lower intensity, and apex reveals the highest level of intensity as shown in Fig. 7. The same preprocessing is applied for the JAFFE dataset with 256 * 256 dimension and JPG format as shown in Fig. 8.

4.3 Training and Testing Processes

To make the training of the ML-DCNNet algorithm, the complete database is divided into three parts: training, testing, and validation groups. The training and testing datasets are divided into 70% and 30% ratio of the total dataset for both datasets in Table 4. The validation set is used to count the number of epochs for a training session. The epochs number with better accuracy for all the categories with the validation set is shown in Fig. 9 (CK+) and Fig. 10 (JAFFE). The overfitting values are commonly known issues in CNN with a limited amount of data. The distribution of train, test, and validation of images into different categories is shown in Table 4 for both datasets.

The overall recognition rates of facial expressions with respect to different parameters on training and testing datasets are explained for CK+ dataset in Table 5. Due to the number of filters and parallel execution of CNNs with different parameters, the proposed method shows accuracies 98.8% and 95.4% for training and test dataset, respectively.

The training and validation accuracy and loss graph which is 98.8% on the different numbers of iterations are shown in Fig. 9. The smooth line shows the training, and the dotted line shows the validation accuracy and loss.

ML-DCNNet shows excellent performance in intensity estimation of intensity segments which is above 90% in each case as shown in Table 6. But on the other hand, the average intensity recognition rate for onset is lower than the other segments, which is due to the reason that onset is the initial state from neutral to apex expression. However, ML-DCNNet shows outstanding performance in this task as the intensity segments recognition, which is an average 99.14%.

4.4 Results with JAFFE

The proposed method was executed in parallel manners both for expression and intensity. This shows overall recognition rates of facial expressions for training 97.7% and testing

