A Survey on Facial Expression Recognition using Deep Learning and Explainable Artificial Intelligence

[1]Madhapura Ravikumar Madhan Kumar, [2]Dr.B G Prasad,

[1]Student, [2]Professor,
[1]Department of Computer Science and Engineering,
[1]B.M.S. College of Engineering, Bengaluru, India

***Abstract*—Facial expressions are nonverbal means of expressing emotions through body movements, eye contact, and gestures without speaking. The way a person expresses his or her feelings is not only a window into the sensitivity of that person, but it also provides insight into his or her mental outlook. This paper describes the various Deep Learning and Machine Learning Models that are employed for the recognition of facial expressions. Different types of classification algorithms are described in this study, namely Support Vector Machines and K-Nearest Neighbors. Various neural networks such as Attentional Neural Network, and Convolutional Neural are implemented and compared accuracy on a FER2013, CK+ dataset. Explainable Artificial Intelligence (XAI) is a topic that has grown considerably. As the use of machine learning has expanded, particularly deep learning, highly accurate models have been developed, but they lack of the explanations and interpretations that would make them useful. With Explainable AI's architecture and tools, machine learning predictions are easier to understand and interpret.**

***Index Terms*—Convolutional Neural Network, Support Vector Machine, K-nearest Neighbors, Explainable Artificial Intelligence, principal component Analysis, Negative matrix factorization, Local ternary patterns, Dynamic LTP, Facial Expression Recognition, Active Appearance Model, Artificial Neuro-Fuzzy Inference System.**

I. INTRODUCTION

Deep learning is a method of machine learning that makes use of artificial neural networks. As a result of its development and improvements, Deep Learning is emerging as the fastest-growing field in machine learning. Neural Networks are used in many projects that are related to Artificial Intelligence for learning, analyzing, processing, and training data.

Artificial neural networks are utilized in deep learning algorithms to process large amounts of data, similar to the human brain. In deep learning, a task would be repeated, similar to the way humans learn from experience. Several real-time applications were made possible by neural networks since they had the capability of learning without any hardcoded features when trained on supervised data. The underlying structures of most models built using machine learning and deep learning are complex, non-linear, and difficult to understand and explain to laypersons. In the field of artificial intelligence, explanation plays a crucial role. The Explainable AI architecture and tools make machine learning predictions easier to understand and interpret. Despite the advantages CNN offers in terms of precision, they are considered unreliable and difficult to interpret because of their black-box nature. As a result, Explainable Artificial Intelligence models such as Local Interpretable Model-agnostic Explanations, Shapley Additive explanations, Layer-wise Relevance Propagation, and Gradient-Class Activation Mapping are employed. Essentially, these methods focus on identifying and highlighting the crucial region elements of the input image that contribute to the classification.

As technology develops, we are increasingly relying on human-machine interactions. One of the most important parts of human emotion recognition that may be utilized to detect interpersonal relationships is facial expression. Nonverbal clues like gestures, facial expressions, and involuntary language can be used to convey general intentions and feelings. Humans can easily understand the expression of a face, but recognizing the facial expression is difficult for machines. There are various applications for facial expression recognition including teaching fields, intelligent transportation, and human-computer interaction. It is also used in healthcare systems to detect and improve the emotional states of people by using emotion analysis.

The following section gives an overview of various Algorithms that have been developed to recognize Facial Expressions with more emphasis on their algorithms. It covers several essential techniques for achieving better performance and efficiency.

