JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

Speech Emotion Recognition using Convolutional Neural Networks and Mel Frequency Cepstral Coefficients

Medipally Nagasri[1], Kalpana K[2], Shirisha[3]

[1,2,3]Assistant Professor, Department of ECE, Malla Reddy Engineering College and Management Sciences, Hyderabad, Telangana.

Abstract

In recent years, significant advancements have been made in artificial intelligence, machine learning, and human-machine interaction. Voice interaction and command-based control of machines have become increasingly popular, with virtual assistants like SIRI, Alexa, Cortana, and Google Assistant integrated into various consumer electronics. However, one of the limitations of machines is their inability to interact with humans as empathetic conversational partners. They often struggle to recognize and respond to human emotions. Emotion recognition from speech has emerged as a cutting-edge research area in the field of human-machine interaction, aiming to create more robust man-machine communication systems. Researchers are actively working on speech emotion recognition (SER) to enhance the quality of human-machine interaction. To achieve this goal, computers should be capable of recognizing emotional states and responding to them in ways that mirror human understanding. The effectiveness of SER systems relies on the quality of extracted features and the choice of classifiers. This project focuses on identifying four basic emotions—anger, sadness, neutrality, and happiness—from speech. It employs audio files of short Manipuri speech from movies as training and testing datasets. The methodology utilizes Convolutional Neural Networks (CNN) for emotion recognition, employing Mel Frequency Cepstral Coefficients (MFCC) as the feature extraction technique from speech data.

Keywords: Mel Frequency Cepstral Coefficient, Convolutional Neural Network, Speech Emotion Recognition, Deep Learning, Human-Machine Interaction, Emotion Recognition, Virtual Assistants.

1. Introduction

Automatic identification of emotions by facial expressions consists of three steps: face recognition, extraction and classification of features or hand movements, facial features, and voice sound that are used to convey emotions and input [1]. Nonetheless, the latest developments of human user interfaces, which have progressed from traditional mouse and keyboard to automated speech recognition technologies to unique interfaces tailored for individuals with disabilities, do not take full account of these important interactive capabilities, sometimes contributing to less than normal experiences When machines were able to understand such emotional signals, they could provide users precise and effective support in ways that are more in line with the desires and expectations of the individual. From psychological science it is generally agreed that human emotions may be divided into six archetypal feelings: shock, terror, disgust, rage, joy and sadness [2]. Facial expression and voice sound play a critical role in communicating certain emotions. Emotion interpretation has arisen as an essential field of research that can provide some useful insight to a number of ends. People communicate their feelings through their words and facial gestures, consciously or implicitly. To interpret emotions may be used several different types of knowledge, such as voice, writing, and visual. Speech and facial expression have been the valuable tool for identifying feelings since ancient times, and have revealed numerous facets, including mentality. It is an enormous and difficult job to determine the feelings beneath these statements and facial expressions [3]. Scientists from multiple disciplines are seeking to find an effective way to identify human emotions more effectively from different outlets, like voice and facial expressions, to tackle this issue.

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

Computer intelligence, natural language modelling systems, etc., have been used to gain greater precision in this responsiveness towards various speeches and vocal-based strategies. Analysis of the feelings may be effective in several specific contexts. One such area is cooperation with the human computers. Computers can make smarter choices and aid consumers with emotion recognition and can also aid render human-robot experiences more realistic [4]. We would explore current emotion recognition methods, emotion modelling, emotion databases, their features, drawbacks, and some potential future directions in this study. We concentrate on evaluating work activities focused on voice and facial recognition to evaluate emotions. We studied different technical sets that were included in current methodologies and technologies. The essential accomplishments in the sector are completed and potential strategies for improved result are highlighted.

Rest of the paper is organized as follows: Section 2 details about literature survey, section 3 details about the proposed methodology, section 4 details about the results with discussion, and section 5 concludes article with references.

