PSYCHOLOGY AND EDUCATION (2021) 58(2): 876-882
An Interdisciplinary Journal
Article Received: 16th October, 2020; Article Revised: 30th December, 2020; Article Accepted: 08th January, 2021

An Efficient and Accurate Real Time Facial Expression Detection Using CNN

B. Srujana*, Dr.M. Sree Devi[2]
[1]KL Deemed to be University, India
[2]KL Deemed to be University, India
*srujanaboddapati@gmail.com,

ABSTRACT

Real Facial expression acknowledgment (RTFER) has become a functioning examination zone that finds a ton of utilizations in territories like human-PC interfaces, human feeling investigation, mental investigation, clinical conclusion and so on Mainstream strategies utilized for this intention depend on math and appearance. Profound convolutional neural networks (CNN) have appeared to beat customary strategies in different visual acknowledgment errands including Facial Expression Recognition. Despite the fact that endeavours are made to improve the exactness of RTFER frameworks utilizing CNN, for functional applications existing strategies probably won't be adequate. This examination incorporates a conventional audit of RTFER frameworks utilizing CNN and their qualities and restrictions which assist us with comprehension and improve the RTFER frameworks further.

Keyword's:
CNN, Feature extraction

I. INTRODUCTION

Feeling acknowledgment is as a rule effectively investigated in the research of Computer Vision. With the new ascent and the advocacy of Machine Learning [1] and also Deep Learning [2] methods, the possibility to construct astute frameworks that precisely perceive feelings turned into a closer reality. In any case, this issue is demonstrated to be increasingly more mind boggling with the advancement of fields that are straightforwardly connected with feeling acknowledgment, for example, brain science and nervous system science. Miniature expressions and electroencephalography (EEG) signals, motions, manner of speaking, facial expressions and encompassing setting are a few terms that have a ground-breaking sway while distinguishing feelings in a human [3]. At that particular point when these factors are sorted out with the restrictions and issues of the current Vision of the Computer's calculations, feeling acknowledgment can get exceptionally mind boggling.

Facial expressions are fundamental focal point of this methodical survey. For the most part, a RTFER framework comprises of the accompanying advances: image securing, pre-preparing, feature extraction, arrangement, or relapse. To have the option to get a legitimate facial expression arrangement, it is exceptionally attractive to give utmost important information to the classifier, in the most ideal conditions. To do that, a regular RTFER framework will right off the bat pre-measure the information image. Single pre-handling step which is basic among most checked on papers is the face identification. Face discovery methods can make jumping boxes which delimit recognized faces, that are the ideal locales of interest (ROIs) to an ordinary RTFER framework. This undertaking is as yet testing, and it isn't ensured that all faces will be identified in a given info image. This is particularly evident when procuring images from a particular uncontrolled climate, where there might be development, unforgiving lighting conditions, dif RTFER ent presents, huge spans, among different variables [4]. At the point when the faces are appropriately identified, a traditional RTFER framework will deal with the recovered ROIs to set up the information that will be taken care of into the classifier. Typically, this pre-handling step is separated into a few sub steps, for example, force standardization for enlightenment changes, commotion channels for image smoothing, information growth (DA) [5] to expand the preparation information, pivot revision for the turned faces, resizing imaged for the different ROI

Article Received: 16th October, 2020; Article Revised: 30th December, 2020; Article Accepted: 08th January, 2021

sizes, trimming of image for a superior foundation sifting, among others.

Then after the pre-preparing, one can recover important features from pre-handled ROIs. There are various features which can be chosen, for example, Actions Units (AUs) [6], movement of particular facial milestones, distance between facial tourist spots, facial surface, angle features, etc. At that point, these features are taken care of into a classifier. For the most part, the classifiers utilized in a RTFER framework are (SVMs) - Support Machine Vectors [7] or Convolutional Neural Networks (CNNs) [8-10].

II. LITERATURE REVIEW

In a paper [10], a hybrid approach in which multi modal information for the facial emotion recognition is used. In the experiment conducted by authors, they chose two different speakers using two different languages. The evaluation was carried out with the three different media clips, (1) Only audio information of emotions, (2) Only video information of the emotions, (3) both video and audio information (original video clip). Video and audio dominance of each type of emotion [11] is recorded and compared. The results of audio and facial expression recognition [12] are provided as input to the weighing matrix. Inside the weighing matrix computations are made and the expression whose computed value is maximum is the result.