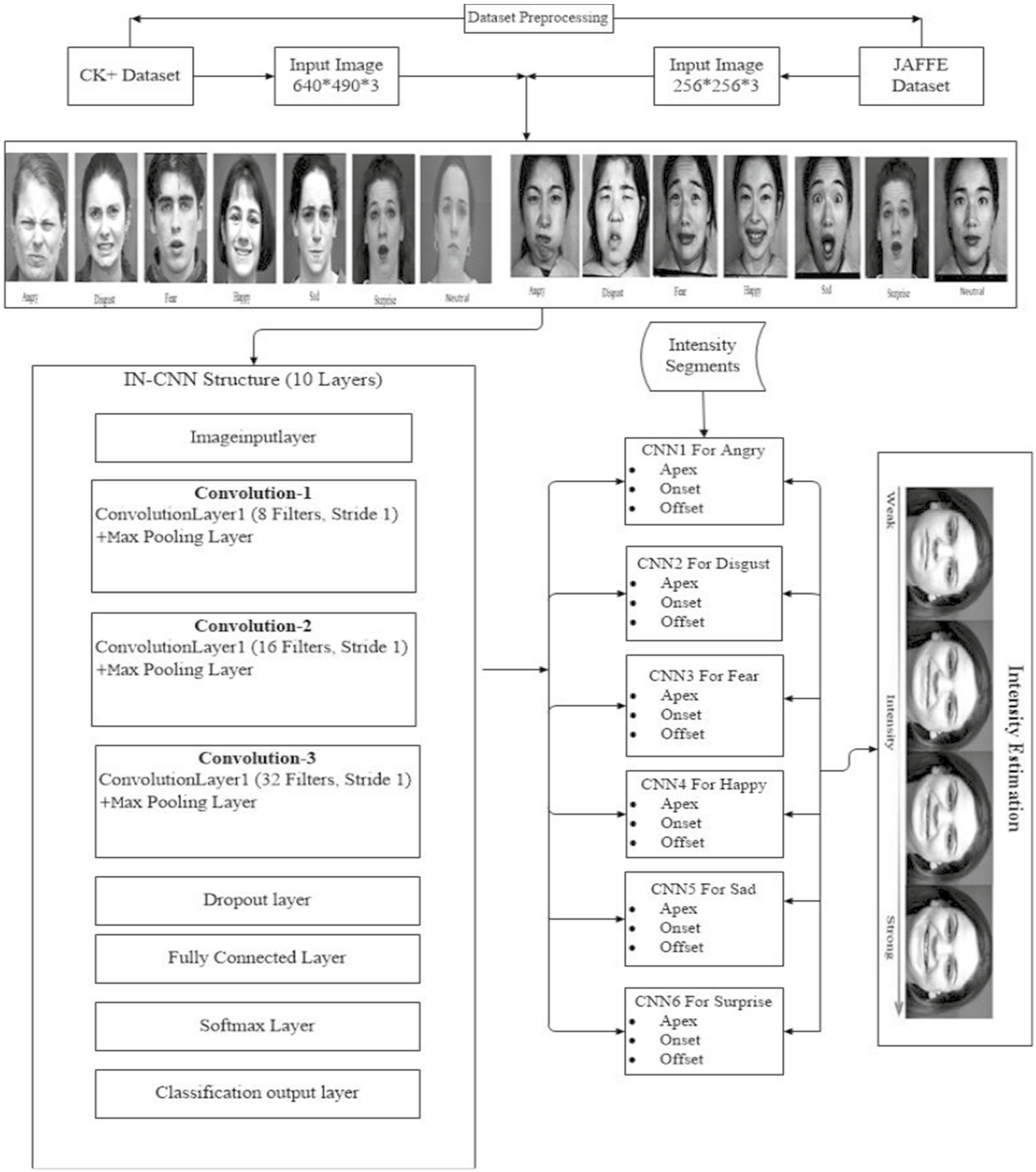

Fig. 4 Layer-wise representation of IN-CNN architecture

95.2% with respect to different parameters on JAFFE dataset and is explained in Table 7.

Figure 10 shows the training and validation accuracy and loss graph, which is 97.7% with learning rate 0.001, no. of iterations and epoch size 60. The smooth line shows the training, and the dotted line shows the validation of the network.

ML-DCNNet shows excellent performance in intensity estimation of intensity segments which is above 90% in each case as shown in Table 8. But on the other hand, the average

Table 3 Layer-wise description of Intensity-Net CNN (IN-CNN)

Sr. no.	Layers description	Parameters
CNN layers description with learning rate = 0.001 and epochs = 30		
1	Image input layer	640 * 490/256 * 256
2	convolution2dLayer	3 * 3, $K = 8$, stride = 1
3	maxPooling2dLayer	3 * 3, stride = 1
4	convolution2dLayer	3 * 3, $k = 16$, stride = 1
5	maxPooling2dLayer	3 * 3, stride = 1
6	convolution2dLayer	3 * 3, $k = 32$, stride = 1
7	maxPooling2dLayer	3 * 3, stride = 1
8	Fully connected layer	3 classes
9	Softmax layer	3 classes
10	Classification layer	3 classes

Table 4 Distribution of images in a dataset

CK+ dataset					JAFFE dataset			
Expression	Total	Training	Test	Validation	Total	Training	Test	Validation
Angry	244	144	36	74	30	30	18	9
Disgust	248	148	37	78	29	29	18	9
Fear	124	100	24	38	32	32	18	10
Happy	184	96	24	55	31	31	18	10
Neutral	138	92	96	38	30	30	18	9
Sad	144	96	24	44	31	31	18	10
Surprise	268	168	42	80	30	30	18	9
	1350	844	283	407	213	213	126	66

Fig. 5 Illustration of six basic and neutral facial expressions from CK+ dataset

Fig. 6 Happy expression intensity from the lowest to highest value

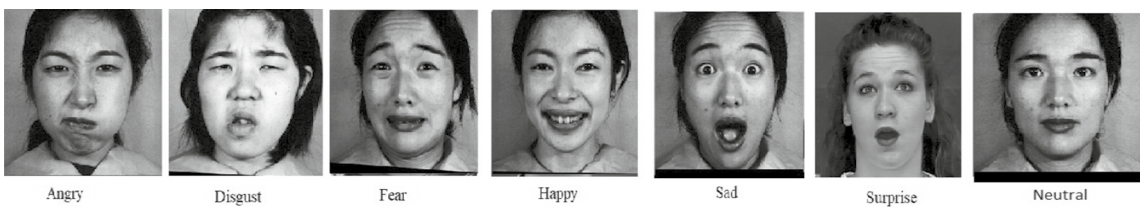

Fig. 7 Illustration of six basic and neutral facial expressions from JAFFE dataset