2. Literature survey

Bhaskar, Shabina, et.al. (2023) [5] developed a Malayalam audio-visual speech expression database of unimpaired people. The experiments were conducted on this newly developed Malayalam audio-visual speech database. A combination of Convolutional Neural Network-Long Short-Term Memory deep learning video processing model is applied for this system. Chen, Yuwei, et.al. (2022) [6] experiment optimizes the traditional CNN MobileNet model and finally constructs a new model framework ms_model_M, which has about 5% of the number of parameters of the traditional CNN MobileNet model. Jayanthi, K., and S. Mohan. Et.al. (2022) [7] proposed integrated framework by the virtue of deep classifier fusion demonstrates an exemplary performance with 94.26% accuracy on comparison with 89% and 91.49% respectively for voice signal and facial expression when considered individually. Farkhod, Akhmedov, et.al. (2022) [8] proposed a graph-based emotion recognition method that adopts landmarks on the upper part of the face. Based on the proposed approach, several pre-processing steps were applied. After pre-processing, facial expression features need to be extracted from facial key points. The main steps of emotion recognition on masked faces include face detection by using Haar–Cascade, landmark implementation through a media-pipe face mesh model, and model training on seven emotional classes.

Ullah, Zia, et.al. (2022) [9] proposed a novel technique termed “Improved Deep CNN-based Two Stream Super Resolution and Hybrid Deep Model-based Facial Emotion Detection”, which consists of three working phases: super-resolution, facial emotion recognition, as well as classification. Improved Deep CNN has been used in the super-resolution phase for two streams. Avula, Himaja, et.al. (2022) [10] proposed model tries to provide speech to the mute. Firstly, hand gestures for sign language recognition and facial emotions are trained using CNN and then by training the emotion to speech model. Finally combining hand gestures and facial emotions to realize the emotion and speech. Lee, Young-Shin, et.al. (2022) [11] proposed a model based on artificial intelligence (AI), that can assist in the diagnosis of depressive disorder. Depressive disorder can be diagnosed through a self-report questionnaire, but it is necessary to check the mood and confirm the consistency of subjective and objective descriptions. Hou, Jie. Et.al. (2022) [12] proposed to evaluate the probability of digital representation, identification, and estimation of feelings. The proposed DL-HEDF analyzes the impact of emotional models on multimodal identification. The paper introduces emerging works that use existing methods like CNN for human emotion identification based on language, sound, image, video, and physiological signals.

Fahn, Chin-Shyurng, et.al. (2022) [13] plenty of care robots have been developed, but humanized care robots that can suitably respond to the individual behaviors of elderly people, such as pose, expression, gaze, and speech are generally lacking. Helaly, Rabie, et.al. (2023) [14] proposed, first, a Deep CNN

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

architecture using Transfer Learning (TL) approach for constructing a highly accurate FER system, in which a pre-trained Deep CNN model is adopted by substituting its dense upper layers suitable with FER, and the model is fine-tuned with facial expression data. Khan, Nizamuddin, et.al. (2022) [15] proposed an enhanced hybrid deep learning model based on STN for facial emotion recognition, which gives the best feature extraction and classification in one go and maximizes the accuracy for a large number of samples on FERG, JAFFE, FER-2013, and CK+ datasets. Wu, Yongzhen, et.al. (2022) [16] multi-modal emotion identification. Based on the modal information of facial expression, the method designs a multi-level convolutional neural network (CNN) model for facial expression emotion identification. Based on electroencephalograhpy (EEG) information modes, the method creates a stacked bidirectional LSTM(Bi-LSTM) model for emotion identification.

Teja, Kuppa Sai Sri, et.al. (2022) [17] SAVEE database is used, which comprises audio and amp; visual features of seven unique types of emotions; and these emotions are identified by using CNN-based systems exploiting facial gestures of actors. Important features from the faces of the actors in the database are extracted and trained using existing deep learning methods namely 3D convnets. Gupta, Vanshika, et.al. (2022) [18] proposed available facial expression databases are used to validate the proposed method. Using an ensemble neural network, the experimental findings indicate high accuracy for diverse datasets on the combined features. Bharathi, S., et.al. (2022) [19] two CNN- models are cascaded to produce the facial expression recognition output. YOLO V5 is used for people detection and custom trained CNN model is used for expression recognition. The suggested model provides better accuracy of 95.57% for seven different facial expressions than the existing models.

