BOHR International Journal of Internet of things, Artificial Intelligence and Machine Learning
2022, Vol. 1, No. 1, pp. 56–64
DOI: 10.54646/bijiam.2022.09
www.bohrpub.com

METHODS

Hybrid facial expression recognition (FER2013) model for real-time emotion classification and prediction

Ozioma Collins Oguine*, Kanyifeechukwu Jane Oguine, Hashim Ibrahim Bisallah and Daniel Ofuani

Department of Computer Science, University of Abuja, Abuja, Nigeria

***Correspondence:**
Ozioma Collins Oguine,
oziomaoguine007@gmail.com

Received: 08 August 2022; **Accepted:** 05 September 2022; **Published:** 27 September 2022

Facial expression recognition is a vital research topic in most fields ranging from artificial intelligence and gaming to human-computer interaction (HCI) and psychology. This paper proposes a hybrid model for facial expression recognition, which comprises a deep convolutional neural network (DCNN) and a Haar Cascade deep learning architecture. The objective is to classify real-time and digital facial images into one of the seven facial emotion categories considered. The DCNN employed in this research has more convolutional layers, ReLU activation functions, and multiple kernels to enhance filtering depth and facial feature extraction. In addition, a Haar Cascade model was also mutually used to detect facial features in real-time images and video frames. Grayscale images from the Kaggle repository (FER2013) and then exploited graphics processing unit (GPU) computation to expedite the training and validation process. Pre-processing and data augmentation techniques are applied to improve training efficiency and classification performance. The experimental results show a significantly improved classification performance compared to state-of-the-art (SoTA) experiments and research. Also, compared to other conventional models, this paper validates that the proposed architecture is superior in classification performance with an improvement of up to 6%, totaling up to 70% accuracy, and with less execution time of 2,098.8 s.

Keywords: deep learning, DCNN, facial emotion recognition, human-computer interaction, Haar Cascade, computer vision, FER2013

Introduction

Humans a natural ability to understand facial expressions. In real life, humans express the emotions on their faces to show their psychological state and disposition at a time and during their interactions with other people. However, the current trend of transferring cognitive intelligence to machines has stirred up conversations and research in the domain of human-computer interaction (HCI) and computer vision, with a particular interest in facial emotion recognition and its application in human computer collaboration, data-driven animation, human robot communication, etc.

Since emotions are physical and instinctive, they immediately cause physical responses to dangers, rewards, and other environmental elements. The objective measurements used to ascertain how people respond to these features include skin conductance (EDA/GSR), voice, body language, locomotion, brain activity (fMRI), heart rate (ECG), and facial expressions. A notable ability for humans to interpret emotions is crucial to effective communication; hypothetically, 93% of efficient conversation depends on the emotion of an entity. Hence, for ideal human-computer interaction (HCI),

Abbreviations: DCNN, deep convolutional neural network; CNN, convolutional neural network; FER, facial emotion recognition; GPU, graphical processing unit; HCI, human computer interaction; ECG, electrocardiography; FMRI, functional magnetic resonance imaging; AFEW, acted facial expressions in the wild; RNN, recurrent neural network; LSTM, long short-term memory; HMM, hidden Markov model; ReLU, rectified linear unit; TFE, tandem facial expression; ANN, artificial neural network; SVM, support vector machine; IRNN, image recognition neural network; FERET, facial recognition technology; SoTA, state-of-the-art.

CC BY NC ND

a high-level understanding of the human emotion is required by machines.

Emotions are a fundamental part of human communication, driven by the erratic nature of the human mind and the perception of relayed information from the environment. There are varied emotions that inform decision-making and are vital components in individual reactions and psychological state. Contemporary psychological research observed that facial expressions are pre-dominantly used to understand social interactions rather than the psychological state or personal emotions. Consequently, the credibility assessment of facial expressions, which includes the discernment of genuine (natural) expressions from postured (deliberate/volitional/deceptive) expressions, is a crucial yet challenging task in facial emotion recognition. This research will focus on educating objective facial parameters from real-time and digital images to determine the emotional states of people given their facial expressions, as shown in **Figure 1**.

Over the last 6 years, advancement in deep learning has propelled the evolution of convolutional networks. Computer vision is an interdisciplinary scientific field that equips the computer with a high-level understanding from images or videos that replicate human visual prowess within a computer perspective. Its aims are to automate tasks, categorize images from a given set of classes, and have the network determine the pre-dominant class present in the image. It can be implied that computer vision is the art of making a computer "see" and impacting it with human intelligence of processing what is "seen." More recently, deep learning has become the go-to method for image recognition and detection, far usurping medieval computer vision methods due to the unceasing improvement in the state-of-the-art performance of the models.

