2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) | 978-1-6654-2642-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/I-SMAC52330.2021.9640832

Survey Analysis on Facial Expression

1st M. Raju Yadav
M.Tech Student
Computer Science and Engineering
GRIET, Hyderabad, Telangana, India

2nd *Dr.*P.Chandra Sekhar Reddy
Professor
Computer Science and Engineering
GRIET, Hyderabad, Telangana, India

***Abstract*— With headways in machine and profound learning calculations, the vision of different significant genuine applications in PC vision turns into a chance. Facial opinion examination is one of the applications. Profound learning has raised face acknowledgment to the first spot on the list of moving investigation fields in the PC vision space. Profound learning-based FER models have as of late been tormented by an assortment of innovative issues like Under-fitting or over-fitting. Driven from the upper than realities, it presents a logical and exhaustive study on present status of-craftsmanship figuring procedures (datasets and calculations) that supply a response to the issues. It conjointly presents a scientific categorization of existing facial slant investigation manners by which quickly. Then, at that point, this examination sums up the latest novel machine and profound learning networks proposed by scientists that are explicitly created for facial recognizable proof upheld static film, just as their professionals and faults. At last, the open inquiries and examination challenges for the plan of a durable face acknowledgment framework are introduced in this investigation.**

Keywords— ***Recurrent Neural Network; rehashed Neural Network; convolution neural network; ImageNet; Ensemble; ResNet;Maxpooling; VGG16.***

I. Introduction

Facial expressions include smiles, sadness, anger, disgust, surprise, and fear. A smile on the human face expresses happiness and a curved-shaped eye. The sad expression is distinguished by a sense of lightness, which is typically expressed by raising crooked brows. Also, frown. Anger in the human face is associated with unpleasant and irritating situations. Smiles, sadness, anger, disgust, surprise, and fear are all examples of facial expressions. A human smile expresses happiness and a curved-shaped eye. The sad expression is characterized by a sense of lightness, which is typically expressed by raising crooked brows. Additionally, frown. Anger on the human face is associated with unpleasant and vexing situations. Dilated brows and thin, elongated eyelids characterize anger. The extraction and classification of features is a critical step in the FER process. Classification is another important process that categories the aforementioned expressions such as smile, sadness, anger, disgust, surprise, and fear. Understand the eyes, mouth, nose, brows, and other facial components, as well as feature extraction based on appearance [1].

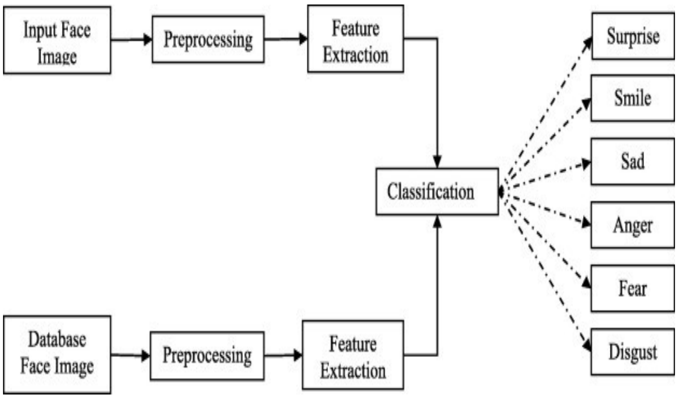

Fig. 1. Architecture of face expression recognition [22].

In this study, we primarily focused on face demeanor recognition using face-parsing components (FP). Given the disadvantage that different parts of the face contain different data measures for face appearance, and that the weighted element varies for different faces, a collaborative setup is intended to work out face demeanor exploitation parts that unit of estimation engaged in appearance exposure [2]. Based on the most recent realities, this paper provides a logical partner degreed comprehensive overview of the current state of-craftsmanship AI strategies that offer a comparative solution [3]. To see the value in the objective, it uses separated local based strategies for inward facial parts and worldwide techniques for external facial parts [4]. A series of fortunate events resulted in Profound Learning's astounding achievement in imaging applications. Unfortunately, this quality has resulted in pernicious applications such as photograph reasonable face trading of gatherings without their consent changes in position are processed and determined by sequentially tracking objects each time.