Singh, Shubham Kumar, et.al. (2022) [20] emotions will be detected and human behavior will be extracted. The various body language approaches like idiomatic expressions, eye stirring and body movement are significant while applying for the association between machines and people. Podder, Tanusree, et.al. (2022) [21] proposed CNN architecture named LiveEmoNet has been jointly trained with wild (FER-2013) and lab-controlled (CK+) datasets for real-time detection, contributing to versatile emotion detection. Azizi, Faza N., et.al. (2022) [22] describes how to use the Convolutional Neural Network (CNN) to identify a person's emotions based on facial expressions observed in facial images. CNN is part of deep learning method that is widely used to analyze information from images and often gives good results. Bhaskar, Shabina, et.al. (2022) [23] introduced a method of visual speech recognition from the face by giving importance to the facial expressions while a person speaking. Facial expression is an important feature for hearing impaired speech recognition because they are more expressive while speaking. Baffour, Prince Awuah, et.al. (2022) [24] revealed the dominance of CNN architectures over other known architectures like RNNs and SVMs, highlighting the contributions, model performance, and limitations of the reviewed state-of-the-art.

3. Proposed Methodology

Emotions are an essential aspect of communication between human beings. There is a very close relationship between emotions, behaviour, and thoughts in such a way that the combination of these aspects governs the way we act and the decisions we make. For this reason, over the past years, there has been a growing interest in this area of scientific research. Automatic recognition of emotions can be applied in several areas to enhance them. For example, human–computer interaction, since detecting the emotional state of a computer system's user will allow generating a more natural, productive, and intelligent interaction. Another area is human–human interaction monitoring, given its allowance to detect conflicts or unwanted situations. This project addresses the automatic emotion recognition from speech, face, and videos as well. The proposed methodology employed deep learning CNN such as the creation of

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

corpora, the feature selection, the design of an appropriate classification scheme, and the fusion with other sources of information, such as text.

Figure 1 shows the proposed block diagram of face and speech-based emotion recognition. RACVDESS dataset is considered to implement this work, which contains both speech and face data files. Then, preprocessing operation is carried out on both datasets performed, which removed the noises from facial images and speech files. Then, MFCC features are extracted only from speech data. Then, CNN model is trained with both speeches based MFCC features and pre-processed facial data. Finally, test face and speech data are applied, and test features are compared with the pre-trained CNN model features. Finally, the predicted emotion is obtained through this AI-CNN model from both face and speech data.

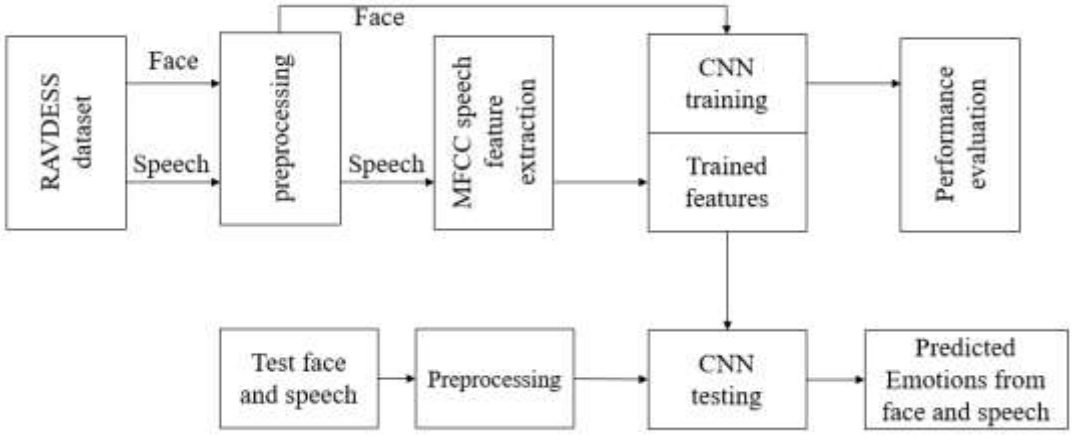

Figure 1. Proposed block diagram.