II. LITERATURE REVIEW

Kavitha. V [1] has described how the impact of dropout on a network's performance may be empirically measured, as well as how dropout might improve validation accuracy. The dropout regularization method, described by Srivastava et al (2014), prevents neural network models from overfitting. The classifier is trained using AdaBoost and Haar feature methods. Adaboost was used to remove some dispensable details, by focusing solely on object-like regions and rejecting the uninteresting background. The focus of applying Haar-like qualities is to identify the difference in contrast levels among pixels. In the final step of edge detection, the image is given as input to the convolutional neural network. The training time is reduced by using maximum pooling to reduce the dimension of the extracted features. Dropout was introduced at the first and second convolution layers and on fully connected layers. In Feedforward Deep Convolutional Neural Networks (FDCNN), Rectified Linear Units are used throughout the hierarchy in all layers as activation functions. The dropout technique is used to address the issue of over-fitting. Training accuracy is affected by variations in dropout rates. Dropout values directly affect dense layer classification. The researcher has computed the difference in validation accuracy with and without dropout. The advantages of applying dropout to a neural network are that the training time as well as the training speed increase as the number of fully-connected layers increases. Validation Accuracy of 58.8% is achieved with dropout and 55.6 % without it.

CNN model takes a picture as input, analyses it, and categorizes it. The CNN model is achieving high accuracy by using more representative models like VGG, GoogleLeNet, and ResNet. The process of using these methods may result in network overfitting during the training. The large structure of the network makes it difficult to train, which will result in low training efficiency. To achieve high accuracy and a lightweight facial expression detection model, Zhou Yue [2] utilized the ExpressionNet convolutional neural network architecture. The ExpressionNet convolutional neural network is a lightweight neural network with a very modular architecture. Caffe, a deep learning framework, was used to train ExpressionNet on the face expression dataset. The highly modular ExpressionNet network is a critical component of the Reducev2 module. Reducev2 is made up of two layers: a dimensional layer and a sampling layer, both of which contain a ReLU active layer. In the ExpressionNet architecture, ReduceV2 modules are layered in 9 levels. In addition to each module, there is a BN layer as well as a ReLU activation function. The last layer of the ExpressionNet architecture utilizes global average pooling instead of fully connected layers to avoid overfitting and achieve a lightweight model. Various methods like AlexNet, ShallowNet, VGG, and ExpressionNet have been applied to the FER2013 dataset for Facial Expression Recognition. Accuracy and model size has been compared. ExpressionNet has taken about 426 seconds to identify all the emotions and this model is having a 15 MB size and an accuracy of 68.4% with fewer parameters. Even though the ExpressionNet Model is slightly larger than ShallowNet, it outperformed all the other methods in terms of accuracy. The advantages of using Expression Net have effectively improved network performance and it also helps in controlling the size of the model.

Rohit Pathar's[3] team has built a model for recognizing various human facial expressions in real-time by capturing images using a webcam. A shallow deep neural network is used to accurately identify human facial emotions. The model makes use of a recently released swish activation feature in the entirely connected layer, which improves its performance and distinguishes it. It is also shown over a range of convolution and max-pooling layer depths. Few parameters and filters are also changed in the network to boost accuracy from 48% to 90%. The shallow model is composed of a single convolution layer which achieved an accuracy of 45.72% on training data and an accuracy of 58.02% on validation data after around 14 epochs. A loss of 1.4191 was observed while training and a loss of 1.3514 were observed on validation data. The deep network which consists of eight convolution layers achieved an accuracy of 92.89% on training data and an accuracy of 90.01% on validation data after around 14 epochs. Also, the loss was 0.1782 on training and the loss was 0.2546 on validation data. When both the models are evaluated on basis of their training and validation accuracy obtained, the Deep Neural Network has outperformed the Shallow model. The limitations of their study are that there is a relatively small number of images for certain emotions, such as disgust, in the FER2013 dataset which leads to average model performance in identifying disgust emotion, and also dataset has been tweaked to make it suitable for recognizing disgust emotion.