The sensor detects the nature of the object, and the object is visualized. The object in space is derived from the application of matching schemes to find the object, and its position in the plane is quantified. Image processing schemes are used to capture objects, and their dimensions are recorded and classified in 2D or 3D [5]. By incorporating various cutting-edge technologies, the system's navigability and usability can be improved. Highlighted text when scrolling with the mouse, and evaluate the layout structure of the website using background music and other similar features. These innovations have the potential to improve the screen reader. A cross-folding recurrent neural network is used to perform FER on film. The projected detail is composed of convolution layers that are then followed by a never-ending neural network that reflects the relationships between the facial film and the persistent organization of the volatile conditions that exist within the unity of the photo region by utilizing concentrate throughout the process. The model is being evaluated over time. Findings from exploratory research that are promising a unit of estimation was acquired in comparison to reformist systems [7]. We concentrate on neural convolution networks, RNN, and DNN, which are less precise than others are. After reviewing all of these documents, we conclude that the vgg16 architecture has a lower precision. I chose the VGG19 architecture with face analysis to improve accuracy. A crossbred Convolution-Recurrent Neural Network method was used for FER in film. The proposed detail consists of Convolution layers followed by a never-ending Neural Network (RNN) that the combined model concentrates the relationships between facial film and by exploitation he persistent organization the fleeting conditions that exist within the photos region unit pondered all through the order. The proposed model of half-and-half is upheld. When compared to reformist systems, promising exploratory outcomes were obtained [7]. We discovered that convolution neural networks, RNN, DNN VGG16, and RESNET architecture have lower accuracy after reviewing all of these papers. I chose the VGG19 architecture with face parsing for improved accuracy.

II. RELATED WORK

Identifying human emotions in photographs or videos has been the goal of facial emotion recognition research since its inception. Recent studies have attempted to recognize faces in photos or videos, but these methods did not use a network framework. Geometry-based approaches and feature-based approaches are the two main methods for extracting features from photos, which differ from feature recognition methods. Furthermore, the appearance is rapidly approaching. In the first scenario, the model focuses on restricting and tracking specified face criteria in order to train the model to classify based on relevant postures. For the purpose of classifying emotions from sequences, the authors proposed a model for tracking a group of points and they detected emotions extracted from feature set forms by way of transformation in only 117 reference points. Bitmaps are used to train the categorization model.

Convolution rehashed Neural Network framework utilizes Convolution layers and rehashed Neural Network (RNN). Relations at spans facial highlights are separated by this model thus the fleeting conditions are considered during the arrangement by abuse-rehashed network [9]. continuous technique referenced is that the Constructive Feed forward Neural Networks inside which include discovery is done by a second DCT (Discrete cos Transform) [10]. On the facial picture and grouping is finished utilizing a useful feed forward one secret layer neural organization. Another technique, Boosted Deep Belief Network, utilizes a combination of highlight selector and classifier in one system [11]. During this model, choices are conjointly tuned and are tip top to shape a more tasteful through a BTD-SFS strategy. Another Hybrid technique, Convolution rehashed Neural Network framework utilizes Convolution layers and rehashed Neural Network (RNN) [9]. Relations at spans facial highlights are separated by this model thus the transient conditions are considered during the characterization by abuse-rehashed network. continuous approach referenced is that the Constructive Feed forward Neural Networks inside which include recognition is done by a second DCT (Discrete cos Transform) [10]. On the facial picture and arrangement is finished utilizing a helpful feed forward one secret layer neural organization. Another system, Boosted Deep Belief Network, utilizes a combination of highlight selector and classifier in one structure [11]. During this model, choices are conjointly tuned and are tip top to shape a more tasteful through a BTD-SFS strategy.

Breuer and Kimmel utilized CNN inward portrayal procedures to realize a model discovered misuse shifted FER datasets, and unquestionably, the capability of organizations prepared on feeling discovery, across all datasets and FER-related errands [12]. Jung et al used two different types of CNN: The first focuses on global look choices from image groupings, while the second focuses on transient number related alternatives from fleeting facial milestone focuses [13]. These two estimation models were joined utilizing another combination procedure to work on the exhibition of facial element acknowledgment. Kahou et al. projected a half and half RNN-CNN system for engendering data over a grouping utilizing partner persistently esteemed covered up layer delineation. The creators fostered a full framework for the 2015 feeling Recognition among the Wild (EmotiW) Challenge, showing that a crossover CNN-RNN style for facial highlights investigation can outflank a formerly utilized CNN technique for collection utilizing fleeting averaging. [14].