3.1 Dataset

For facial emotion detection model, we have used 28,709 images with 7 different emotions includes angry, happy, neutral, sad, disgusted, fearful, and surprised. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is used for speech emotion detection model. The data rate, sample frequency, and format of speech audio-only files from the RAVDESS is 16bit, 48kHz, and .wav. This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

3.2 Image and Speech Pre-processing

Digital image processing is the use of computer algorithms to perform image processing on digital images. As a subfield of digital signal processing, digital image processing has many advantages over analogue image processing. It allows a much wider range of algorithms to be applied to the input data — the aim of digital image processing is to improve the image data (features) by suppressing unwanted distortions and/or enhancement of some important image features so that our AI-Computer Vision models can benefit from this improved data to work on. To train a network and make predictions on new data, our images must match the input size of the network. If we need to adjust the size of images to match the network, then we can rescale or emotion data to the required size.

we can effectively increase the amount of training data by applying randomized augmentation to data. Augmentation also enables to train networks to be invariant to distortions in image data. For example, we

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

can add randomized rotations to input images so that a network is invariant to the presence of rotation in input images. An augmented Image Datastore provides a convenient way to apply a limited set of augmentations to 2-D images for classification problems.

3.3 MFCC feature extraction.

Pre-emphasis is the initial stage of extraction. It is the process of boosting energy in high frequency. It is done because the spectrum for voice segments has more energy at lower frequencies than higher frequencies. This is called spectral tilt which is caused by the nature of the glottal pulse. Boosting high-frequency energy gives more info to Acoustic Model which improves phone recognition performance. Figure 2 shows the MFCC feature extraction process block diagram.

Step 1: The given speech signal is divided into frames (~20 ms). The length of time between successive frames is typically 5-10ms.

Step 2: Hamming window is used to multiply the above frames to maintain the continuity of the signal. Application of hamming window avoids Gibbs phenomenon. The Hamming window is multiplied to every frame of the signal to maintain the continuity in the start and stop point of frame and to avoid hasty changes at end point. Further, a hamming window is applied to each frame to collect the closest frequency component together.

Step 3: Mel spectrum is obtained by applying Mel-scale filter bank on DFT power spectrum. Mel-filter concentrates more on the significant part of the spectrum to get data values. Mel-filter bank is a series of triangular band pass filters similar to the human auditory system. The filter bank consists of overlapping filters. Each filter output is the sum of the energy of certain frequency bands. Higher sensitivity of the human ear to lower frequencies is modeled with this procedure. The energy within the frame is also an important feature to be obtained. Compute the logarithm of the square magnitude of the output of Mel-filter bank. Human response to signal level is logarithm. Humans are less sensitive to small changes in energy at high energy than small changes at low energy. Logarithms compresses dynamic range of values.

Step 4: Mel-scaling and smoothing (pull to right). Mel scale is approximately linear below 1 kHz and logarithmic above 1 kHz.

Step 5: Compute the logarithm of the square magnitude of the output of Mel filter bank.

Step 6: DCT is a further stage in MFCC which converts the frequency domain signal into time domain and minimizes the redundancy in data which may neglect the smaller temporal variations in the signal. Mel-cepstrum is obtained by applying DCT on the logarithm of the mel-spectrum. DCT is used to reduce the number of feature dimensions. It reduces spectral correlation between filter bank coefficients. Low dimensionality and 17 uncorrelated features are desirable for any statistical classifier. The cepstral coefficients do not capture the energy. So, it is necessary to add energy feature. Thus twelve (12) Mel Frequency Cepstral Coefficients plus one (1) energy coefficient are extracted. These thirteen (13) features are generally known as base features.

Step 7: Obtain MFCC features. The MFCC i.e., frequency transformed to the cepstral coefficients and the cepstral coefficients transformed to the MFCC by using the equation (1).

$$mel(f) = 2595 \times \log 10 \left(1 + \frac{f}{700}\right) \quad (1)$$

Where f denotes the frequency in Hz. The Step followed to compute MFCC. The MFCC features are estimated by using the following equation (2).

$$C_n = \sum_{n=1}^{K} (logS_k) \left[n \left(K - \frac{1}{2}\right) \frac{\pi}{K}\right] wheren = 1,2, \ldots . K \quad (2)$$

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

Here, K represents the number of Mel cepstral coefficient, C0 is left out of the DCT because it represents the mean value of the input speech signal which contains no significant speech related information. For each of the frames (approx. 20 ms) of speech that has overlapped, an acoustic vector consisting of MFCC is computed. This set of coefficients represents as well as recognize the characteristics of the speech.