A gap identified by this research is with regards to the fact that most datasets of preceding research consist of well-labeled (posed) images obtained from a controlled environment, usually postured. According to Lopez et al., this anomaly increased the likelihood of model overfitting, given insufficient training data availability, ultimately causing a relatively lesser efficiency in predicting emotions in uncontrolled scenarios (1). Consequently, this research also identified the importance of lighting in facial emotion recognition (FER), highlighting that poor lighting conditions could decline the model's efficiency.

This research will use convolutional neural networking (CNN) to model some critical extracted features used for facial detection and classify the human psychological state into one of the six emotions or a seventh neutral emotion. It will also employ the Haar Cascade model for real-time facial detection. It is worthy of note that given the hand-engineered nature of the features and model dependency on prior knowledge, a resultant comparatively higher model accuracy is required.

Empirical review of related work

Over the years, several scholars have embarked on research to tackle this novel challenge. Ekman and Friesen highlighted seven basic human emotions independent of the culture a human being is born into (anger, fear, disgust, happiness, sad, surprise, and neutral) (2). Sajid et al., a study using the facial recognition technology (FERET) dataset, recently established the significance of facial asymmetry as a marker for age estimation. The study highlighted the simplicity with which the right face's asymmetry may be identified relative to the left face (3). Below are some reviews of works of literature pertinent to this research.

A model was trained on individual frames from movies and still digital images using a CNN-RNN mixed architecture by Kahou et al. (4). The acted facial expressions in the wild (AFEW) 5.0 dataset was used for the video clips, and a combination of the FER2013 and Toronto Face Database (5) was used for the photographs. Long short-term memory (LSTM) units were replaced by IRNNs, which are composed of rectified linear units (ReLUs) (6). IRNNs were appropriate because they offered a straightforward solution to the vanishing and exploding gradient problem. The accuracy of the study as a whole was 0.528.

Face detection still has a serious problem with how faces appear in poses. The explanation for the variation in face position look was offered by Ratyal et al. Using subject-specific descriptors, three-dimensional pose invariant approach was used (7, 8).

The inter-variability of emotions between individuals and misclassification are two issues with still image-based FERs that Ming Li et al., to address (9), suggest a neural network model. Their model consists of two convolutional neural networks; the first is trained on facial expression databases, while the second is a Deep ID network that learns features relevant to identity. These two neural networks were integrated into a tandem facial expression of TFE feature and delivered to the fully linked layers to produce a new model. Arousal and valence emotion annotations employing face, body, and context information were used by Mou et al. to infer group emotion (10). To identify group-level emotion in photos, Tan et al. combined two different forms of CNNs, namely, individual facial emotion CNNs and global image-based CNNs (11). Different pooling methods such as average and max pooling are used to downsample the inputs and aid in generalization (12, 13). Dropout, regularization, and data augmentation were used to prevent overfitting. Batch normalization was developed to help prevent gradient vanishing and exploding (14, 15).

As can be inferred from the literature highlighted above, several innovative research conducted by other scholars has emphasized continuous upscaling of accuracy without consideration for efficiency simultaneously. This research paper proposes a more efficient and accurate model with improved generalization, as discussed in subsequent

FIGURE 1 | FER2013 sample training set images.

sections. **Table 1** summarizes previous reported classification accuracies FER2013. Most reported methods perform better than the estimated human performance (∼65.5%). In this work, a state-of-the-art accuracy of 70.04% was achieved.

Theoretical background

The emotions are divided into groups according to surprise, wrath, happiness, fear, disgust, and neutrality in order to solve the FER problem (21). It creates a classification algorithm utilizing the features that were extracted. In the past, researchers have used classifiers for identifying facial expressions, including support vector machine (SVM), artificial neural network (ANN), hidden Markov model (HMM), and K-nearest neighbor (KNN). This research employs a modified hybrid model, which comprises two models (CNN and Haar Cascade). Following are explanations of several model components that constitute an architecture for learning various degrees of abstraction and representations. In the output layer for seven emotion classifications, these elements comprise convolutional layers, dropout, ReLU activation function, categorical cross entropy loss, Adam optimizer, and softmax activation function, as illustrated in **Figure 2**. This section will also emphasize the modifications made to these components through hyperparameter tuning that enhanced the proficiency and accuracy of the hybrid model.

Convolutional layer

The central component of a convolutional neural network that does the majority of the computational work is the conv layer. The convolution layer applies a filter f_k with a kernel size of n × m to an input x in order to perform a convolution (21) for a given input. There are n × m input connections. The following equation can be used to compute the result.

$$C\left(X_{U,V}\right) = \sum_{i=n/2}^{n/2} \sum_{j=-n/2}^{n/2} f_k\left(i,j\right) x_{(u-i,v-j)} \tag{1}$$

TABLE 1 | Summary of previous reported accuracies for the FER2013 dataset.