The VGG-16 layer stores a 224 by 224 RGB image as data. There are a lot of convolutional (conv.) layers in the image, and the channels are set to catch the notions of left/right, up/down, and Center with a little open field of 33 (the smallest size to capture the concepts of left/right, up/down, and Center). In one design, it also uses an 11-channel convolutional filter, which can be viewed as a simple switch in the information channels (trailed by non-linearity). The convolutional step is set to one pixel, and the convolutional spatial cushioning is set to zero. The layer input is designed with the goal of saving the spatial goal after convolution, for example, one pixel for 33 convolutions. Levels (Figure 2). After a portion of the convolutional layers, five max-pooling layers perform spatial pooling (not all the conv. layers are trailed by max pooling). Step 2 is used to maximise pooling across a 22-pixel window. VGG16 had been working on NVIDIATitan Black GPUs for quite some time. VGG16 had been working on NVIDIATitan Black GPUs for quite some time.

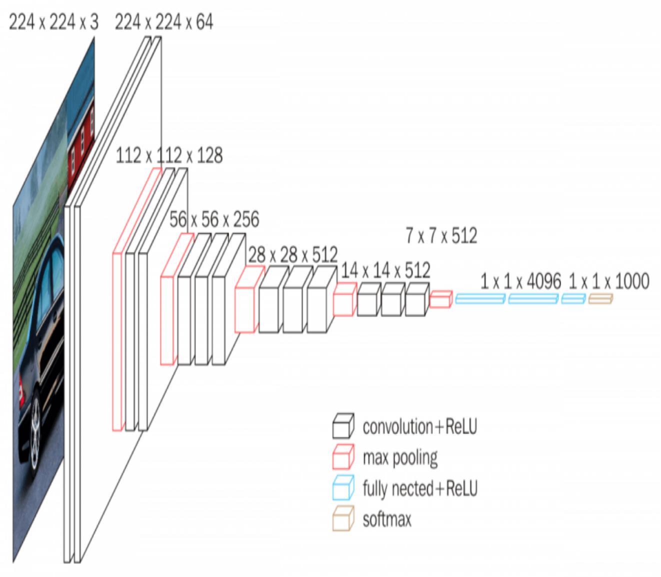

Fig. 2. VGG-16 Architecture diagram [15].

As information, our VGG16 is taking care of a 48x48 RGB image. By removing the typical RGB from each pixel, we achieve phenomenal preprocessing. To handle the image, a convolutional layer stack with 3x3 channels is used. In one of the arrangements, we also use 1x1-convolution channels, which can be thought of as a direct redesign of the information channels (saw through way of method of non-linearity). Because the convolution step is set at one pixel, and because the convolutional layer data is padded spatially, the spatial assurance is preserved after convolution. The spatial clustering procedure is completed by inspecting some of the convolution layers through the lens of five maximum clustering layers (now not all convolution layers are observed through manner of way of maximum clustering). In a two-pixel window, stride2 collects the most information. Three related (FC) layers detect a pile of convolutional layers: the first has 4096 channels each, the third performs the 7-way ILSVRC type and therefore incorporates seven channels, and the last has 4096 channels each (one for each class). The ideal reorientation. All companies use a similar structure for the associated tiers. Non-linearity rectification is enabled on all hidden layers (ReLU). VGG16 is made up of 16 weight layers, 13 folding layers with a 3x3 clean out period, and 3 associated layers. Stride and padding are set to a minimum of one pixel for all convolution layers. Each convolution layer is divided into five groups, with each group being observed using a maxpooling layer (Figure 2). Step 2 performs maxpooling in a 2x2 window. The number of filters with within the convolution layer company starts at 6 4 with within the primary company and increases by a factor of after each maxpooling layer until it reaches 512[15]. Keras VGG16 is used as the implementation.

ResNet 50 is a cutting-edge architecture for convolutional neural networks. Its architecture is similar to that of networks like VGG-16, but it adds the capability of identity mapping (Figure3).

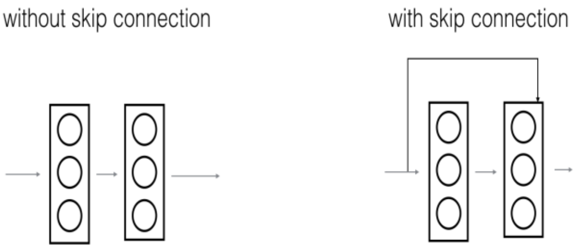

Fig. 3. ResNet residual block diagram with skip connection [15].