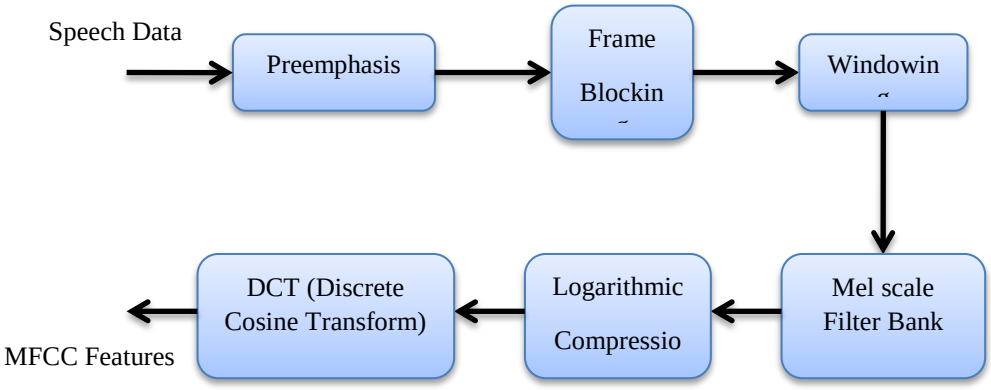

Figure 2. MFCC operation diagram

3.4 CNN model

Figure 3 shows the deep CNN model for emotion detection using speech recognition and Figure 4 shows the proposed deep CNN model for emotion detection from facial expressions. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system. It is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system. In addition, it shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output. Moreover, it may be used to represent a system at any level of abstraction, and it may be partitioned into levels that represent increasing information flow and functional detail.

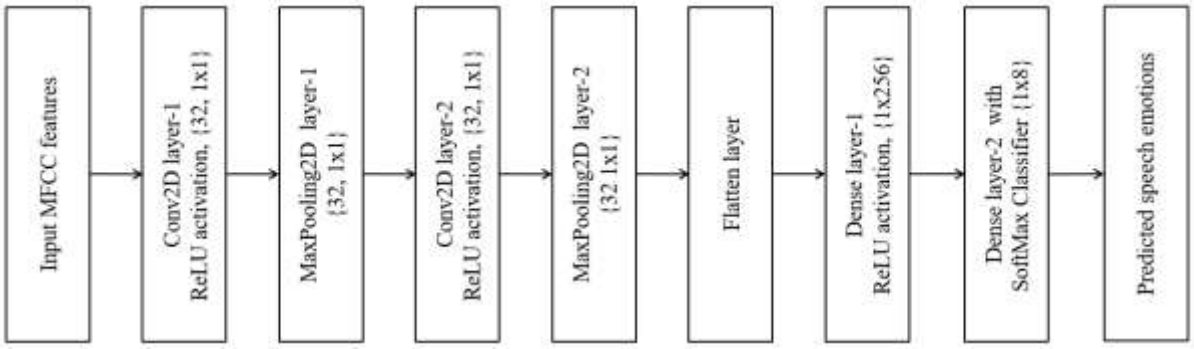

Figure 3: Proposed deep CNN model for emotion detection using speech recognition.

