

Deep-SFER: deep convolutional neural network and MFCC an effective speech and face emotion recognition

Ravi Gummula, Vinothkumar Arumugam, Abilasha Aranganathan

Department of Electronics and Communication Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, India

Article Info

Article history:

Received Mar 13, 2024

Revised Aug 22, 2024

Accepted Aug 31, 2024

Keywords:

Convolutional neural network

Facial emotion recognition

Facial expression recognition

Image recognition

MFCC

Speech emotion recognition

ABSTRACT

There has been a lot of progress in recent years in the fields of expert systems, artificial intelligence (AI) and human machine interface (HMI). The use of voice commands to engage with machinery or instruct it to do a certain task is becoming more common. Numerous consumer electronics have SIRI, Alexa, Cortana, and Google Assistant built in. In the field of human-device interaction, emotion recognition from speech is a complex research subject. We can't imagine modern life without machines, so naturally there's a need to create a more robust framework for human-machine communication. A number of academics are now working on speech emotion recognition (SER) in an effort to improve the interaction between humans and machines. We aimed to identify four fundamental emotions: angry, unhappy, neutral and joyful from speech in our experiment. As you can hear below, we trained and tested our model using audio data of brief Manipuri speeches taken from films. This task makes use of convolutional neural networks (CNNs) to extract functions from speech in order to recognize different moods using the Mel-frequency cepstral coefficient (MFCC).

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Ravi Gummula

Department of Electronics and Communication Engineering

Dr. M.G.R. Educational and Research Institute

Chennai, India

Email: ravi.gummula@gmail.com

1. INTRODUCTION

Hand gestures, facial features and vocal inflections are all potential inputs and emotions that may be automatically detected via the three steps of feature extraction and classification [1]. Despite significant advancements since the dawn of the mouse and keyboard, automated speech recognition and accessible user interfaces for individuals with disabilities, state-of-the-art human user interfaces often fail to fully leverage these vital interactive capabilities, resulting in less than satisfactory experiences. Machines that could interpret these emotional cues may tailor their assistance to each individual's unique tastes and requirements. According to studies in psychology, there are six basic human emotions: happiness, sadness, fear, wrath, disgust, and shock. There are certain emotions that can only be expressed via specific facial gestures and voice intonation [2]-[4].

Emotion interpretation is a promising new field that could provide answers to many mysteries [5]. A person's facial expressions convey both spoken and nonverbal messages about their emotional state [6], [7]. It is possible to interpret emotional states from a variety of sources, such as verbal, textual and visual cues. A long-standing talent has allowed us to comprehend many things, including people's minds: the capacity to interpret their feelings via their words and expressions. Reading between the lines of these sentiments and

words is an incredibly difficult and time-consuming endeavour. A number of scientific disciplines are now collaborating on a solution to this problem: an improved method of emotion detection from a variety of sources, including facial expressions and voice [8].

The introduction of artificial intelligent (AI) and natural language modelling systems have improved the accuracy of this reaction to various vocal-based techniques and utterances [9]. When in doubt, it could be wise to examine feelings. One such issue is the use of human-computer cooperation. Computers can make better decisions, help consumers identify emotions, and facilitate the development of more realistic interactions between humans and robots. We would examine the current state of emotion recognition methods, emotion models and emotion databases, going over their pros and cons and what the future holds in terms of these areas [10]. One of our main areas of research is emotion assessment via the use of voice and facial recognition. We examined the many technical components that comprise contemporary methods and instruments. The field's most significant accomplishments have been completed and potential methods for improved outcomes have been highlighted.

The purpose of using convolutional neural networks (CNNs) in deep learning [11] is to construct a trustworthy system capable of automatically identifying and categorising human emotions in voice and facial expressions [12]. This interdisciplinary approach incorporates auditory and visual cues to provide a more complete picture of people's emotions. CNNs are used by the model to effectively evaluate facial expressions [13]. CNNs are very effective at processing spatial data. Emotional components in the auditory domain may be extracted simultaneously using speech data. The primary goals are to make the model more generalizable across different datasets to make emotion identification more accurate and to lay the stage for future advancements in areas such as sentiment analysis, mental health monitoring and smart machine-human interactions [14]. The literature survey is shown in Table 1.