Fig. 8 Neutral to apex images of fear expression

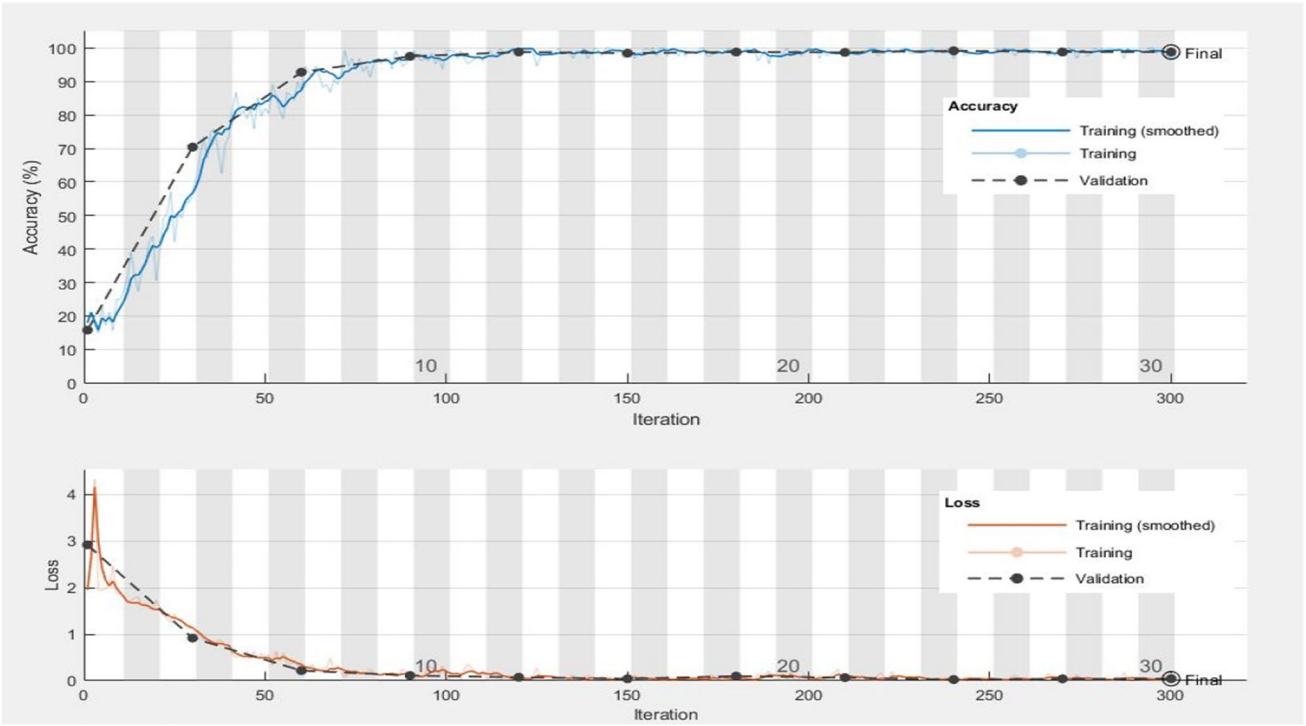

Fig. 9 Graphical representation of training and validation accuracy and loss (CK+) for different numbers of iterations

intensity recognition rate for an apex is lower and the offset is higher. However, the proposed algorithm shows outstanding performance in this task as the intensity segments recognition which is an average 93.43%.

4.5 Resultant Confusion matrix

The confusion matrix in Figs. 11 and 12 shows the ML-DCNNet verified results of the recognition rate of different facial expression classes for CK+ and JAFFE training datasets which show the accuracy 98.8% and 97.7% and for testing datasets 95.4% and 95.2%, respectively. The confusion matrix in Figs. 13 and 14 shows the verified results of intensity segments for all the facial expressions on testing datasets. The rows show the predicted values of the classes, and the columns explained the true class values. The diagonal cells show the total number of observations that are correctly classified. The off-diagonal cells explain the incorrect classification of observations. Each cell includes the total number of observations and their percentage. The predicted values percentage that is correctly and incorrectly classified for each class is presented in the far-right column of the confusion matrix. These values are also called precision (positive predictive value) and false discovery rate, correspondingly. The correct and incorrect classification percentages of all the classes are explained in the row at the bottom end. These values are frequently called the recall (or true positive rate) and false negative rate, correspondingly. The overall accuracy is shown in the bottom-right cell of the confusion matrix."