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

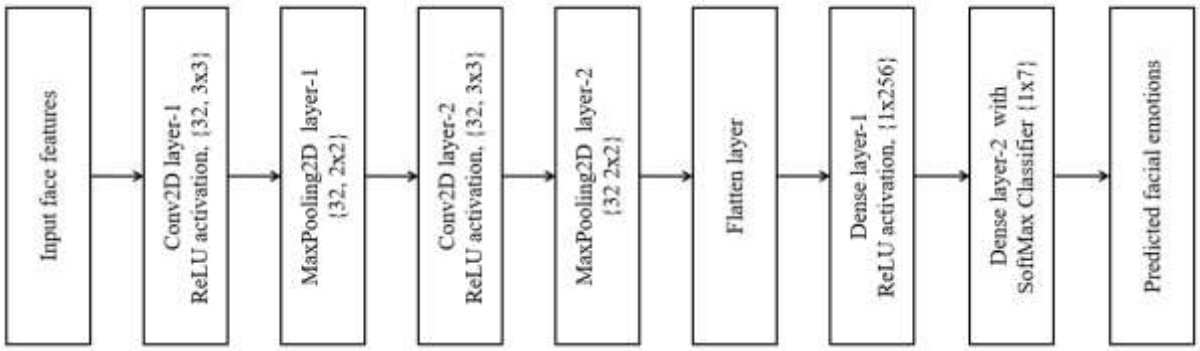

Figure 4: Proposed deep CNN model for emotion detection from facial expressions.

4. Results and Discussion

Figure 5 shows both datasets are processed and we can see total number of images and audio files available in both datasets with various emotion classes. The training of CNN with Facial images got 96.52% accuracy and CNN speech Emotion we got 96.72% accuracy as shown in Figure 6.

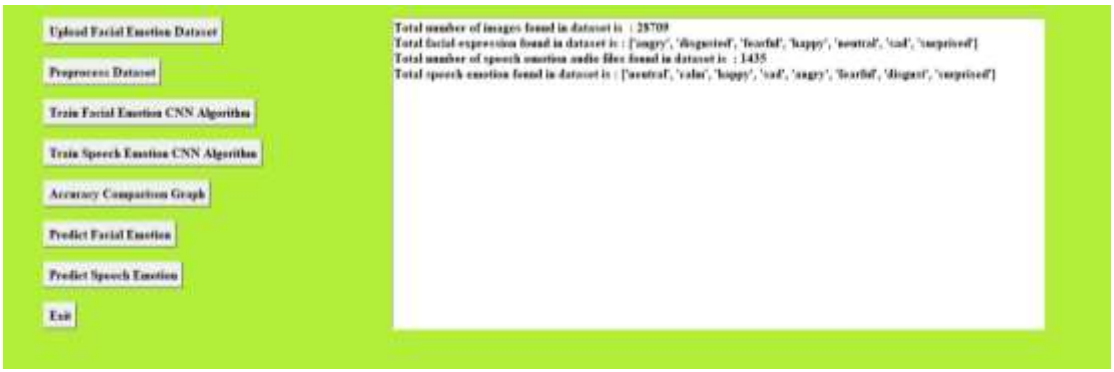

Figure 5. Records found in dataset.

Figure 7 illustrate the sample test images of emotion prediction from given facial expressions, where it includes all the emotions such as sad, angry, neutral, disgusted, surprised, and fearful. Figure 8 discloses the obtained prediction accuracy and loss performance using proposed deep CNN from facial expression, speech, and videos. From both the figures, it is observed that proposed deep CNN obtained superior performance for emotion prediction from videos as compared to both facial expression and speech inputs.

Figure 6. Accuracy estimation.

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

5. Conclusion

Emotion interpretation has arisen as an essential field of research that can provide some useful insight to several ends. People communicate their feelings through their words and facial gestures, consciously or implicitly. To interpret emotions may be used several different types of knowledge, such as voice, writing, and visual. Therefore, this work proposed a deep CNN model for emotion prediction from speech, and facial expression with enhanced prediction accuracy and reduced loss. In addition, the speech CNN model utilized MFCC as feature extraction from given speech samples.

References

[1]. Jaiswal, Akriti, A. Krishnama Raju, and Suman Deb. "Facial emotion detection using deep learning." 2020 international conference for emerging technology (INCET). IEEE, 2020.

[2]. Ozdemir, Mehmet Akif, et al. "Real time emotion recognition from facial expressions using CNN architecture." 2019 medical technologies congress (tiptekno). IEEE, 2019.

[3]. Agrawal, Ishika, et al. "Emotion Recognition from Facial Expression using CNN." 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC). IEEE, 2021.

[4]. Abdullah, Sharmeen M. Saleem Abdullah, et al. "Multimodal emotion recognition using deep learning." Journal of Applied Science and Technology Trends 2.02 (2021): 52-58.