Table 1. Literature survey

Year	Author	Title	Method	Abstract
2014	Kudiri <i>et al.</i> [15]	Emotion detection machine learning technique.	Human computer interaction, relative bin frequency coefficients (RBFC), relative sub-1 image features (RSF), support vector machine (SVM).	Estimation of human emotions using ECG images and EEG Signals. Identifying the positive and negative emotions.
2017	Ruiz <i>et al.</i> [16]	Human emotion detection through facial expressions.	SVM, RBFC.	Facial expressions are two types namely, Deliberate and non-deliberate facial expression.
2019	Khalil <i>et al.</i> [17]	Emotion detection through speech and facial expressions.	SER, deep learning, deep neural network, deep Boltzmann machine, recurrent neural network, CNN.	The review covers databases used, emotions extracted and contributions mode.
2020	Begaj <i>et al.</i> [18]	Emotion recognition based on facial expressions using CNN.	CNN, facial expression recognition, facial emotion recognition, image recognition.	Over the last few years, there have been an increasing number of studies in emotion recognition because of the importance and the impact that it has in the interaction of humans with computers.

2. LITERATURE SURVEY

2.1. The 2014 emotion detection through speech and facial expressions

There have been numerous studies done regarding predicting emotions through emotive speech. Emotion detection via speech is growing in popularity since speech-based input sensors are much cheaper than other technologies. Speech may be recognized in the loudest situations using the human auditory frequency range of 20 Hz to 20,000 Hz. When managing with emotion in noisy situations through speech, pitch variations and tempo of speaking in regard to time become significant. Nevertheless, the pitch of the speech signal can shift due to a number of environments. Additionally, the increasing dimensionality of the voice signals increases the time complexity.

2.2. The 2017 state of human emotion recognition via verbal and facial expressions

It has proven difficult to use computers to estimate human emotions ever since people began taking part in verbal secessions. Using a proposed hybrid method that incorporates both spoken and facial emotions, this study estimates the typical emotional states of a subject during a conversational separation, including anger, sadness, satisfaction, boredom, disgust, and astonishment. For audio and video data in particular, we use relative bin frequency coefficients and family member sub-image-based features. When it comes to classification, SVM with radial basis bit is the way to go. According to this study's findings, the

recommended blend approach and feature removal with voice and face are two of the most well-known factors impacting the emotion detection system. There are few factors that might influence the emotion detecting system, although their impact is minimal. Through the use of purposeful facial expressions, it was shown that the bimodal emotion detection system outperformed the unimodal method. An appropriate database is used to address the problem. The results indicated that compared to the other systems, the recommended emotion recognition method performed much better when it came to fundamental emotional courses.

2.3. Emotion recognition in speech with the use of deep learning techniques in 2019

A vital yet challenging aspect of human-computer interaction (HCI) is the ability to perceive recognition from voice signals. Several methods, including several well-established speech assessment and classification algorithms, have been used in the speech emotion recognition (SER) literature to de-emotionalize signals. A new alternative to traditional procedures in SER based on deep learning techniques has just been proposed. This article provides an overview of deep discovering techniques and discusses recent research that has used these methods to recognize emotions in spoken language. The assessment delves into the data sources used, emotions eliminated, funds allocated to SER and any associated restrictions.

2.4. In 2020 emotion recognition using convolutional neural networks

Facial emotion recognition has been the subject of an increasing sample of studies in recent years due to its significance and influence on HCI. The use of deep learning methods is becoming essential due to the rise in the number of challenging datasets. This study investigates the challenges of emotion acknowledgment datasets and experiments with various criteria and designs of CNNs to identify seven human emotions: angst, fear, contempt, mockery, joy, despair, and shock in facial expressions. The key dataset for our work is multi-emotion facial expression dataset (iCV MEFED) which is pretty new, intriguing and very difficult.