4.6 Comparison with Other Deep Learning Techniques

This section provides a comparison of the results of all the experiments with the existing results. It also explains the facial expression recognition results and intensity estimation results of intensity segments for onset, offset, apex for both the datasets CK+ and JAFFE. Table 9 shows the average results of all the facial expressions based on the

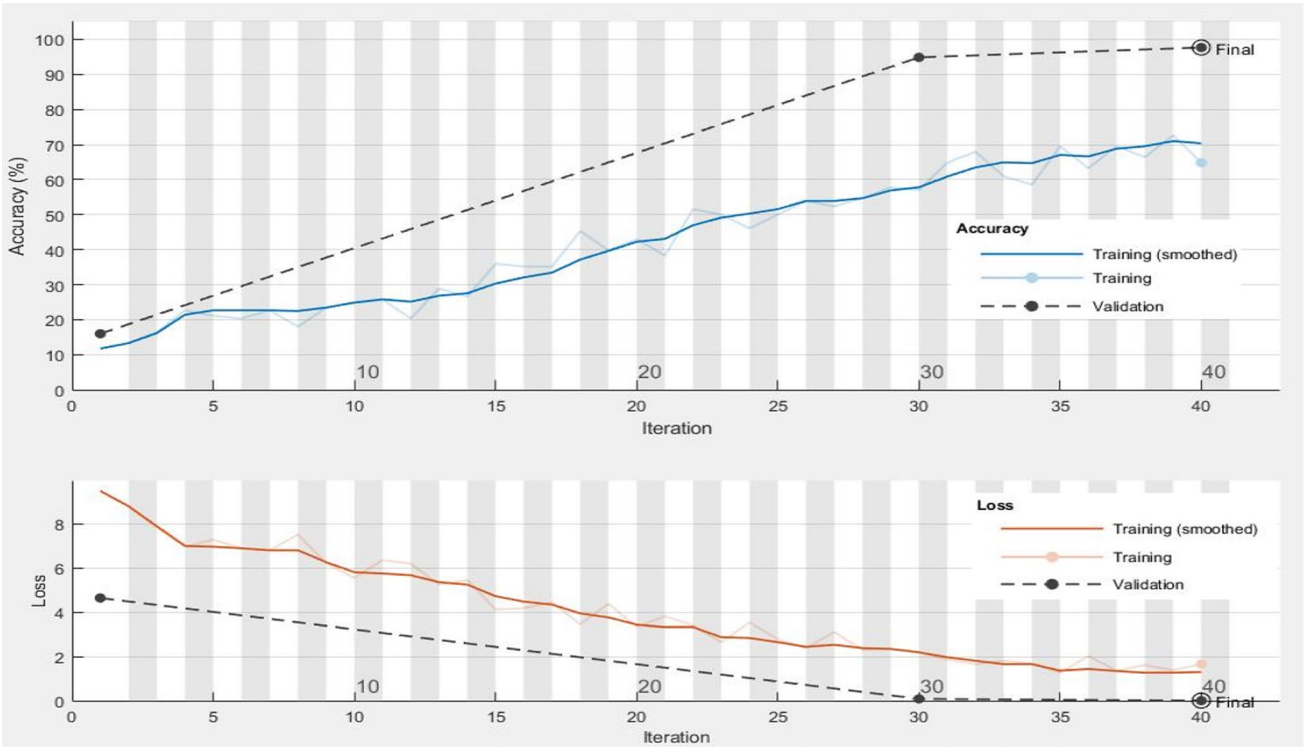

Fig. 10 Graphical representation of training and validation accuracy and loss (JAFFE) for different numbers of iterations

Table 5 Overall accuracy of facial expressions recognition on CK+ dataset

Method	Training accuracy (%)	Test accuracy (%)
ML-DCNNet	98.8	95.4

Table 6 Intensity segment estimation recognition rate (%) on CK+ dataset

Intensity segments	Accuracy (%)
Apex	99.21
Onset	98.88
Offset	99.33
Average	99.14

Table 7 The overall accuracy of facial expressions recognition on JAFFE dataset

Method	Training accuracy (%)	Test accuracy (%)
ML-DCNNet	97.7	95.2

Table 8 Intensity segment estimation recognition rate (%) on JAFFE dataset

Intensity segments	Accuracy
Apex	90.06
Onset	92.45
Offset	97.91
Average	93.43

CK+ dataset. It explains the facial expression recognition rates for each expression class. The ML-DCNNet algorithm uses the multi-level parallel CNN's structure that reduces the complexity in terms of the selection of facial features. It shows the average recognition rate of 98.8%, which is the highest one as compared to the already existing state-of-the-art algorithms recognition rates of facial expressions.

The recognition rates of intensity segments (apex, onset, offset) are also explained in Table 10. This Intensity-Net CNN part of the algorithm shows the highest results of the recognition rate of intensity segments which is 99.14% as compared to the already existing state-of-the-art algorithms.