Methods	Accuracy rating
CNN (16)	62.44%
GoogleNet (17)	65.20%
VGG + SVM (18)	66.31%
Conv + Inception Layer (19)	66.40%
Bag of words (20)	67.40%
CNN + Haar Cascade (This work)	70.04%

Max pooling layer

By using the max function, it significantly decreases the input Saravanan et al. (22). Let m represent the filter's size and x_i represent the input. **Figure 2** depicts the application of this layer, and the equation can be used to determine the result.

$$M\left(x_i\right) = \max\left(r_{i+k,i+/1}/\left|k\right| \le m/2, \left|l\right|\right) < m/2, k, l \in N \tag{2}$$

Rectified linear unit (ReLU) activation function

It determines the output for a specific value of the network's or a neuron's input p, as shown in Equation (3). ReLU was utilized for this research because it has no exponential function to calculate and has no vanishing gradient error. **Figure 2** shows the concatenated convolutional layers parallelly processed via ReLU activation functions to upscale accuracy and obtain facial features of images flawlessly.

$$f(x) = \begin{cases} 0, & \text{for } x < o \\ 1 & \text{for } x \ge 0 \end{cases} \tag{3}$$

Fully connected layer

This is a multilayer perceptron, and it changes every neuron from the previous levels into a neuron in the current layer. The following equation serves as its mathematical representation:

$$f(x) = \sigma(p \times x) \tag{4}$$

where *a* is the activation function, and p is the resultant matrix of size n × k. k is the dimension of x, and n is the number of neurons in a fully connected layer.

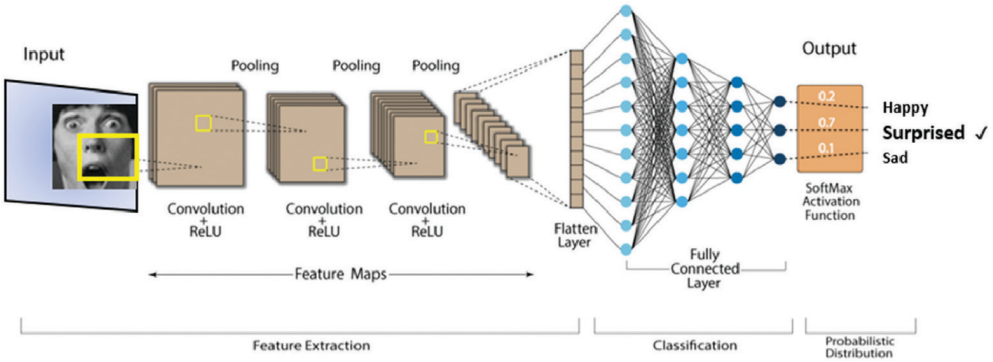

FIGURE 2 | Proposed CNN model structure.

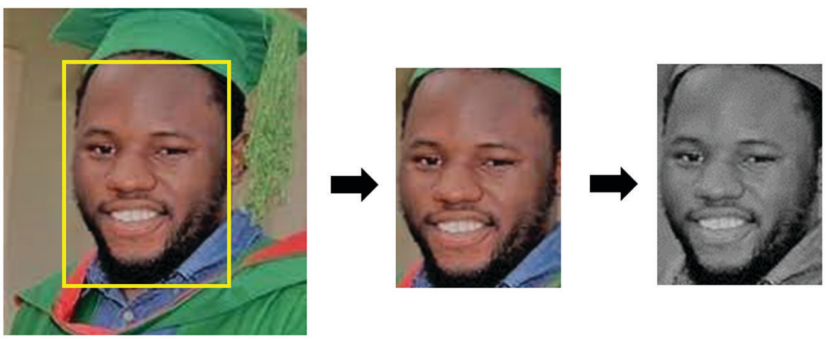

FIGURE 3 | Facial detection and conversion to grayscale.

Output Layer: It represents the class of the given input image and is one hot vector (21). It is expressed in the following equation.

$$C(x) = \left(i \left| \exists v_{j \neq i} and\ x_j \leq x_i \right.\right) \tag{5}$$

Softmax layer

Its main functionality is back propagation of error (21). Let the dimension of an input vector be P. Then, it can be represented as a mapping function as expressed in the following equation.

$$s(x) : R^P \rightarrow [0, 1]^P \tag{6}$$

and the output for each component $j(1 < j < P)$ is given as the following equation.