2.5. The 2014 emotion detection through speech and facial expressions

There have been numerous studies done regarding predicting emotions through emotive speech [15]. Emotion detection via speech is growing in popularity since speech-based input sensors are much cheaper than other technologies. Speech may be recognized in the loudest situations using the human auditory frequency range of 20 Hz to 20,000 Hz. When managing with emotion in noisy situations through speech, pitch variations and tempo of speaking in regard to time become significant. Nevertheless, the pitch of the speech signal can shift due to a number of environments. Additionally, the increasing dimensionality of the voice signals increases the time complexity.

2.6. The 2017 state of human emotion recognition via verbal and facial expressions

It has proven difficult to use computers to estimate human emotions ever since people began taking part in verbal secessions. Using a proposed hybrid method that incorporates both spoken and facial emotions, this study estimates the typical emotional states of a subject during a conversational separation, including anger, sadness, satisfaction, boredom, disgust and astonishment. For audio and video data in particular, we use relative bin frequency coefficients and family member sub-image-based features. When it comes to classification, SVM with radial basis bit is the way to go. According to this study's findings [16], the recommended blend approach and feature removal with voice and face are two of the most well-known factors impacting the emotion detection system. There are few factors that might influence the emotion detecting system, although their impact is minimal. Through the use of purposeful facial expressions, it was shown that the bimodal emotion detection system outperformed the unimodal method. An appropriate database is used to address the problem. The results indicated that compared to the other systems, the recommended emotion recognition method performed much better when it came to fundamental emotional courses.

2.7. Emotion recognition in speech with the use of deep learning techniques in 2019

A vital yet challenging aspect of HCI is the ability to perceive recognition from voice signals. Several methods, including several well-established speech assessment and classification algorithms, have been used in the SER literature to de-emotionalize signals. A new alternative to traditional procedures in SER based on deep learning techniques has just been proposed. This article provides an overview of deep discovering techniques and discusses recent research that has used these methods to recognise emotions in spoken language [17]. The assessment delves into the data sources used, emotions eliminated, funds allocated to SER and any associated restrictions.

2.8. In 2020 emotion recognition using convolutional neural networks

Facial emotion recognition has been the subject of an increasing sample of studies in recent years due to its significance and influence on HCI [18]. The use of deep learning methods is becoming essential due to the rise in the number of challenging datasets [19]. This study investigates the challenges of emotion acknowledgment datasets and experiments with various criteria and designs of CNNs to identify seven human emotions: angst, fear, contempt, mockery, joy, despair and shock in facial expressions [20]. The key dataset for our work is multi-emotion facial expression dataset (iCV MEFED) which is pretty new, intriguing and very difficult.

3. METHOD

3.1. Problem statement

Given that humans engage in conversational sessions, it has been difficult to evaluate human emotions using a computer [21]. This research project proposes a hybrid system that uses harsh faces and voice to mimic the basic concepts (angry, sad, joyful, bored, disgusted, and shocked) of an arsonist during a conversational secession. Family member in relative sub-image and frequency coefficients tyres that are based on specialised materials are used for visual and auditory data [22]. Support classification is accomplished using vector equipment with a radial basis kernel. Based on the results of this research, the recommended function extraction using voice and face is a major factor influencing the emotion discovery system, especially when combined with the proposed approach. There are a few factors that might influence the emotion detecting system, but they won't have much of an effect. Compared to a unimodal system that uses intentional faces to determine emotions, the bimodal approach is less effective [23]. A relevant data source is used to tackle the problem. According to the results, the suggested emotion discovery method outperformed the competition when it came to basic psychological categories.

3.2. Proposed system

The importance and impact of facial emotion recognition in HCI has led to an upsurge in research into the topic in recent years. It is now necessary to use deep learning algorithms due to the increasing diversity of demanding datasets [24]. In this paper, we take a look at the problems with feeling acknowledgment datasets and try various criteria and CNN architectures to detect seven human emotions: angst, concern, contempt, joy, sadness, and surprise in facial expressions [25]. Because it is novel, interesting and challenging, we have chosen iCV MEFED as the primary dataset for our research.

3.3. Convolutional neural network

The CNN is commonly used in deep learning for image recognition and computer vision tasks. It's great at capturing spatial relationships in data. Figure 1 explains the CNN process.