Table 11 shows the results of all the facial expressions based on the JAFFE dataset. It also explains the facial expression recognition rates for each expression class. It

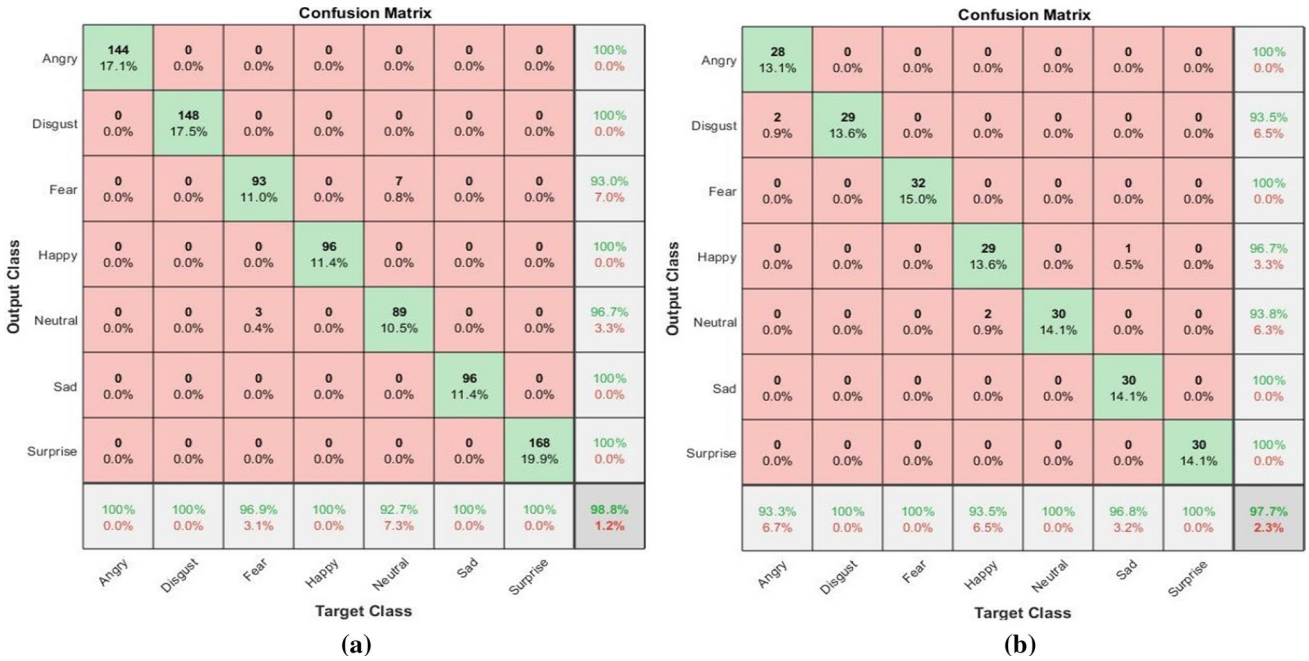

Fig. 11 Confusion matrix of seven classes of facial expressions using ML-DCNNet model: **a** CK+ and **b** JAFFE on training dataset

Fig. 12 Confusion matrix of seven classes of facial expressions using ML-DCNNet model: **a** CK+ and **b** JAFFE on the test dataset

also demonstrates the average of all the recognition rates of all the expressions. The ML-DCNNet algorithm uses the multi-level parallel CNN's structure that reduces the complexity in terms of the selection of facial features. It shows the average recognition rate of 97.7%, which is the highest one as compared to the already existing state-of-the-art algorithms recognition rates of facial expressions.