Many models like neural networks, K-Nearest Neighbors, Support Vector Machines, and smaller networks were also employed to recognize facial expressions. As part of pre-processing Gabor Filter was applied along with Local Binary pattern operators to the dataset to extract features. By using the Gabor filter, the training model can extract necessary features and important patterns present in the data. Hypotheses indicate the increase in the accuracy of Recognition with the use of the Log Gabor filter and the time taken to process the image with the help of the Log Gabor filter is seen to be slightly higher when compared to the Gabor filter. The Local Binary Pattern (LBP) labels all pixels present in the image with the specific threshold considering the neighboring pixels with respect to each pixel in focus and generates the result in a binary number. The classifier distinguishes different classes of images based on the features and patterns present in each class of the image. KNN is used as an initial classifier in some models, where the output of the KNN algorithm is fed as input for a different algorithm like the Hidden Markov Model for further classification. E.Kodhai [4] compared the accuracy of KNN, Hidden Markov, SVM, Deep CNN, and Shallow CNN. The shallow CNN is comprised of a single convolutional layer followed by three fully connected layers. On the other hand, Deep CNN has comprised of higher layers in this case 8 convolutional neural networks are used that outperform shallow CNN in terms of performance due to better feature extraction Capabilities. Deep CNN attains an accuracy of 89% which is higher than the accuracy of 45% attained by the shallow CNN. In terms of precision, the support vector machine outperforms the K-Nearest Neighbor algorithm. The Convolutional Neural Network (CNN) is preferred because of its accuracy over other approaches such as SVM, KNN, and Gaussian Process Verifier for the Recognition of Facial Expression.

Jiequan Li [5] has proposed a system for automating facial expression recognition. It uses row images as input where the first step of face identification and recognition is performed. Opensource Faint, a Java framework for face identification and recognition, includes the face detection function. Java Native Interface (JNI) is used, allowing Faint to call the C++ implemented OpenCV face detection program directly from Java. As part of OpenCV's face detection algorithms, Haar-like features and the Adaptive AdaBoost algorithm are used to determine classification. The difference between a total dark region and a total light region is used to quantify the existence of oriented contrasts between regions of an image. The Faint project also integrates the face recognition function. The Face Detection function accepts the face image as input, which is fed to a dimensionality reduction algorithm called Principal Component Analysis (PCA) to extract the necessary facial features by reducing the dimensions of the input image. In their proposed system the PCA and the negative matrix factorization (NMF) algorithm are compared. Following this, K-Nearest Neighbor (KNN) algorithm is used to determine the appropriate facial expression. Users can choose between PCA or NMA methods and can specify the number of neighbors (k) in the KNN classification algorithm. The Principal Component Analysis method is an optimal linear dimensionality reduction strategy when the classification algorithm is applied to the reconstruction of K- Nearest Neighbor mean square error (MSE). In PCA, solving the eigenvalue problem for the covariance matrix of data leads to the first principal component vector which is for the largest eigenvalue, and the second principal component vector for the next largest eigenvalue. The non-negative matrix factorization (NMF) approach is a way of obtaining a non-negative representation of data. The Taiwanese Facial Expression Image Database (TFEID 2008) is used as a testing database that contains about 40 images on each expression category and the Indian face database comprises images of 40 distinct individuals each giving eleven different expressions to test the face recognition system employing Faint open source. The test focused on identifying

different types of facial expressions such as neutral, surprise, happiness, sadness, and disgust. For one test, a random selection of half of the images from each category is chosen, yielding about 20 images from each expression. As a test set, the remaining 120 images will be used. Training with the NFM method is usually long, while the PCA method is much faster. To test the KNN algorithm, set the parameter K to an odd value and change the value from 1 to 17. Comparing NFM and PCA recognition rates for each facial expression, both methods are around 75%, but PCA outperforms NFM in categories such as "Surprise" and "Sadness & Disgust." However, NFM has a higher Recognition rate for "Neutral" and "Happiness". In the NFM, reducing data redundancy is an advantage, which makes it more suitable for use with large, static datasets, and in the PCA, a short training time makes it more suitable for use with a dynamic dataset.