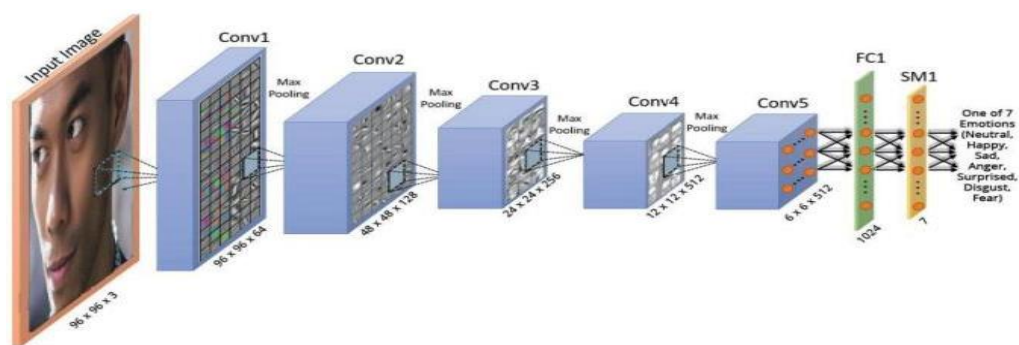


Figure 1. CNN process

3.3.1. Here are the key aspects of CNNs

CNNs are designed with a number of layers such as fully connected, pooling and convolutional ones. Convolutional layers capture feature hierarchies by applying filters (kernels) to incoming data. CNNs are commonly used for image and video recognition, medical image analysis, object detection, facial recognition, and other purposes. They are also used in non-visual activities which includes natural language processing and time-series analysis. These attributes contribute to CNNs exceptional feature extraction and representation skills, making them a cornerstone of modern computer vision and related fields.

Backpropagation is a mechanism for updating the weights in a network based on an error. It entails computing the gradient of the loss function in relation to each weight and changing the weights accordingly. optimization algorithms: stochastic gradient descent (SGD), Adam, and RMSprop are three popular optimizers, each with its own strategy for modifying learning rate and other parameters.

3.3.2. Layers that use convolution

In convolution, features are extracted by sliding a filter across the input data, while spatial correlations are preserved. By learning hierarchical representations, many filters can identify a wide range of characteristics. By lowering the number of spatial dimensions and the computational effort, pooling layers lower sample feature maps. Max pooling (choosing the maximum value) and average pooling are two common ways to pool data.

3.3.3. Layers that are fully connected

In order to aggregate learnt features for use in regression or classification tasks, fully connected layers are used. Each neuron in one layer is connected to every neuron in the layer below and above it via these layers. Fully connected layers, also known as dense layers, perform a significant function in neural networks, in particular for regression and classification. These layers are referred to be "fully connected" because each neuron in one layer communicates with every neuron in the layers exclusively below and above it. This design enables the network to capture and aggregate complicated interactions between learned features. They identify patterns or sequences in the data, such as edges in an image or temporal interactions in time series. Following the feature extraction procedure, fully connected layers are used to aggregate all of the knowledge gained and generate the final forecast. This information aggregation makes fully connected layers necessary for programs such as regression and classification. In each fully connected layer, each neuron computes a weighted sum of the previous layer's outputs, adds a bias term, and sends the result employing an activation function. Fully connected layers are additionally helpful in dimensionality reduction. As data passes through these layers, the high-dimensional feature space usually becomes flattened, making it easier to identify or determine outcomes. For classification tasks, the final fully connected layer is frequently constructed up of as many neurons as classes, with a softmax activation function producing probabilities for each.

3.4. Mel frequency cepstral coefficient

Mel-frequency cepstral coefficient (MFCC) is a widely used technique for extracting audio features and the flow chart is shown in the Figure 2. The primary goals are:

- Eliminate pitch information which is vocal fold excitation.
- Separate the retrieved characteristics from each other.
- Modify to accommodate the way people hear different volumes and frequencies of sound.
- Record how context is changing.