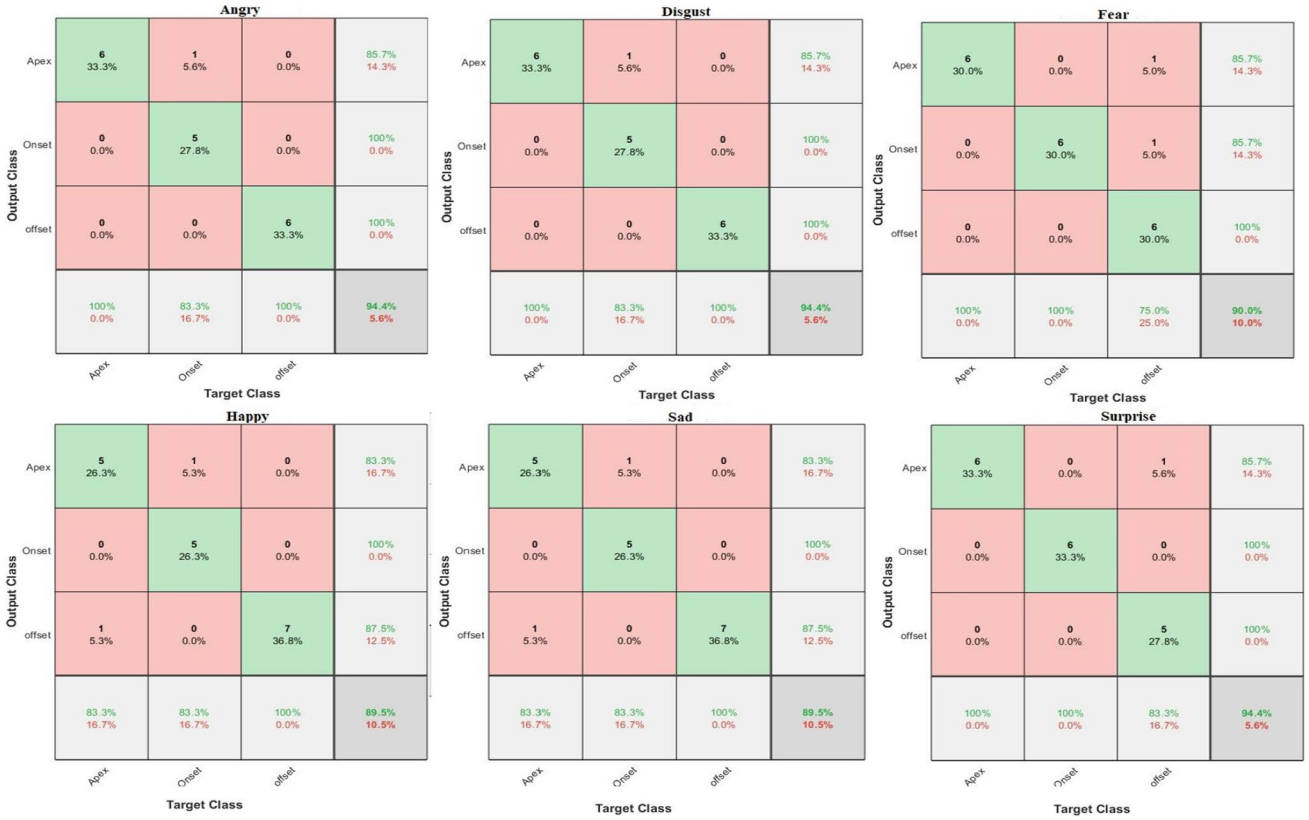

Fig. 13 Confusion matrix of three classes for the intensity of facial expressions for CK+ on the test dataset

The recognition rates of intensity segments apex, onset, and offset are also explained in Table 12. This Intensity-Net CNN part of the algorithm shows the results of the recognition rate of intensity segments which is 93.47% as compared to the already existing state-of-the-art algorithms.

5 Conclusion

This research work implemented a multi-level convolutional neural network for recognition of facial expression along with intensity estimation. The proposed method employs in two phases. In the first phase, Expression-Net CNN architecture focuses only on facial expression recognition. It uses various layers of the CNN model with different parameters and filters for the classification of facial expressions. In the second phase, Intensity-Net CNN architecture is used for intensity segments estimation of facial expressions. It uses different layers, parameters, and filters for intensity estimation. The intensity is estimated in terms of three segments: onset, offset, and apex values of expressions. Due to no. of parameters and filters involved in makes the model less complex, that's why the proposed model has shown better accuracy of 98.8% for CK+ and 97.7% for JAFFE dataset, which is quite satisfactory as compared to the other state-of-the-art techniques. This research considers only the face frontal pose of image sequences. As future work, the proposed method will be used for multi-pose image sequences or videos of facial expressions to find the intensity of facial expressions.