A dimensional emotion model based on the Valence-Arousal scale has been demonstrated by Shuang Liu's team [6]. CNN networks are used to identify the valence dimension of facial expressions, resulting in nine types of classifications. The probability of a network output's valence dimension equals the fusion of its valence value and its corresponding probability. The CNN network structure comprises 9 hidden layers followed by single input and output layer. Grayscale images of 48*48 facial expressions are used in the input layer. To estimate Local feature perception and sparse features of facial expression images, four convolution layers and three pooling layers are used, resulting in 64 feature maps of 6*6 pixels each. The full connection layer uses the feature map as input for local feature integration so that the network can learn from it. In addition to combining L2 regularization with the dropout technique, the Adam algorithm is used to optimize CNN network model performance. The facial emotion recognition system designed using the CNN architecture achieves a Root Mean Square Error value of 0.0857±0.0064. Although the model can accurately distinguish the emotional categories (sadness, serenity, and happiness), further improvement is needed since there are not enough images of facial expressions with valence dimensions.

Khadija Lekdioui [7] has described a technique to recognize facial expressions based on the Intraface algorithm where the whole face is decomposed into ROIs. These ROIs are then resized and partitioned into blocks in the preprocessing stage, before being used in feature extraction to generate face feature descriptors. The Intraface algorithm detects 49 landmarks around the eyebrows, mouth, and nose using the Supervised Descent Method. The supervised Descent Method generalizes better to untrained situations because it is a nonparametric model. As part of the training process, SDMs learn some generic descent directions. SDM minimizes the NLS objective function by using the learned descent directions. The feature is extracted from each ROI using a local binary pattern, a compound local binary pattern, a local ternary pattern, and a dynamic LTP. Local Binary Patterns encrypt image pixels into a binary value string to analyze textures. CLBP operator adds extra bits to the original LBP code to represent the magnitude of gray value differences between centers and neighbors. In addition to the Local Binary Pattern (LBP), the Local Ternary Pattern (LTP) exists as a subset of the Local Binary Pattern (LBP). Depending on the distance between neighboring pixels and the central pixel, the Local Ternary Pattern (LTP) can take three values.

Formula 1	Formula 2	Formula 3	Formula 4	Formula 5
$t = p_c \mathrm{x} \delta$	$t = \frac{\sum_{i=0}^{N-1} \sqrt{P_i}}{N}$	$t = \frac{\sum_{i=0}^{N-1} P_i}{N}$	$t = \frac{\sqrt{\sum_{i=0}^{N-1} P_i}}{N}$	$t = \sqrt{\sum_{i=0}^{N-1} \frac{p_i}{N}}$

Table 1

Table 1 shows the formula based on the LTP operator where pc is the central pixel value, pi is the ith Neighbor pixel value and N represents the number of neighbor pixels (N=8). δ is a scaling factor (source- paper [7]). Based on the obtained feature vector of the face image, multiclass support vector machines are then used to perform the recognition task. CK and Feed datasets are used to perform facial expression recognition. F-score is used to calculate the recognition rate. A comparison of three ROI face decompositions is carried out with different block sizes and block numbers. The first decomposition uses the entire face as one ROI, the second uses six ROIs, and the third uses seven ROIs. Descriptors are also evaluated along with the LTP parameter(t) by applying all descriptors to all the decompositions. Using dynamic LTP threshold formula 1 with δ =0.02 and threshold formula 5, the seven ROIs with face decomposition were able to achieve a 94.11% recognition rate for the CK dataset and 87.57% recognition rate for the FEED dataset, outperforming the other ROIs.

A system for automating facial expression recognition was described by Anagha S. Dhavalikar[8]. There are three modules in the system which includes detection of a face, feature extraction, and facial expression recognition. The first step in detecting faces begins with determining the color of skin with the YCbCr model, compensating for lighting, and performing morphological operations on the image. The most basic morphological operations in image processing are dilation and erosion. Dilation increases the pixels at object boundaries in images, while erosion reduces those pixels. The Active Appearance Model (AAM) uses the detected Face points to extract the features of the face. A first-stage face detection test has been conducted using 105 image samples and 95 images have been correctly identified. In the form of a data file, the Active Appearance Model locates the facial points on each image and stores the relative x-y coordinates of those points. Although Euclidean distance may benefit images that are static and offers a Recognition rate that ranges from 90% to 95%. To improve image uncertainty in real-time, an Artificial Neuro-Fuzzy Inference System was developed. The ANFIS model produced an accurate recognition rate of around 100% when trained on large amounts of samples.