Feature is the acoustic property of the voice signal. The acoustic features of the speech signal are preserved throughout analysis by extracting just a limited quantity of data from the signal [26].

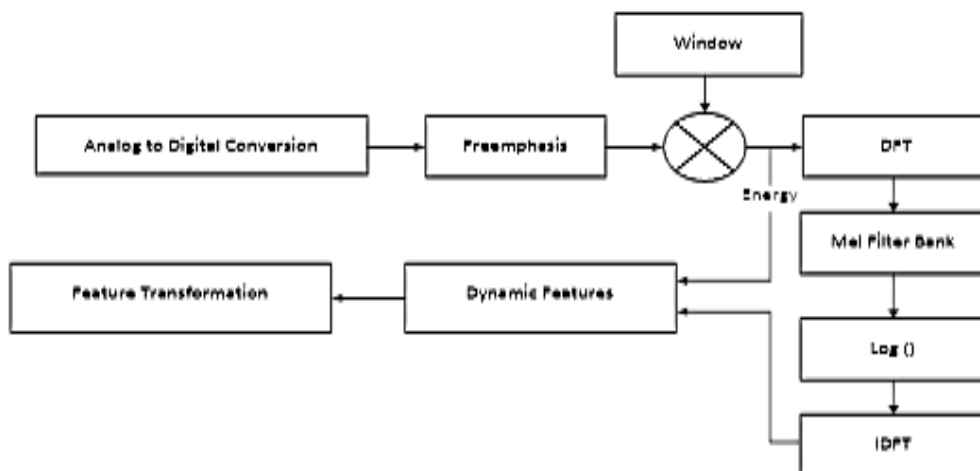


Figure 2. MFCC flow chart

3.5. Face emotion using CNN

The face emotion using CNN process steps are explained in Figure 3.

- (a) First thing to do (Step 1) gather information: obtain a collection of images of faces labelled with matching emotions (such as happiness, sorrow and anger). Get the images ready for processing by doing things like resizing, normalising pixel values and adding metadata if necessary.
- (b) Style of the model (Step 2): CNN style is explained here.
 - Layer for input: takes pictures with a certain dimensionality. A layer of convolutions: deactivate the picture functionality. After every convolutional layer, there should be activation functions (like rectified linear unit (ReLU)). Layer merging (e.g., max pooling) to reduce spatial dimensions and important leading functions.
 - Flattening layer: before feeding the output of the convolutional layers into fully connected layers, flatten them.
 - Interconnected layers: use characteristics extracted by convolutional layers to do categorization. In the result layer, there are as many nerve cells as there are classes (e.g., emotional states).
- (c) Forward propagation (Step 3): apply the CNN's predefined layers to the input pictures. Determine the network's output by using the learned biases and weights.
- (d) Determine loss (Step 4): to find the difference between the actual and projected labels, use a loss feature that is excellent for category tasks, such as categorical cross-entropy. Fifthly, when working with factors like weights and predispositions, use back propagation to determine the loss function's slopes. Revise the requirements with the help of an optimization programme like RMSprop, Adam or SGD.
- (e) Data training (Step 5): put the CNN through its paces with the categorised dataset. Repeat iteratively across picture batches, doing forward propagation, loss computation and back propagation. Continue until the merging requirements are met or until a specified number of dates have passed.
- (f) Evaluation (Step 6): try out the seasoned model on a fresh batch of validation data to see how well it performs. Evaluate parameters like F1-score, recall, accuracy and precision.
- (g) Verifying (Step 7): take a look at the experienced model on its generalisation efficiency on a different test. To verify the design's efficacy on hidden data, compute comparable evaluation measures.
- (h) Optimization and fine-tuning (Step 8): adjust design style and hyper parameters (such as learning rate and batch dimension) in light of test results. To avoid overfitting, use approaches such as failure or L2 regularisation.
- (i) Execution (Step 9): use fresh face images to train the skilled version to make inferences. Incorporate the model into systems or apps for practical use, such as discovering emotions in videos or using it in real-time.