[5]. Bhaskar, Shabina, and T. M. Thasleema. "LSTM model for visual speech recognition through facial expressions." Multimedia Tools and Applications 82.4 (2023): 5455-5472.

[6]. Chen, Yuwei, and Jianyu He. "Deep learning-based emotion detection." Journal of Computer and Communications 10.2 (2022): 57-71.

[7]. Jayanthi, K., and S. Mohan. "An integrated framework for emotion recognition using speech and static images with deep classifier fusion approach." *International Journal of Information Technology* 14.7 (2022): 3401-3411.

[8]. Farkhod, Akhmedov, et al. "Development of Real-Time Landmark-Based Emotion Recognition CNN for Masked Faces." Sensors 22.22 (2022): 8704.

[9]. Ullah, Zia, et al. "Improved Deep CNN-based Two Stream Super Resolution and Hybrid Deep Model-based Facial Emotion Recognition." Engineering Applications of Artificial Intelligence 116 (2022): 105486.

[10]. Avula, Himaja, R. Ranjith, and Anju S. Pillai. "CNN based Recognition of Emotion and Speech from Gestures and Facial Expressions." 2022 6th International Conference on Electronics, Communication and Aerospace Technology. IEEE, 2022.

[11]. Lee, Young-Shin, and Won-Hyung Park. "Diagnosis of depressive disorder model on facial expression based on fast R-CNN." Diagnostics 12.2 (2022): 317.

[12]. Hou, Jie. "Deep Learning-Based Human Emotion Detection Framework Using Facial Expressions." Journal of Interconnection Networks 22.Supp01 (2022): 2141018.

[13]. Fahn, Chin-Shyurng, et al. "Image and Speech Recognition Technology in the Development of an Elderly Care Robot: Practical Issues Review and Improvement Strategies." Healthcare. Vol. 10. No. 11. MDPI, 2022.

[14]. Helaly, Rabie, et al. "DTL-I-ResNet18: facial emotion recognition based on deep transfer learning and improved ResNet18." Signal, Image and Video Processing (2023): 1-14.

[15]. Khan, Nizamuddin, Ajay Vikram Singh, and Rajeev Agrawal. "Enhanced Deep Learning Hybrid Model of CNN Based on Spatial Transformer Network for Facial Expression Recognition." International Journal of Pattern Recognition and Artificial Intelligence 36.14 (2022): 2252028.

[16]. Wu, Yongzhen, and Jinhua Li. "Multi-modal emotion identification fusing facial expression and EEG." Multimedia Tools and Applications (2022): 1-19.

[17]. Teja, Kuppa Sai Sri, et al. "3D CNN Based Emotion Recognition Using Facial Gestures." Evolution in Computational Intelligence: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021). Singapore: Springer Nature Singapore, 2022.

[18]. Gupta, Vanshika, and Vikas Sejwar. "Facial Expression Recognition with Combination of Geometric and Textural Domain Features Extractor using CNN and Machine Learning." 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS). IEEE, 2022.

[19]. Bharathi, S., K. Hari, and M. Senthilarasi. "Expression Recognition using YOLO and Shallow CNN Model." 2022 Smart Technologies, Communication and Robotics (STCR). IEEE, 2022.

[20]. Singh, Shubham Kumar, et al. "Deep Learning and Machine Learning based Facial Emotion Detection using CNN." 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2022.

[21]. Podder, Tanusree, Diptendu Bhattacharya, and Abhishek Majumdar. "Time efficient real time facial expression recognition with CNN and transfer learning." Sādhanā 47.3 (2022): 177.

[22]. Azizi, Faza N., Arrie Kurniawardhani, and Irving V. Paputungan. "Facial Expression Image based Emotion Detection using Convolutional Neural Network." 2022 IEEE 20th Student Conference on Research and Development (SCOReD). IEEE, 2022.

[23]. Bhaskar, Shabina, and T. M. Thasleema. "CNN Based Feature Extraction for Visual Speech Recognition in Malayalam." Proceedings of Data Analytics and Management: ICDAM 2021, Volume 2. Springer Singapore, 2022.

[24]. Baffour, Prince Awuah, et al. "A Survey on Deep Learning Algorithms in Facial Emotion Detection and Recognition." Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi 7.1 (2022): 24-32.