$$S(x)_j = \frac{e^x j}{\sum_{i=1}^{p} e^x i} \tag{7}$$

Proposed methodology

With the need for real-time object detection, several object detection architectures have gained wide adoption by most researchers. However, the hybrid architecture put forward by this research uses both the Haar Cascade face detection, a popular facial detection algorithm proposed by Paul Viola and Michael Jones in 2001, and the CNN model together. In **Figure 2**, the CNN architecture initially needs to extract input pictures of 48 × 48 × 1 (48 wide, 48 high, 1 color channel) from dataset FER2013. The network starts with an input layer equal to the input data dimension of 48 × 48. It also consists of seven concatenated convolutional layers parallelly processed via ReLU activation functions to upscale accuracy and obtains facial features of images flawlessly, as shown in **Figure 2**. This input is shared and the kernel size is the same across all submodels for feature extraction. Before a final output layer allows classification, the outputs from these feature extraction sub-models are flattened into vectors, concatenated into a large vector matrix, and transferred to a fully connected layer for assessment. A detailed step of the methodology is described in the architecture below.

The suggested CNN model, as shown in **Figure 2**, consists of a batch normalization layer, seven convolutional layers with independent learnable filters (kernels), each with a size of [3 × 3], and a local contrast normalization layer to eliminate the average from nearby pixels. A max pooling layer is also included in order to flatten and create dense layers while also reducing the spatial dimension of the image. It is followed by a layer that is entirely connected and Softmax, which categorizes seven emotions. The fully connected layer received a dropout of 0.5 to reduce over-fitting, and

TABLE 2 | Summary of the FER-2013 dataset.

	Surprise	Fear	Angry	Neutral	Sad	Disgust	Happy	Total
Train_count	3171	4097	3995	4965	4830	436	7215	28709
	Surprise	**Fear**	**Angry**	**Neutral**	**Sad**	**Disgust**	**Happy**	
Test_count	831	1024	958	1233	1247	111	1774	7178
Total Count of the test dataset								**35887**

the rectified linear units (ReLU) activation function was applied to all layers. After that, a Softmax output layer that can classify is connected to the concatenation of two comparable models.

Real-time classification

The output of the DCNN model is saved as a JSON string file. A JavaScript Object Notation (JSON) was deemed suitable for this research because it stores and allows faster data exchange, according to Ingale et al. (23). The model.tojson() function of python is used to write the output of the trained model into JSON. A pre-trained Haar Cascade XML file for frontal face detection of real-time facial feature classification was imported. A multiscale detection approach was implemented, and parameters such as coordinates of bounding boxes around detected faces and detectMultiScale (grayscale inputs) and scale factor were tuned for better efficiency.

In this research, Open CV's Haar Cascade detects and extracts the face region from the webcam's video feed through the flask app. This process follows a video conversion to grayscale, and the detected face is contoured or enclosed within a region surrounding the face (see **Figure 3**).

Experimental details

Dataset

The Kaggle repository's facial emotion recognition (FER2013) dataset is used in this study. The FER2013 dataset has 35,887 photos in total, of which 28,709 are tagged images and the rest 7,178 are part of the test set. The photographs in the FER2013 dataset are categorized into seven universal emotions: happy, sad, angry, fear, surprise, disgust, and neutral. The photos are 48 × 48 pixels in size and are grayscale. **Table 2** provides a summary of the dataset. The models are trained on an Nvidia Tesla K80 GPU running on Google Cloud (Colab Research). It is a group-based alternative to an iPython or Jupyter Notebook.

Pre-processing

Each image sample underwent a single application of the face detection and registration techniques. They are required for this procedure in order to correct any pose and illumination variations that may have resulted from the real-time facial detection operation. Keras library was used in this research for the pre-processing of both the test and train images before passing them to the deep convolution neural network (DCNN). This process includes cropping of detected faces and scaling of the detected images. For data cleaning purposes, all image target sizes were resized to a dimension of 48 × 48 × 1, converted into a grayscale color channel, and pre-processed in a batch size of 120. Square boxes were also used for facial landmark detection to apply illumination correction.

TABLE 3 | Summary of the proposed DCNN layers.

Layer (Type)	Output shape	Param #
Rescaling (Rescaling)	(None, 48, 48, 1)	0
sequential (Sequential)	(None, 48, 48, 1)	0
conv2d_7 (Conv2D)	(None, 46, 46, 32)	320
conv2d_8 (Conv2D)	(None, 44, 44, 64)	18496
max_pooling2d_4 (MaxPooling2	(None, 22, 22, 64)	0
dropout_3 (Dropout)	(None, 22, 22, 64)	0
conv2d_9 (Conv2D)	(None, 20, 20, 64)	36928
conv2d_10 (Conv2D)	(None, 18, 18, 64)	36928
conv2d_11 (Conv2D)	(None, 16, 16, 128)	73856
max_pooling2d_5 (MaxPooling2	(None, 8, 8, 128)	0
conv2d_12 (Conv2D)	(None, 6, 6, 128)	147584
conv2d_13 (Conv2D)	(None, 4, 4, 256)	295168
max_pooling2d_6 (MaxPooling2	(None, 2, 2, 256)	0
max_pooling2d_7 (MaxPooling2	(None, 1, 1, 256)	0
dropout_4 (Dropout)	(None, 1, 1, 256)	0
flatten_1 (Flatten)	(None, 256)	0
dense_2 (Dense)	(None, 1024)	263168
dropout_5 (Dropout)	(None, 1024)	0
dense_3 (Dense)	(None, 7)	7175
Total params:		879,623
Trainable params:		879,623
Non-trainable params:		0