Leftover Networks, also known as Reset, are a type of neural organization that serves as the foundation for some PC vision tasks. This model won the Magnet Challenge in 2015.The ability to prepare profound neural organizations was Reset's key breakthrough. With over 150 layers. Prior to Reset, profound neural organization preparation was difficult due to the issue of blurring gradients. Reset was the first to suggest the idea of a jump affiliation. The leap association is depicted in the outline below. The left image shows collapsed layers stacked consistently; the right image shows similar collapsed layers stacked, but we are currently adding the first contribution to the yield of the convolution square. It is what it is: a leap association. Hop comes to work in this environment for two reasons: They reduce the slope blurring issue by allowing this other angle alternate way. Essentially equal to, if not worse than, the base level [15].

III. DATA SET

The Extended Cohn-Kanade Dataset (CK+) [16]: CK+ contains 594 video groupings on each show and non-presented (unconstrained) feelings, just as different kinds of data. The 123 members range in age from eighteen to thirty years of age, with most of the World Health Organization unit of estimation being female. Picture groupings can be taken apart for activity units and original feelings. It gives conventions and benchmark results to facial element following, AUs, and feeling acknowledgment. The photos have part goals of 640 × 480 and 640 × 490 with 8-bit exactness for dim scale esteems.

JAFFE [Japanese Female Facial Expressions] [17]: The JAFFE information contains 213 photos of ten unique female Japanese models showing seven face feelings (six fundamental facial feelings and one unbiased). Each picture was given a score dependent on six emotive words used to misuse sixty Japanese individuals. Every facial picture's essential size is 256 pixels by 256 pixels.

ImageNet is a data set containing over 15 million labelled high-goal images organized into approximately 22,000 classes. The images were compiled from the internet and labelled by people who volunteered on Amazon's Mechanical Turk. Since 2010, the Pascal Visual Object Challenge has included a yearly competition known as the ImageNet Large-Scale Visual Recognition Challenge. ILSVRC makes use of a subset of ImageNet, with approximately 1200 images in each of the 1200 classifications. There are approximately 1.3 million preparing images, 52,000 approval images, and 152,000 testing images in total. ImageNet contains images of varying sizes and objectives. As a result, the images' goal was reduced to 256x256. A rectangular image is rescaled, and the resulting image is edited to remove the focal 256x256 fix [15].

IV. Result And Analysis

The accuracy of the models in the test data for different spans of the noticed sequence is shown in Table 1 and Figure 4. The precision pattern is mostly up, demonstrating that giving the model a fleeting setting improves characterization exactness. Except for the two lip models, who have the pattern up to the most extreme number of edges, everyone has 45 edges. The edge model clearly performs the worst, whereas the CNN and RNN Lips lip models perform exceptionally well. In higher edges, the RNN lip model marginally outperforms the CNN lip model. The main point to emphasize is that in outlines greater than 25, there is a reasonable and supported expansion in lip model accuracy when contrasted with the relating face models.

TABLE I. The Models' Data Correctness Was Examined Throughout A Range Of Frame Values [18].

Frames	**CNN**	**CNN Lip**	**RNN**	**RNN Lip**	**Frame**
5	0.429	0.446	0.423	0.444	0.413
10	0.413	0.448	0.443	0.452	0.421
15	0.458	0.458	0.458	0.442	0.431
20	0.459	0.442	0.462	0.475	0.437
25	0.465	0.509	0.452	0.470	0.439
30	0.475	0.516	0.465	0.513	0.441
35	**0.479**	0.485	0.476	0.530	0.449
40	0.459	0.509	**0.485**	0.505	0.450
45	0.452	0.479	0.473	0.533	**0.457**
50	0.465	0.492	0.479	0.519	0.449
55	0.466	0.508	0.482	0.538	0.444
60	0.462	0.544	0.470	**0.594**	0.447
65	0.448	**0.569**	0.482	0.563	0.454

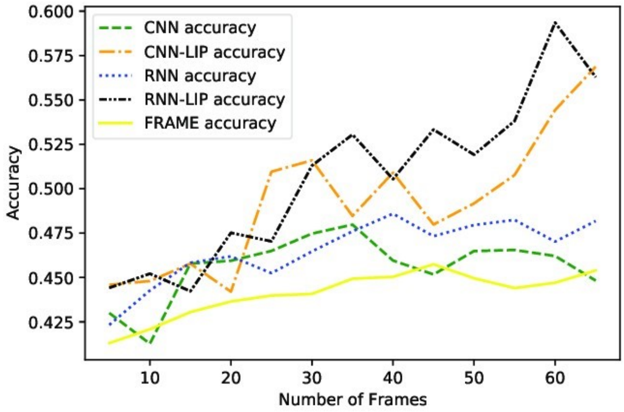

Fig. 4. Shows a plot of the models' accuracy on test data [18].