According to a paper [13], the problem which is solved is about Emotion detection using face expressions [14]. Microsoft Kinect was used for 3D modelling of the face. Microsoft Kinect has 2 cameras. One works with visible light and also other one works with infrared light. It gives three-dimensional co-ordinates of specific face muscles. Facial Action Coding System (FACS) was used to return special coefficients called Action Units (AU). There are 6 Action Units. These Action Units (AU) represent different region of face. Six men of the age group 26-50 participated and tried to mimic the emotions specified to them. Each person had 2 sessions and each session had 3 trials. 3-NN had an accuracy of 96%.MLP had an accuracy of 90%

According to the paper [15], CERT can detect 19 different facial actions, 6 different prototypical emotions and 3D head orientation using Facial Action Unit Coding System (FACS) and three emotion modules. It follows 6 stages: (1) Face Detection using Gentle Boost as boosting algorithm, (2) Facial Feature Detection – Specific location estimates are estimated by combining log likelihood ratio and feature specific prior at that location, and these location estimates are refined using Linear regressor, (3) Face Registration – affine wrap is made and L2 Norm is minimized between wrapped facial feature position and canonical position from GENKI dataset, (4) Feature Extraction – feature vector is obtained using Gabor filter on face patch from previous patch, (5) Action Unit Recognition – feature vector is fed to Support Vector machine to obtain Action Unit Intensities, (6) Expression Intensity and Dynamics – Empirically CERT outputs significantly correlates with facial actions.

In a paper [16], Psychological theories state that all human emotions can be classified into six basic emotions: sadness, happiness, fear, anger, neutral and surprise. Three systems were built- one with audio, another with face recognition [17] and one more with both. The performances of all the systems were compared. Features used for speech-global prosodic features, for facial data from 102 markers on face.

In a paper [18], a brief explanation of how feature selection (or) the feature extraction has been employed and how the pre-processing step has been done to point out more prominent features were explained. And in the next stage it was clearly stated that the classification was applies by the use of some particular features which are said to be a subset [20]. Both feature level and also the decision level integration were implemented. According to that paper the feature selection's [19] quality will determine the accuracy of the recognition directly, whereas the feature selection wants some more professional knowledge and also the recognition rate takes more time, and is low also laborious. The result proved that performance of both the systems was similar. However, recognition rate for specific emotions presented significant errors. The type of integration to be used is depends on nature of the application.

Article Received: 16th October, 2020; Article Revised: 30th December, 2020; Article Accepted: 08th January, 2021

Papers [21-23], were useful to identify patterns in the Database.

III.METHODOLOGY& IMPLEMENTATION

We have used the data of 7 expressions from the [Fer2013] dataset. There exists 28709 48×48 grayscale images in total in training set, and a total of 3589 48×48 grayscale images exist in the test set. we normalized them by deducting the mean of training images from each image. In order to classify the expressions, we 've mainly taken use of the features which were generated by the convolution layers by the use of raw pixel data. The training set was given as the input to the CNN architecture, with the parameters of both the Python and TensorFlow, for the training of the architecture on the platform of Anaconda. For the training, we made use of all images in training set with 1001 epochs along with a batch size of 128 and is cross-validated. For the validation of our model in each of iteration, we've used validation set, for the evaluation of the efficient performance of model, we've used- test set.

The complete CNN architecture had divided into one- input layer, 3 convolution layers and a pooling, and some fully connected layers for getting prediction of the results of 6 the expressions.