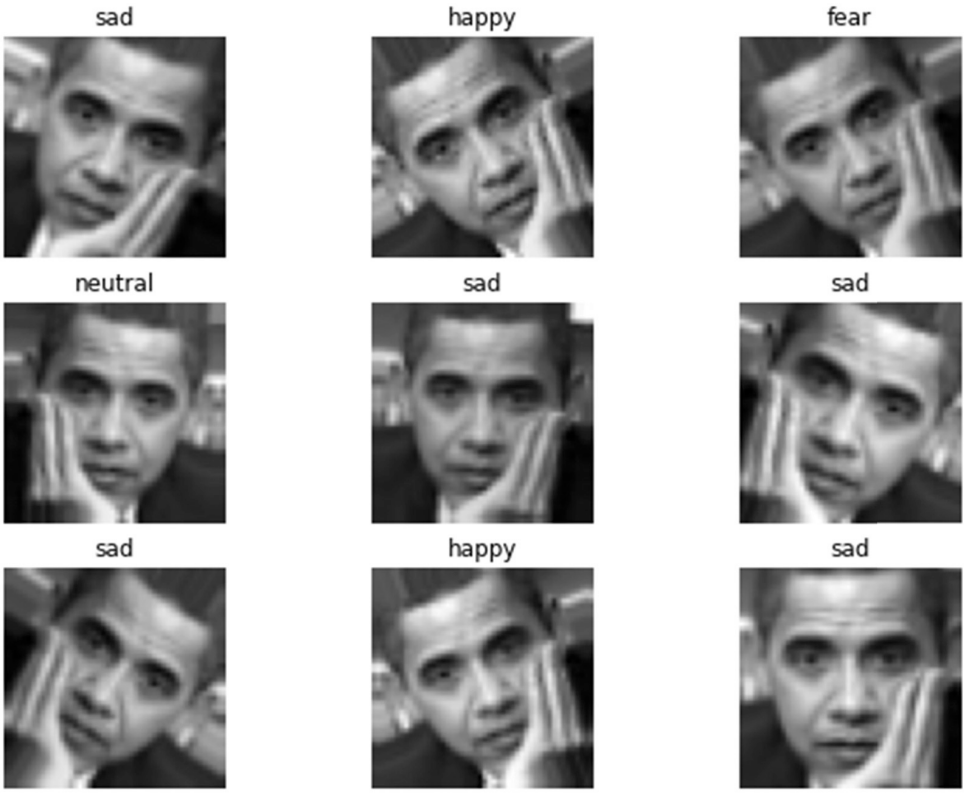

FIGURE 4 | Result of image data augmentation.

Training and validation

This research used the Keras library in python to accept images to train. To minimize the losses of the neural network during training, this research used a Mini-Batch Gradient Descent algorithm. This type of Gradient Descent algorithm was suitable because of its capabilities in finding the coefficients or weights of neural networks by splitting the training dataset into small batches, i.e., a training batch and a validation batch, using data augmentation techniques. From Keras, a sequential model was implemented to define the flow of the DCNN, after which several other layers, as described in the proposed methodology above and shown in **Figure 2**. With an increase in the convolutional layers, a concurrent increase in the kernel size was also adopted, which was used in parallel with the ReLU activation function for training the model. Dropouts were used at several layers of the DCNN to prevent the overfitting of the model. A Soft Max activation function and an Adam optimizer were used to improve the classification efficiency of the model. This research also adopted a categorical cross-entropy loss function. A summary of the DCNN structure is shown in **Table 3**.

Image data augmentation. For efficient generalization and optimized model performance, augmentation was used to expand the size of the training dataset by creating several modified variations of the images using the Image Data Generator Class in Keras. Here a target image was converted in an index of 0 to augmented variations, then looping it through nine instances of different emotions to see what predictability would look like during the test, as illustrated in **Figure 4**.

Hyperparameter tuning. A hyperparameter is a parameter whose value is used to control the training process of a model (neural network). Hyperparameter optimization entails choosing a set of optimal hyperparameters for a learning algorithm to improve the generalization or predictability of a model.