The below table 2 summarizes the various Deep Learning methods and algorithms used in the field of Recognizing facial expressions by various researchers on different datasets, along with their accuracy results.

Related work	Algorithm	Dataset used	Accuracy %
C.Szeged [9]	GoogLeNet CNN (22 layers)	FER-2013	63
Naveen kumar [11]	VGG16 CNN (16 layers)	FER-2013	68.2
He, Kaiming [10]	ResNet CNN (152 layers)	FER-2013	71
Kumar, Kumar&san yal [12]	CNN	FERC-2013	90
Kavitha, Dr.Chetan [1]	FDCNN	FER-2013	76.6
Zhou Yue [2]	Expression Net CNN	FER-2013	68.4
Rohit Pathar[3]	Shallow CNN(1 Layer)	FER-2013	45.72
	Deep CNN(8 Layers)	FER-2013	92.89
JiequanLi [5]	PCA& NMA	TFEID	75
Khadija Lekdioui[7]	Dynamic LTP	CK+	94
	Dynamic LTP	FEED	87.57
Kulkarni,Baagal	Gabor	FACES	82
	Log Gabor	FACES	87
Shan,Guo,Lu&Bie	KNN	JAFFE	65.11
E.kodhai,2020[4]	Support Vector Machine	FER-2013	85
	KNN	FER-2013	82.87
	KNN	CK+	77.27
Minaee &Abdolrashidi[14]	Attentional CNN	FER-2013	70

Table 2

III. DATASET INFORMATION

- **FER-2013:** The FER2013 has presented as a database in 2013 ICML challenges for representative learning enabling one to visualize the image sections that contribute to the classification. It has been prepared by Aaron Courville and Pierre-Luc Carrier. The FER2013 is an image search database that Google collected automatically from its Google image-based search API on a large scale. The dataset contains grayscale pictures of 48x48 pixels of faces. All images have the same amount of space occupied by faces because they are automatically registered. Images are grouped based on the emotion expressed by their faces. The dataset is split into three parts: 28,709 pictures of training data, 3589 pictures of testing data, and 3589 pictures of validation data. There are 7 categories in the dataset namely, anger, happiness, sadness, surprise, disgust, neutral, and fear.
- **JAFFE:** JAFFE dataset includes a total of 219 sample images of Japanese individuals. Each image is labeled with six basic facial expressions (happy, fear, surprise, anger, disgust, and sad).
- **CK+:** CK+ dataset contains 593 samples from 123 different expressions. Each sample consists of 10 to 60 frames and are stored in array of pixels. On each image there are seven basic facial expressions (Sadness, Fear, Disgust, Happy, Anger, Contempt, and Surprise).
- **EmotioNet:** The dataset includes 950,000 images from the Internet with emotion keywords, which are set as AUs and then automatically annotated and classified into 23 basic or compound emotion categories.

IV. CONCLUSION

Facial Emotion Expressions (FER) are extremely valuable in a variety of real-life situations since they provide crucial information about a person. To identify a facial expression, many different methods and techniques are used. The paper presented the datasets and algorithms that are used for facial expression recognition and compared the accuracy of the models. This study also explained the impact of dropout on performance and discussed how it might improve the accuracy of validation.The Explainable AI approach, along with deep learning, helps make machine learning predictions more understandable and is effective in illuminating them.

REFERENCES

[1] Chetan, H. "Performance Dependency of Facial Emotion Recognition System on Dropout and Learning Rate." In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India,2020, pp. 71-81.

[2] Yue, Zhou, Feng Yanyan, Zeng Shangyou, and Pan Bing. "Facial expression recognition based on convolutional neural network." In 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2019, pp. 410-413.

[3] Pathar, Rohit, Abhishek Adivarekar, Arti Mishra, and Anushree Deshmukh. "Human emotion recognition using convolutional neural network in real time." In 2019 1st International Conference on Innovations in Information and Communication Technology ICIICT), Chennai, India. pp. 1-7. 2019.