Taking a gander at the exactness distinction between the lip models and their comparing face models, as displayed in Figure 5, we can see an increment in precision from 25 housings, which seems, by all accounts, to be an immediate vertical example, similarly as the model count increments. When applied to the data in Figure 5, the direct relapse model yields a measurably critical incline of 0.0015, addressing the increase in accuracy for each additional casing in the data. The increased number of frames, as well as the upper limit of the precision metric, raises concerns about the persistence of basic emotional expressions in spontaneous conversations. The available data's precision gains, on the other hand, appear to increase linearly over the duration of the input sequence, with positive values starting at sequence. While crediting understandings to the activity of a discovery framework, for example, a neural organization is a perilous endeavor; the justification behind lip models does not deteriorate and develops straightly: joint-related data expands the helpfulness of the face. Assuming that the articulatory features of speaking subjects do not carry affect information, they cause the network to become articulation-invariant. Lip models can classify better with a larger number of frames because of providing more information at the start [18].

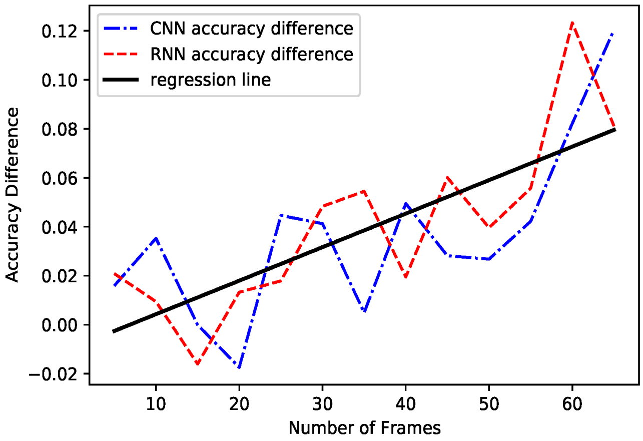

Fig. 5. Differences in the accuracy of the lip-only model and the facial-only model based on the testing data [18].

Table 1 shows the outcomes of the Waggle dataset's SVM (standard), VGG16, ResNet50, and co-learning models. The precision of our gauge SVM was 32 percent, while the precisions of VGG16 and ResNet50 were 59.2 and 65.1 percent, respectively. The model includes character derivation layers, which can help it outperform VGG16 in terms of exactness, accuracy, and recovery. The ensemble-learning model, which combined VGG16 and ResNet50, achieved 67.2 percent accuracy, 2.1 percent higher than either VGG16 or ResNet50 alone. The DEF dataset outperforms the Waggle dataset in terms of overall accuracy, as well as accuracy and retrieval. SVM had a precision of 38 percent, while VGG 16 and ResNet50 had precisions of 70.8 percent and 74 percent, respectively. With an accuracy of 75.8 percent, the ensemble approach outperformed individual deep learning models. KEF and Waggle both use the same four models. Surprisingly, despite the fact that the dataset was much smaller, each of the four models outperformed wrangle in the DEF. We believe this is due to the DEF dataset's design and consistency in terms of subject positions and the number of models for each point and feeling. The images in the DEF dataset are also of higher quality. Regardless of the larger picture goal, there were instances of text being displayed in the image's association with the Waggle dataset [15].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP= True Positive, TN= True Negative FP= False Positive, FN= False Negative.

TABLE II. PERFORMANCE OF THE KDEF DATASET (ACCURACY, PRECISION, AND RECALL) FOR SVM, VGG-16, RESNET50, AND ENSEMBLE LEARNING MODELS[15].

	Accuracy	**Precision**	**Recall**
SVM	**37.9%**	**50.1%**	**54.9%**
VGG-16	**71.4%**	**81.9%**	**79.4%**
ResNet50	**73.8%**	**83.3%**	**80.7%**
Ensemble	**75.8%**	**85.0%**	**82.3%**

VI. CONCLUSION

In this we observed that accuracy of the data percentage is less with the CNN, RNN, DNN and VGG16 architecture of parsed images and by using VGG19 architecture we can improve the accuracy percentage along with dense layer and we can demonstrate that our technique outperformed the cutting-edge strategies. From all the Architectures VGG16 have achieved 71.4% is very less when compared with ensemble, but very high when compared with CNN, RNN and DNN and with parsed image VGG19 on facial expression we can improve accuracy of facial expression.