For this research, the random search technique is used wherein random combinations of a range of values for each hyperparameter have been used to find the best possible combination. The effectiveness of the random

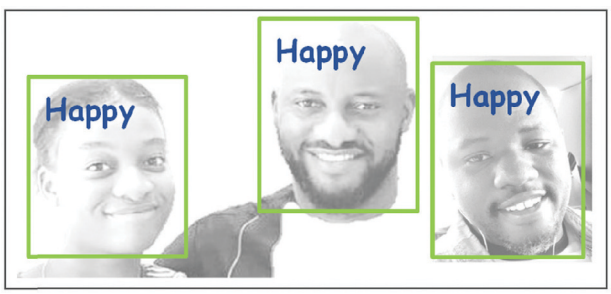

FIGURE 5 | Result of real-time test sample associated with the HAPPY emotion.

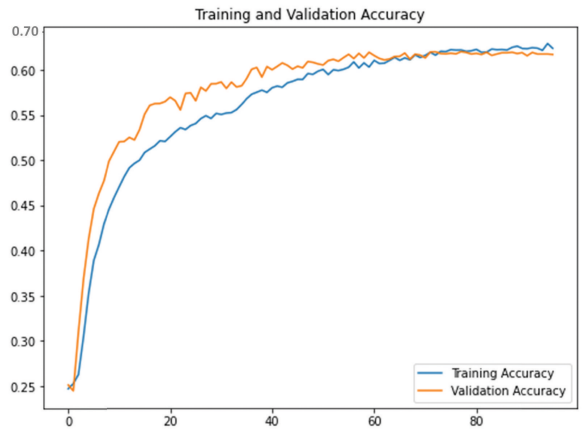

FIGURE 6 | Plots of the training and validation accuracy.

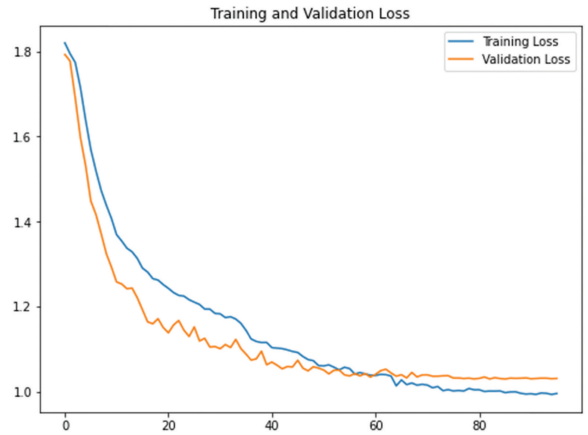

FIGURE 7 | Plot of the training and validation loss.

search technique for hyperparameter optimization has been demonstrated in Ref. (24).

Result analysis and discussion

The algorithm and facial emotion recognition model proposed by this research is based on two principal ideas. First, the utilization of high-capacity deep convolutional neural networks for feature extraction and emotion classification, which employs a single classifier for detecting faces from multiple views over both real-time scenarios and digital images or video frames. This research also sought to optimize the computational complexity of the deep convolutional neural network (DCNN) by modifying the architecture through the addition of layers to improve pattern identification in real-time or digital images. The additional layers apply more convolution filters to the image to detect features of the image. To further enhance the model's predictive efficiency and accuracy, the number of training Epoch was increased to 80.

The proposed model for emotion recognition uses three steps—face detection (see **Figure 3**), features extraction, and emotion classification—and achieves better results than the previous model. In the suggested method, validation accuracy rises as calculation time decreases, and validation loss is significantly decreased. The FER-2013 dataset, which includes the seven main emotions, was used to test the proposed DCNN model (sad, fear, happiness, angry, neutral, surprised, and disgust).

Figure 5 indicates the test sample result associated with happy emotions from the digital test image. The proposed model also predicted the identical emotion with decreased computation time compared to preceding models discussed in the literature review.

Table 5 describes the metrics used in measuring the success of the CNN model developed for this research. Given the necessary modifications made as proposed earlier, it was observed that, on average, the proposed model had a

FIGURE 8 | Confusion metrics of the model on emotion prediction.

predictive accuracy of 70%. The weighted average of the test dataset is also 70%.

After the training was done, the model was evaluated and the training and validation accuracy (**Figure 6**) and loss (**Figure 7**) were computed.

Figures 6, 7 show the CNN model's learning curve, which plots the model's learning performance over experience or time. After each update, the model was evaluated on the training dataset and a hold-out validation dataset. Learning curves were preferred as a graphic metric for this research because of their wide adoption in machine learning for models that optimize their internal parameters incrementally. From the learning curve in **Figure 7**, it can be observed that the plot of training loss and validation loss declines to the point of stability with a generalization gap of minimal difference. Also, it can be inferred in **Figure 6** that the plot of training accuracy and validation accuracy surges with an increase in training sequence and batch size with a minimal generalization gap. Thematically, the model can be a good fit and is proposed to generalize efficiently.