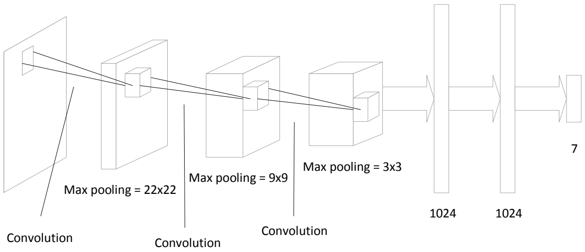
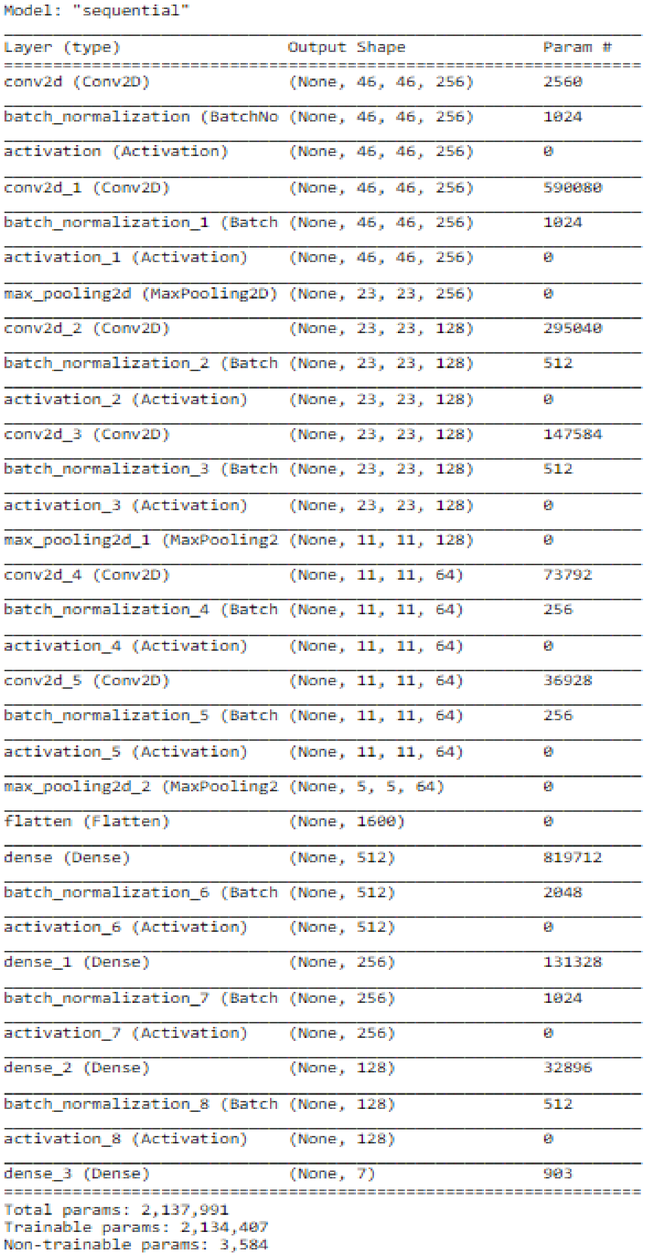


Figure 1. CNN architecture for the detection of face expressions from FER2013.