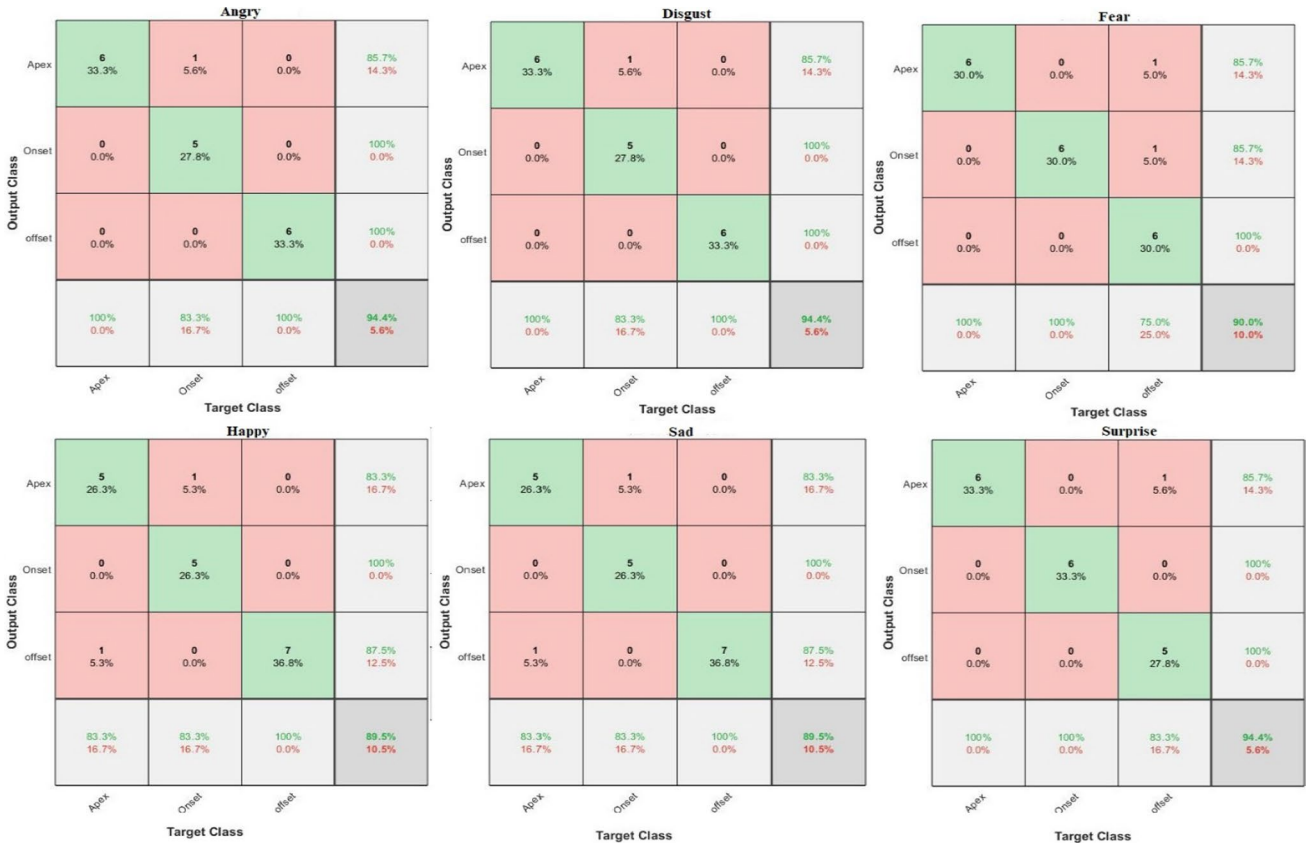

Fig. 14 Confusion matrix of three classes for intensity of facial expressions for JAFFE on the test dataset

Table 9 Recognition rate (%) comparison with state-of-the-art methods on CK+ dataset 2017

Cited year	Method	Accuracy (%)
2020	ML-DCNNet	98.8
2019	ACN [10]	98.0
2018	3DCNN [11]	98.77
2017	CNN [22]	96.10
2016	PHRNN +MSCNN [23]	98.50
2015	CNN [24]	83
2014	CNN [25]	87.98

Table 10 Overall intensity rate (%) comparison with state-of-the-art methods on CK+ dataset

Cited year	Method	Accuracy (%)
2020	Proposed ML-DCNNet	99.14
2019	ACN [10]	98.0
2018	3DCNN [11]	98.77
2017	CNN [22]	96.10
2016	PHRNN +MSCNN [23]	98.50
2015	CNN [24]	83
2014	CNN [25]	87.98

Table 11 Recognition rate (%) comparison with state-of-the-art methods on JAFFE dataset

Cited year	Method	Accuracy (%)
2020	ML-DCNNet	97.7
2019	ACN [10]	92.8
2018	CNN + SVM [26, 27]	95.31
2017	CNN [22]	96.10
2016	PHRNN +MSCNN [23]	98.50
2015	CNN [24]	83
2014	CNN [25]	87.98