Figure 8 shows the CNN model's confusion matrix on the FER2013 testing set. It was inferred that the model

TABLE 4 | Describes all the hyperparameters used by the proposed model and their corresponding values.

Hyperparameter	Value	Description
Batch size	120	The number of images (from the training set) to be propagated through the network at a go.
Number of epochs	80	The number of complete passes through the training set.
Optimizer	Adam	Adam Optimization algorithm.
Learning rate (lr)	0.001	Controls the speed at which weights of the network are updated during training.
FC1 neurons	1024	Total number of neurons in the first fully connected layer
FC2 neurons	512	Total number of neurons in the second fully connected layer
Dropout	0.5	Dropout rate between fully connected layers
Convolution kernel size	3 × 3	Size of the kernel in each convolution layer
MaxPooling kernel size	2 × 2	Size of the kernel in each MaxPooling layer
MaxPooling strides	2	Kernel stride in each MaxPooling layer

TABLE 5 | Showing the evaluation metrics for the proposed CNN.

	Precision	recall	F1-score	support
Angry	0.62	0.64	0.63	958
Disgust	0.73	0.34	0.45	111
Fear	0.56	0.42	0.48	1024
Happy	0.91	0.91	0.91	1774
Neutral	0.64	0.71	0.67	1233
Sad	0.56	0.62	0.59	1247
Surprise	0.79	0.81	0.72	831
Accuracy			0.70	7178
Macro avg	0.69	0.65	0.66	7178
Weighted avg	0.70	0.70	0.70	7178

generalizes best in the classification of "happiness" and "surprise" emotions. In contrast, it performs reasonably average when classifying "disgust" and "fear" Emotion. This reduction in classification accuracy in "disgust" and "fear" emotions can be ascribed to the reduced number of training set samples for the two classes. The complacency between "fear" and "sadness" may be due to the inter-class similarities of the dataset.

Conclusion

This paper proposed a deep CNN model for real-time facial expression recognition based on seven emotional classes ("neutral," "happy," "sad," "angry," "surprised," "fear," and "disgusted"). The structure of the proposed model has good generality and classification performance. First, a variety of well-classified, high-quality databases were acquired. Then, the face region is detected, cut, and converted into a grayscale image of one channel to remove unnecessary information. Image data augmentation that increases the number and variations of training images is applied to solve the problem of overfitting. Hyperparameter tuning was employed to achieve a state-of-the-art classification performance accuracy of 70.04%.

In the proposed hybrid architecture, an optimal structure to reduce execution time and improve the classification performance of real-time and digital images was developed. This was done by adjusting the number of feature maps in the convolutional layer, layers in the neural network model, and numerous training epochs. Cross validation experiments showed that the proposed convolutional neural network (CNN) architecture has better classification performance and generality than some state- of-the-arts (SoTA). The Haar Cascade model employed for real-time facial detection showed better classification performance than other implementations. Experimental results confirmed the effectiveness of data pre-processing and augmentation techniques.

Some shortcomings in this research were a deficiency in predicting the "disgust" and "angry" emotions due to insufficient training dataset for the two-class categories. Another issue the proposed model faces is the reduced generality of real-time predictions. A major causative factor stems from the postured nature of the training images and environmental conditioning (lightening) of real-time test images. A significant impact of this research in human-computer interaction (HCI) will be the upscaling of software and AI systems to deliver an improved experience to humans in various applications, for instance, in Home robotic systems in the recommendation of moodbased music. This research will enhance psychological prognosis by assisting psychologists in detecting probable suicides and emotional traumas, and it can also be employed in the marketing, healthcare, and gaming industry.

Future research proposal

Recognizing human facial expressions is an important practice since it helps to determine a subject's mood or emotional state while they are being observed. Numerous situations could benefit from the straightforward idea of a machine being able to recognize a person's emotional state. As such, numerous potentials remain untapped in this area. State-of-the-art real-time algorithms (Faster R-CNN, HOG + Linear SVM, SSDs, and YOLO) are encouraged to be employed in future research as well. In addition, advancement to this research may include the utilization of regression in analyzing bio-signals and physiological

multimodal features in determining the levels and intensities of emotion classes discussed herein.

Data availability statement

This research paper analyzes the Facial Emotion Recognition (FER-2013) dataset housed by the Kaggle repository (https://www.kaggle.com/datasets/msambare/fer2013).