The 1st layer of the convolution uses a convolution kernel which is of 5X5, whereas the 2nd convolution layer makes uses the kernel of same size, and the 3rd convolution layer makes uses of a size of convolution kernel (4X4). And then Max-pooling layer will be employed for pooling of each layer. The CNN architecture for our FER would be as shown in fig 1. After some sufficient number of the training iterations which are of above architecture, final parameters of training result were saved to later use. Fig.2 shows architecture working.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 46, 46, 256)       2560
_________________________________________________________________
batch_normalization (BatchNo (None, 46, 46, 256)       1024
_________________________________________________________________
activation (Activation)      (None, 46, 46, 256)       0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 46, 46, 256)       590080
_________________________________________________________________
batch_normalization_1 (Batch (None, 46, 46, 256)       1024
_________________________________________________________________
activation_1 (Activation)    (None, 46, 46, 256)       0
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 23, 23, 256)       0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 23, 23, 128)       295040
_________________________________________________________________
batch_normalization_2 (Batch (None, 23, 23, 128)       512
_________________________________________________________________
activation_2 (Activation)    (None, 23, 23, 128)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 23, 23, 128)       147584
_________________________________________________________________
batch_normalization_3 (Batch (None, 23, 23, 128)       512
_________________________________________________________________
activation_3 (Activation)    (None, 23, 23, 128)       0
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 11, 11, 128)       0
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 11, 11, 64)        73792
_________________________________________________________________
batch_normalization_4 (Batch (None, 11, 11, 64)        256
_________________________________________________________________
activation_4 (Activation)    (None, 11, 11, 64)        0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 11, 11, 64)        36928
_________________________________________________________________
batch_normalization_5 (Batch (None, 11, 11, 64)        256
_________________________________________________________________
activation_5 (Activation)    (None, 11, 11, 64)        0
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 5, 5, 64)          0
_________________________________________________________________
flatten (Flatten)            (None, 1600)              0
_________________________________________________________________
dense (Dense)                (None, 512)               819712
_________________________________________________________________
batch_normalization_6 (Batch (None, 512)               2048