Table 12 Overall intensity rate (%) comparison with state-of-the-art methods on JAFFE dataset

Cited year	Method	Accuracy (%)
2020	ML-DCNNet	93.47
2019	ACN [10]	92.8
2018	CNN + SVM [26]	95.31
2017	CNN [22]	96.10
2016	PHRNN +MSCNN [23]	98.50
2015	CNN [24]	83
2014	CNN [25, 28]	87.98

Acknowledgements The authors acknowledge the Ministry of Education and the Deanship of Scientific Research, Najran University. Kingdom of Saudi Arabia, under code number NU/ESCI/19/001.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Rukmani, P.; Teja, G.K.; Vinay, M.S.: Industrial monitoring using image processing, IoT and analyzing the sensor values using big data. Proc Comput Sci **133**, 991–997 (2018)
2. Mohammad Mahoor, B.H.: Facial expression recognition using enhanced deep 3D convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2017)
3. Kanade, T.: Picture processing system by computer complex and recognition of human faces (1974).
4. Snoek, J.; Larochelle, H.; Adams, R.P.: Practical bayesian optimization of machine learning algorithms. In: Advances in neural information processing systems (2012)
5. Sandbach, G.; et al.: Static and dynamic 3D facial expression recognition: a comprehensive survey. Image Vis. Comput. **30**(10), 683–697 (2012)
6. Fasel, B.; Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recogn. **36**(1), 259–275 (2003)
7. Krizhevsky, A.; Sutskever, I.; Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems (2012)
8. Verma, R.; et al.: Quantification of facial expressions using high-dimensional shape transformations. J. Neurosci. Methods **141**(1), 61–73 (2005)
9. Jolliffe, I.: Principal Component Analysis. Springer, Berlin (2011)
10. Kamarol, S.K.A.; et al.: Joint facial expression recognition and intensity estimation based on weighted votes of image sequences. Pattern Recogn. Lett. **92**, 25–32 (2017)
11. Minaee, S.; Abdolrashidi, A.: Deep-emotion: Facial expression recognition using attentional convolutional network. arXiv preprint arXiv:1902.01019 (2019)
12. Zhao, J.; Mao, X.; Zhang, J.: Learning deep facial expression features from image and optical flow sequences using 3D CNN. Vis. Comput. **34**(10), 1461–1475 (2018)
13. Valstar, M.F., et al.: Fera 2015-second facial expression recognition and analysis challenge. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). IEEE (2015)
14. Valstar, M.F., et al.: Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In: 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017). IEEE (2017)
15. Moghadam, S.M.; Seyyedsalehi, S.A.; Amini, N.: Nonlinear synthesis of expression variation dynamics on video using deep dynamic bottleneck neural networks. In: 2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME). IEEE (2017)
16. Quan, C.; Qian, Y.; Ren, F.: Dynamic facial expression recognition based on K-order emotional intensity model. In: 2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014). IEEE (2014)
17. Chang, K.-Y.; Chen, C.-S.; Hung, Y.-P.: Intensity rank estimation of facial expressions based on a single image. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics. IEEE (2013)
18. Murtaza, M.; et al.: Analysis of face recognition under varying facial expression: a survey. Int. Arab J. Inf. Technol. **10**(4), 378–388 (2013)
19. Tian, J.; Ma, K.-K.: A survey on super-resolution imaging. SIViP **5**(3), 329–342 (2011)
20. Shin, H.-C.; et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
21. Abdel-Hamid, O., et al.: Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2012)
22. Lucey, P., et al.: 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)
23. Lyons, M.J., et al.: The Japanese female facial expression (JAFFE) database. In: Proceedings of Third International Conference on Automatic Face and Gesture Recognition (1998).
24. Chen, X., et al.: Convolution neural network for automatic facial expression recognition. In: 2017 International Conference on Applied System Innovation (ICASI). IEEE (2017)
25. Mollahosseini, A.; Chan, D.; Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2016)
26. Li, W., et al.: A deep-learning approach to facial expression recognition with candid images. In: 2015 14th IAPR International Conference on Machine Vision Applications (MVA). IEEE (2015)
27. Ijjina, E.P.; Mohan, C.K.: Facial expression recognition using kinect depth sensor and convolutional neural networks. In: 2014 13th International Conference on Machine Learning and Applications. IEEE (2014).
28. Shima, Y.; Omori, Y.: Image augmentation for classifying facial expression images by using deep neural network pre-trained with object image database. In: Proceedings of the 3rd International Conference on Robotics, Control and Automation (2018)