REFERENCES

[1] I.Michael Revina, W.R. Sam Emmanuel, A Survey on Human Face Expression Recognition Techniques, Journal of King Saud University - Computer and Information Sciences, Volume 33, Issue 6, 2021, Pages 619-628.

[2] Y. Lv, Z. Feng and C. Xu, "Facial expression recognition via deep learning," 2014 International Conference on Smart Computing, Hong Kong, China, 2014, pp. 303-308.

[3] Keyur Patel, Dev Mehta, Chinmay Mistry , Rajesh Gupta, Sudeep Tanwar, Neeraj Kumar, Mamoun Alazab "Facial sentiment analysis using ai techniques: state-of-the-art, taxonomies, and challenges",2017.

[4] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, Lu Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5654-5663

[5] Sungheetha, Akey, and Rajesh Sharma. "3D Image Processing using Machine Learning based Input Processing for Man-Machine Interaction." Journal of Innovative Image Processing (JIIP) 3, no. 01 (2021): 1-6.

[6] Manoharan, J. S. (2019), "A smart image processing algorithm for text recognition, information extraction and vocalization for the visually challenged", Journal of Innovative Image Processing, 1(1): 30 – 38.

[7] Xiaoming Zhao, Shiqing Zhang. (2016) A Review on Facial Expression Recognition: Feature Extraction and Classification. IETE Technical Review 33:5, pages 505-517.

[8] Deepak Kumar Jain, Pourya Shamsolmoali, Paramjit Sehdev, Extended deep neural network for facial emotion recognition,Pattern Recognition Letters, Volume 120, 2019, Pages 69-74.

[9] Neha Jain, Shishir Kumar, Amit Kumar, Pourya Shamsolmoali, Masoumeh Zareapoor, Hybrid deep neural networks for face emotion recognition, Pattern Recognition Letters. 115(2018). 101-106. https://.org/10.1016/j.patrec.2018.04.010.

[10] L. Ma and K. Khorasani, Facial expression recognition using constructive feedforward neural networks, in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 34(2004).1588-1595.10.1109/TSMCB.2004.825930.

[11] P. Liu, S. Han, Z. Meng and Y. Tong, Facial Expression Recognition via a Boosted Deep Belief Net- work, 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus. 2014.1805-1812 . 10.1109/CVPR.2014.233.

[12] Breuer, R.; Kimmel, R. A deep learning perspective on the origin of facial expressions. arXiv 2017, arXiv:1705.01842 .

[13] Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint fine-tuning in deep neural networks for facial expression recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–12 December 2015; pp. 2983–2991.

[14] Ng, H.W.; Nguyen, V.D.; Vonikakis, V.; Winkler, S. Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 17th ACM International Conference on Multimodal Interaction, Emotion Recognition in the Wild Challenge, Seattle, WA, USA, 9–13 November 2015; pp. 1–7.

[15] Poonam Dhankhar (2019) "ResNet-50 and VGG-16 for recognizing Facial Emotions", International Journal of Innovations in Engineering and Technology (IJIET)", Volume 13, ISSN: 2319-1058.

[16] Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.

[17] Lyons, M.J.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with Gabor wave. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 200–205.

[18] Bursic, Sathya & Boccignone, Giuseppe & Ferrara, Alfio & D'Amelio, Alessandro & Lanzarotti, Raffaella. (2020). Improving the Accuracy of Automatic Facial Expression Recognition in Speaking Subjects with Deep Learning. Applied Sciences. 10. 4002. 10.3390/app10114002.

[19] Smys, S., and Jennifer S. Raj. "Assessment of Fire Risk and Forest Fires in Rural Areas Using Long Range Technology." Journal of Electronics 2, no. 01 (2020): 38-48.

[20] Manoharan, J. S. (2017), "Super-resolution reconstruction model using Compressive Sensing and Deep Learning", International Journal for research and development in Technology", 7(4): 884 – 889.

[21] Joe, Mr C. Vijesh, and Jennifer S. Raj. "Location-based Orientation Context Dependent Recommender System for Users." Journal of trends in Computer Science and Smart technology (TCSST) 3, no. 01 (2021): 14-23.

[22] Michael Revina, W.R. Sam Emmanuel, "A Survey on Human Face Expression Recognition Techniques", Journal of King Saud University - Computer and Information Sciences", Volume 33, Issue 6, July 2021, Pages 619-628.