_________________________________________________________________
activation_6 (Activation)    (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 256)               131328
_________________________________________________________________
batch_normalization_7 (Batch (None, 256)               1024
_________________________________________________________________
activation_7 (Activation)    (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 128)               32896
_________________________________________________________________
batch_normalization_8 (Batch (None, 128)               512
_________________________________________________________________
activation_8 (Activation)    (None, 128)               0
_________________________________________________________________
dense_3 (Dense)              (None, 7)                 903
=================================================================
Total params: 2,137,991
Trainable params: 2,134,407
Non-trainable params: 3,584
```

Figure2. Model summary with all parameters

Video of a person is recorded using web camera. The video is converted to frames and provided as an input to a classifier to get the desired emotion. This system has three modules.

Article Received: 16th October, 2020; Article Revised: 30th December, 2020; Article Accepted: 08th January, 2021

i. Pre-processing: Real-time video is captured using the camera at the rate of 30 frames per second (fps). The frames are in BGR (Blue Green Red) format. It is converted into greyscale format which makes computing easy.

ii. Face detection*:* A pre-trained classifier called Haar-cascade provided by OpenCV is used for face detection. Haar features exists mainly for the extraction of the face features. There exist 4 categories of them which can be combined effectively to from some feature templates. And there are 2 kinds of the rectangles in each feature template one is black and one is white. while extraction of the features, the template will be covering a particular patch(area) of the image, and then, it calculates sum of both the pixel values of 2 types of the rectangles one after another. And then it finally deducts the sum of both pixel values of both kinds of rectangles so that to obtain the feature of that template, which are said to be Haar features. And then finally, returns face coordinates. These co-ordinates are used to crop the image and obtain only the face.

iii. Classifier: Previously built Convolution Neural Network (CNN) model is used and the license to use this CNN model is provided in [25]. The code available in the link [25] will be useful to train the model. The trained data is used to predict emotion. A list with probabilities of all 6 emotions is obtained as an output. The required output is the maximum of these values and the corresponding emotion is predicted as the final output.

After all of the three modules, using OpenCV's video capture () our facial emotion analysis window appears to capture emotion of the person in front of the webcam and detects his/her expression in Realtime.

Confusion matrix:

Confusion Matrix is applied and plotted to obtain and know which emotion is usually get confused with each other among all the emotions.