[4] Kodhai, E., A. Pooveswari, P. Sharmila, and N. Ramiya."Literature Review on Emotion Recognition System." In 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2020, pp. 1-4.

[5] Li, Jiequan, and M. Oussalah. "Automatic face emotion recognition system." In 2010 IEEE 9th International Conference on Cyberntic Intelligent Systems, Reading, UK, 2010, pp. 1-6.

[6] Liu, Shuang, Dahua Li, Qiang GAO, and Yu Song. "Facial Emotion Recognition Based on CNN." In 2020 Chinese Automation Congress (CAC), Shanghai, China, 2020, pp. 398-403.

[7] Lekdioui, Khadija, Yassine Ruichek, Rochdi Messoussi, Youness Chaabi, and Raja Touahni. "Facial expression recognition using face-regions." In 2017 international conference on advanced technologies for signal and image processing (ATSIP), Fez, Morocco2017, pp. 1-6.

[8] Dhavalikar, Anagha S., and R. K. Kulkarni. "Face detection and facial expression recognition system." In 2014 International Conference on Electronics and Communication Systems (ICECS), Coimbatore, India2014 pp. 1-7.

[9] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov,Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich."Going deeper with convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1-9.

[10] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[11] Naveen Kumar H N, Jagadeesha S, Amith K Jain. "Optimization in Feature Extraction schemes on Static Images to improve the performance of Automatic Facial Expression Recognition Systems". International Journal of Computer Sciences and Engineering, Vol.7, Issue.6, pp.1104-1109, 2019.

[12] Kumar, GA Rajesh, Ravi Kant Kumar, and Goutam Sanyal. "Facial emotion analysis using deep convolution neural network." In 2017 International Conference on Signal Processing and Communication (ICSPC), Coimbatore, India, 2017 pp. 369-374.

[13] Kulkarni, Ketki R., and Sahebrao B. Bagal. "Facial expression recognition." In 2015 Annual IEEE India Conference (INDICON), New Delhi, India, 2015, pp.1-5.

[14] Minaee, Shervin, Amirali Abdolrashidi, and Yao Wang. "Face recognition using scattering convolutional network." In 2017 IEEE signal processing in medicine and biology symposium (SPMB), pp. 1-6. IEEE, 2017.

[15] Balasubramanian, Balaji, Pranshu Diwan, Rajeshwar Nadar, and Anuradha Bhatia. "Analysis of Facial Emotion Recognition." In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 945-949.

[16] Albawi, Saad, Tareq Abed Mohammed, and Saad Al- Zawi. "Understanding of a convolutional neural network." In 2017 International Conference on Engineering and Technology, Australia,2018, pp. 115118.
[17] Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient- based localization." In Proceedings of the IEEE international conference on computer vision, Venice Italy, 2017, pp. 618-626.
[18] L. Xu, M. Fei, W. Zhou, and A. Yang, "Face Expression Recognition Based on Convolutional Neural Network". In 2018 Australian & New Zealand Control Conference (ANZCC), Melbourne, Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 2006, pp. 223-230.
[19] Vilone, Giulia, and Luca Longo. "Explainable artificial intelligence: a systematic review."arXiv preprint arXiv:2006.00093 (2020).
[20] Tian, Y-I., Takeo Kanade, and Jeffrey F. Cohn. "Recognizing action units for facial expression analysis." IEEE Transactions on pattern analysis and machine intelligence 23, no. 2 (2001): 97115.
[21] Bartlett, Marian Stewart, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. "Fully automatic facial action recognition in spontaneous behavior." In 7th International (ICET), Antalya, Turkey, 2017, pp. 1-6.
[22] Yu, Zhiding, and Cha Zhang. "Image based static facial expression recognition with multiple deep network learning." In Proceedings of the 2015 ACM on international conference on multimodal interaction, New York, United States, 2015, pp. 435-442.