Author contributions

All authors made significant contributions to conception and design, data acquisition, analysis, and interpretation; participated in drafting the article and critically revising it for important intellectual content; agreed to submit it to the current journal; and gave final approval of the version to be published.

References

1. Montesinos Lopez OA, Montesinos Lopez A, Crossa J. *Overfitting, Model Tuning, and Evaluation of Prediction Performance in Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Cham: Springer (2022). doi: 10.1007/978-3-030-89010-0_4
2. Ekman P, Friesen WV. Constants across cultures in the face and emotion. *J Personal Soc Psychol.* (1971) 17:12. doi: 10.1037/h0030377
3. Sajid M, Ratyal NI, Ali N, Zafar B, Dar SH, Mahmood MT, et al. The impact of asymmetric left and asymmetric right face images on accurate age estimation. *Math Probl Eng.* (2019) 2019:1–10. doi: 10.1155/2019/8041413
4. Kahou SE, Michalski V, Konda K, Memisevic R, Pal C. Recurrent neural networks for emotion recognition in video. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. New York, NY (2015).
5. Dhall A, Goecke R, Lucey S, Gedeon T. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE Multimed.* (2012) 19:3441. doi: 10.1109/MMUL.2012.26
6. Le QV, Jaitly N, Hinton GE. A simple way to initialize recurrent networks of rectified linear units. *arXiv* [Preprint] (2015) doi: 10.48550/arXiv.1504.00941
7. Ratyal NI, Taj IA, Sajid M, Ali N, Mahmood A, Razzaq S. Three-dimensional face recognition using variance-based registration and subject-specific descriptors. *Int J Adv Robot Syst.* (2019) 16:1729881419851716. doi: 10.1177/1729881419851716
8. Ratyal NI, Taj IA, Sajid M, Ali N, Mahmood A, Razzaq S, et al. Deeply learned pose invariant image analysis with applications in 3D face recognition. *Math Probl Eng.* (2019) 2019:3547416. doi: 10.1155/2019/3547416
9. Li M, Xu H, Huang X, Song Z, Liu X, Li X. Facial Expression Recognition with Identity and Emotion Joint Learning. *IEEE Trans Affect Comput.* (2018) 12:544–50.
10. Mou W, Celiktutan O, Gunes H. Group-level arousal and valence recognition in static images: Face, body and context. *Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Ljubljana (2015).
11. Tan L, Zhang K, Wang K, Zeng X, Peng X, Qiao Y. Group Emotion Recognition with Individual Facial Emotion CNNs and Global Image-based CNNs. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. Glasgow (2017).
12. Giusti A, Ciresan DC, Masci J, Gambardella LM, Schmidhuber J. Fast image scanning with deep max-pooling convolutional neural networks. *Proceedings of the 2013 IEEE International Conference on Image Processing, ICIP 2013*. Piscataway, NJ (2013).
13. Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. *Proceedings of 2017 International Conference on Engineering and Technology, ICET 2017*. Antalya (2018).
14. Han B, Sim J, Adam H. BranchOut: Regularization for online ensemble tracking with convolutional neural networks. *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Piscataway, NJ (2017).
15. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*. France (2015).
16. Liu K, Zhang M, Pan Z. Facial Expression Recognition with CNN Ensemble. *Proceedings of the 2016 International Conference on Cyberworlds, CW 2016*. Chongqing (2016).
17. Ionescu RT, Popescu M, Grozea C. Local Learning to Improve Bag of Visual Words Model for Facial Expression Recognition. *Proceedings of the ICML Workshop on Challenges in Representation Learning*. Delhi (2013).
18. Georgescu MI, Ionescu RT, Popescu M. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access* (2019) 7:2917266.
19. Giannopoulos P, Perikos I, Hatzilygeroudis I. Deep learning approaches for facial emotion recognition: A case study on FER- 2013. In: Hatzilygeroudis I, Palade V editors. *Advances in Hybridization of Intelligent Methods. Smart Innovation, Systems and Technologies*. Berlin: Springer (2018).
20. Mollahosseini A, Chan D, Mahoor MH. Going deeper in facial expression recognition using deep neural networks. *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*. Piscataway, NJ (2016).
21. Handa A, Agarwal R, Kohli N. Article: Incremental approach for multi-modal face expression recognition system using deep neural networks Journal. *Int J Comput Vis Robot.* (2021) 11:1–20.
22. Saravanan A, Perichetla G, Gayathri K. Facial emotion recognition using convolutional neural networks. *arXiv* [preprint]. (2019) doi: 10.48550/arXiv.1910.05602
23. Ingale S, Kadam A, Kadam M. Facial Expression Recognition Using CNN with Data Augmentation. *Int Res J Eng Technol.* (2021) 7:771–9.
24. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res.* (2012) 13:281–305.