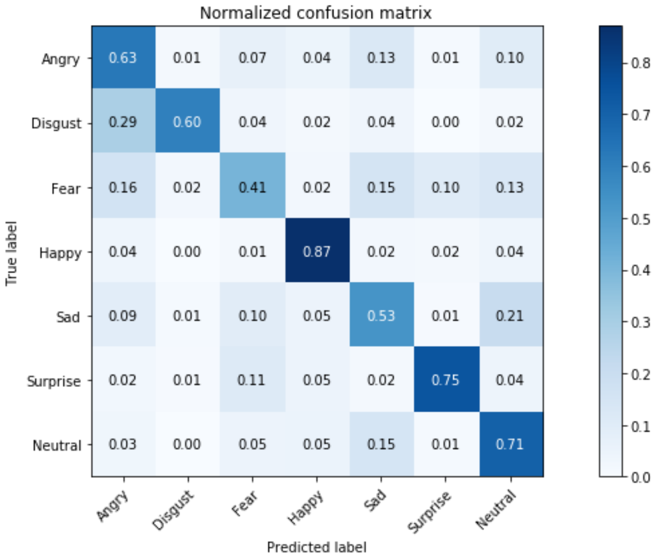
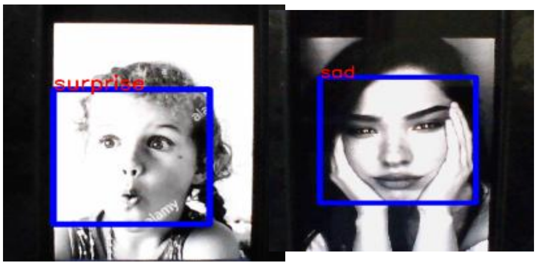


Figure 3. Confusion matrix of different emotions

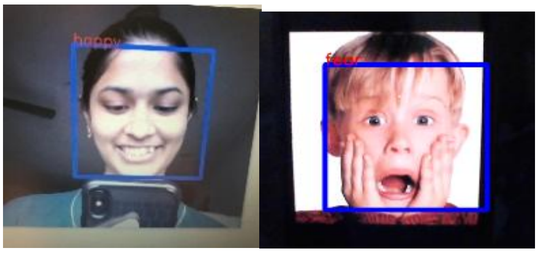
Our precision is too higher for the emotions Surprised, Happy, Neutral, and also Angry. And also consistent within our observations using our real-time model. Both recall and the precision vary greatly by the class. And, precision is too low for the classes – Afraid and even recall is high for classes such as Happy. From our visual inspection of the dataset, we noticed, some emotion pairs - Disgusted and Surprised (or) Surprised and Afraid are almost similar.

IV. Results and Discussion:



{a} {b}

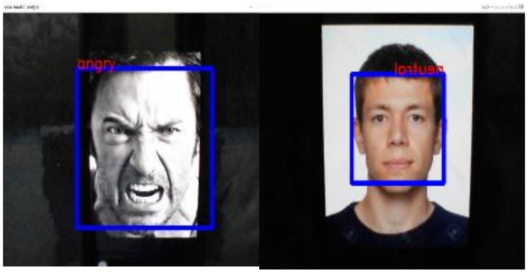
Fig.4.RealTimeFaceExpression(a)Surprise(b)sad



{c} {d}

Fig.5.Real-Time Face Expression(c)Happy(d) Fear

Article Received: 16th October, 2020; Article Revised: 30th December, 2020; Article Accepted: 08th January, 2021

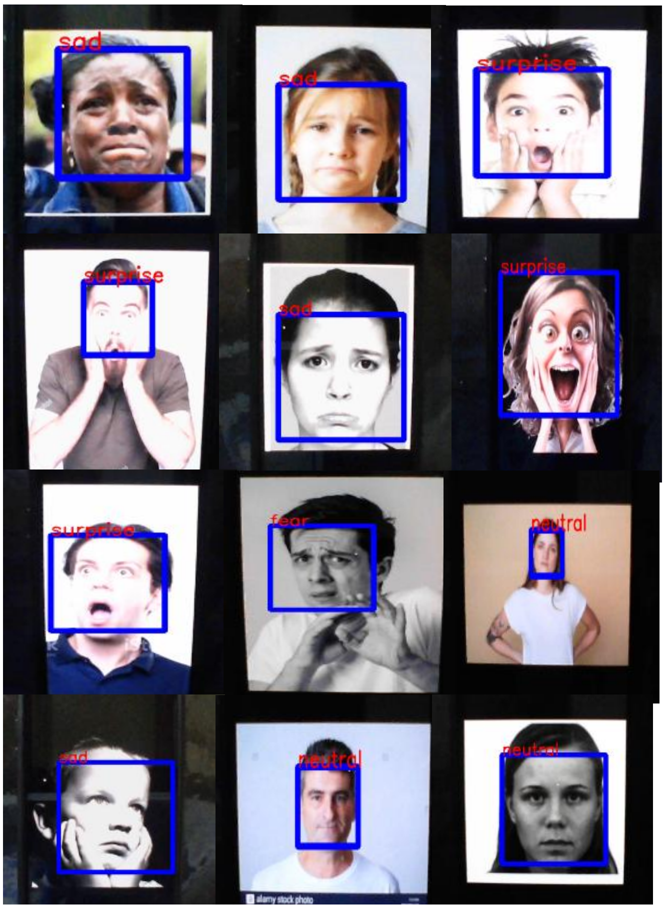


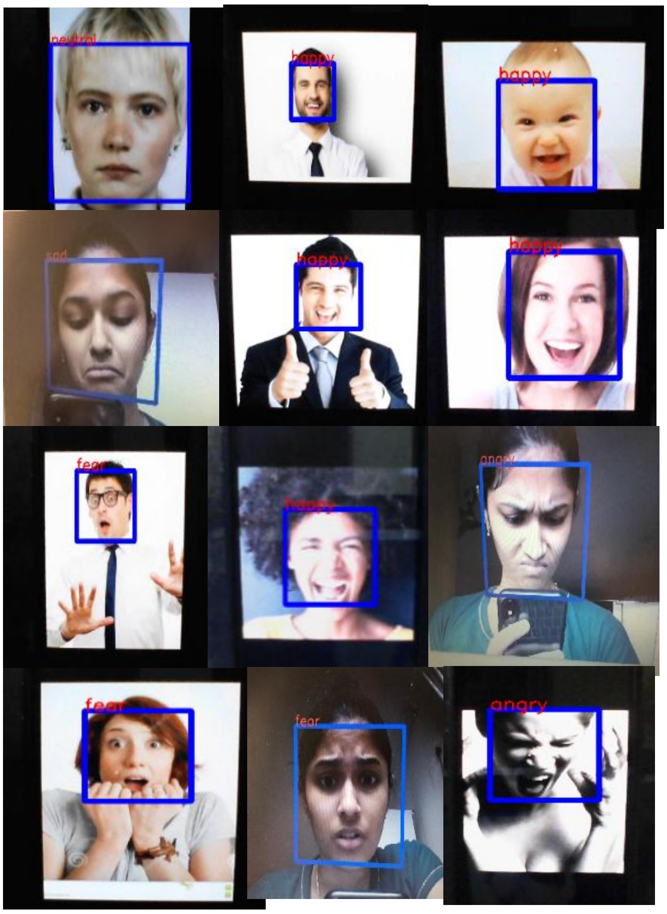
{e} {f}

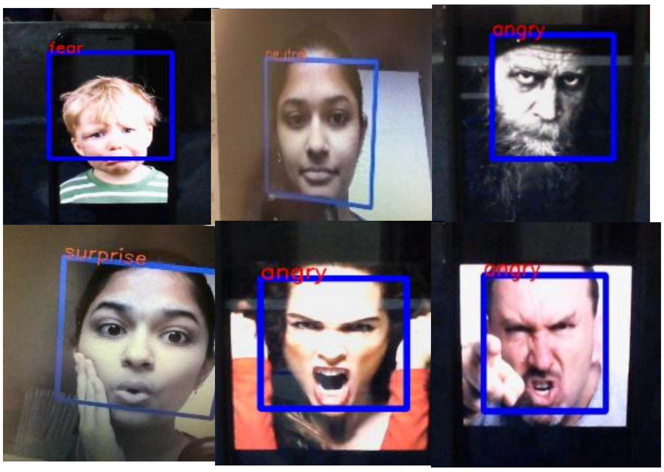
Fig.6. Real Time Face Expression (e) Angry (f) Neutral

Fig 4,5,6 shows experimental results of our Realtime facial expression recognition system of 6 individual face expressions - surprise, sad, happy, fear, angry, neutral respectively. However, we also considered few grey scale images and captured the emotion using webcam by simply showing the images to the webcam. And also taken some captured images from videos and given as inputs too. Here are some random outputs for our RTFER system.

Expression detection of random faces:







V. CONCLUSION:

In this paper, procedure to predict emotions of a person by processing the frames of video through various stages, such as pre-processing, face detection, and classifier using CNN is showcased and succeeded in detecting facial expression of a person in Realtime. Our RTFER accurately detects one's expressions and showcase their emotions at a rate of 85% on an average recognition rate of 85%.

REFERENCES

[1]. Bishop, C.M. Pattern Recognition and Machine Learning; Springer: New York, NY, USA, 2006.

[2]. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015, 521, 436. [CrossRef] [PubMed]

[3]. Coan, J.A.; Allen, J.J. Frontal EEG asymmetry as a moderator and mediator of emotion. Biol. Psychol. 2004, 67, 7–50. [CrossRef] [PubMed]

[4]. Zafeiriou, S.; Zhang, C.; Zhang, Z. A survey on face detection in the wild: Past, present and future. Comput. Vis. Image Underst. 2015, 138, 1–24. [CrossRef]

[5]. Perez, L.; Wang, J. The effectiveness of data augmentation in image classification using deep learning. arXiv 2017, arXiv:1712.04621.

[6]. Tian, Y.I.; Kanade, T.; Cohn, J.F. Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. 2001, 23, 97–115. [CrossRef]

[7]. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992; ACM: New York, NY, USA, 1992; pp. 144–152. [CrossRef]

[8]. Liu, W.; Wang, Z.; Liu, X.; Zeng, N.; Liu, Y.; Alsaadi, F.E. A survey of deep neural network architectures and their applications. Neurocomputing 2017, 234, 11–26. [CrossRef]

[9]. N.-H. Chang, Y.-H. Chien, H.-H. Chiang, W.-Y.Wang, and C.-C. Hsu, "A robot obstacle avoidance method using merged CNN framework," in Proc. Of the 2019 International Conference on Machine Learning and Cybernetics (ICMLC), Kobe, Japan, July 7-10, 2019.

[10]. S. Alizadeh and A. Fazel, "Convolutional neural networks for facial expression recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26, 2017.

[11]. Edwards, Jane, Henry J. Jackson, and Philippa E. Pattison. "Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review." Clinical psychology review22.6: 789-832, 2002.

[12]. S. Li and W. Deng, "Reliable crowdsourcing and deep locality preserving learning for unconstrained facial expression recognition," IEEE Trans. Image Process., vol. 28, no. 1, pp. 356-370, Jan. 2019.

[13]. Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski and Remigiusz J. Rak. "Emotion recognition using facial expressions". International Conference on Computational Science (ICCS), 12- 14 June, 2017.

[14]. M.Pantic and J. M. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," IEEE Trans. Systems, Man andCybernetics, vol. 34, no. 3, pp. 1449-1461, 2004.

[15]. Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. "The Computer Expression Recognition Toolbox (CERT)", Face and Gesture 2011, 21-25 March 2011.

[16]. Björn Schuller, Stephan Reiter, Ronald Müller, Marc Al-Hames, Manfred Lang, Gerhard Rigoll. "Speaker Independent Speech Emotion Recognition by Ensemble

Article Received: 16th October, 2020; Article Revised: 30th December, 2020; Article Accepted: 08th January, 2021

Classification", 2005 IEEE International Conference on Multimedia and Expo, 6-6 July 2005.

[17]. B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26, 2017

[18]. Potharaju, S.P., Sreedevi, M., Amiripalli, S.S "Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance". Clinical Epidemiology and Global Health. June 2019.

[19]. Potharaju, S.P., Sreedevi, M., Amiripalli, S.S. "An ensemble feature selection framework of sonar targets using symmetrical uncertainty and multi-layer perceptron (su-mlp)". Advances in Intelligent Systems and Computing. Jan 2019 pp.247-256

[20]. Vijay Kumar- G. Sreedevi-M., Bhargav- K and Mohan Krishna- P.-2018 Incremental mining of popular patterns from transactional databases. International journal of Engineering and Technology-7 –pp 636-641.

[21]. Vijay Kumar- G. Sreedevi-M. Vamsi Krishna-G and Sai Ram-N. 2018 Regular frequent crime pattern mining on crime datasets. International journal of Engineering and Technology-7 –pp 972-975.

[22]. G. Vijay Kumar, T. Krishna Chaitanya, M. Pratap, "Mining Popular Patterns from Multidimensional Database", Indian Journal of Science and Technology, Vol 9(17), DOI:10.17485/ijst/2016/v9i17/93106, May 2016.

[23]. Potharaju, S.P., Sreedevi, M. "A novel LtR and RtL framework for subset feature selection (reduction) for improving the classification accuracy". Advances in Intelligent Systems and Computing. 18 Dec 2018.

[24]. Kaggle,URL: https://www.kaggle.com/c/challengesin-representation-learning-facialexpression-recognition-challenge/data[Last accessed:Dec 2018]

[25]. License Link, URL: https://github.com/oarriaga/face_classification/blob/master/LICENSE.