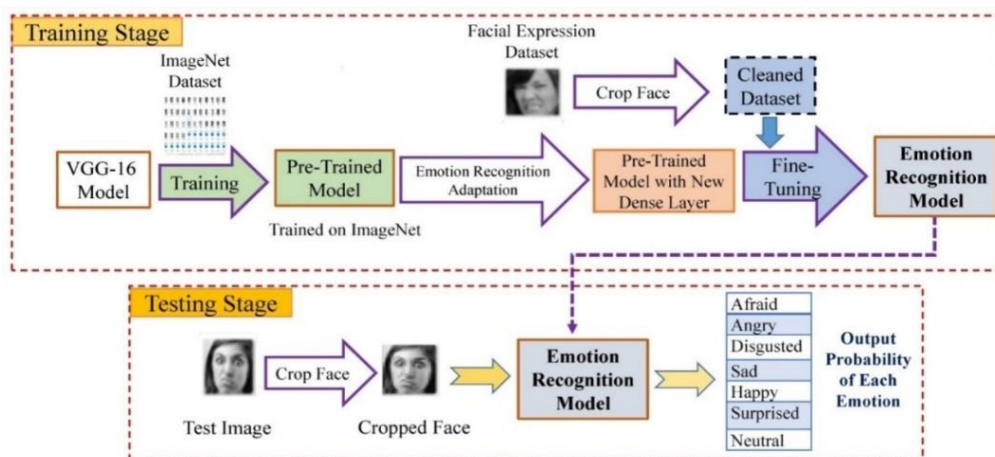


Figure 3. Face emotion using CNN

3.6. Speech emotion using CNN and MFCC

All the stages of this process are explained in Figure 4.

- (a) Stage 1 data preparation: North American accent phrases may be found in the ryerson audio-visual database of emotional speech and song (RAVDESS). Expressions of pleasure, sadness, anger, surprise and disgust are all part of the language we use. There are two tiers of emotional intensity (regular and

solid) and a third, more subtle, level of emotion. Collect a collection of audio clips labelled with various emotional states.

- Get the audio files ready: the audio files must be prepared before anything else. An individual's distinct character in the file's third digit, which represents the emotion key, allows the sound framework to recognise different emotions. Five distinct emotions: calm, angry, worried, disgusted, and surprise which are part of the data set. Transform the unprocessed audio samples into a suitable format such as waveform audio file (WAV) files. Edit the audio samples such that MFCCs are not present.
- (b) Stage 2 creating the model: when deciding on speech tags, classification according to the number of classes is required. The following categories are: classification: good and poor excellent: unwell, content. Poor: anxiety, depression, and irritability. Anger, sadness, contentment, fear, and tranquilly are the classes. Emotional state: fury, depression, joy, worry, irritability, temper, and dread.
- Define the CNN algorithm here: the input layer is capable of receiving MFCCs.
 - Layers of convolution: use one-dimensional convolutional filters to pick up on patterns in the surrounding area.
 - Layers for pooling: reduce spatial dimensions by down sampling attribute maps. Divide the input result into completely connected layers by flattening the layer. Execute category based on drawn out attributes: fully connected layers. The output layer consists of nerve cells that are activated by softmax and correspond to different emotion types.
- (c) Stage 3 data training: here, you'll build the CNN model's layers, choose the activation function, make the appropriate decision and tell Softmax to split the discussion into many teams. Make use of training data to hone the model and then test it with new data. Evaluate the estimated cost in relation to the actual costs. The accuracy of the setting is shown by this comparison. Divide the dataset into three parts: training, recognition and testing.
- (d) Stage 4 first, we need to train the CNN model: perform iterations using sets of MFCCs. Initiate back propagation, forward breeding and loss computation. Use an optimisation formula (such as Adam or SGD) to update the parameters. Fourth piece of advice: test the veteran model on the validation data set: determine measures like F1-score, recall, accuracy and precision.
- (e) Stage 5 assessing: take a look at the tested model using the training set: find analytical metrics to see how well generalisation works.
- (f) Stage 6 testing: make available the trained model for use in inferring fresh audio samples in real time: take audio samples as input. Get MFCCs out of the data. Input the MFCCs into the expert design to foretell the emotion.

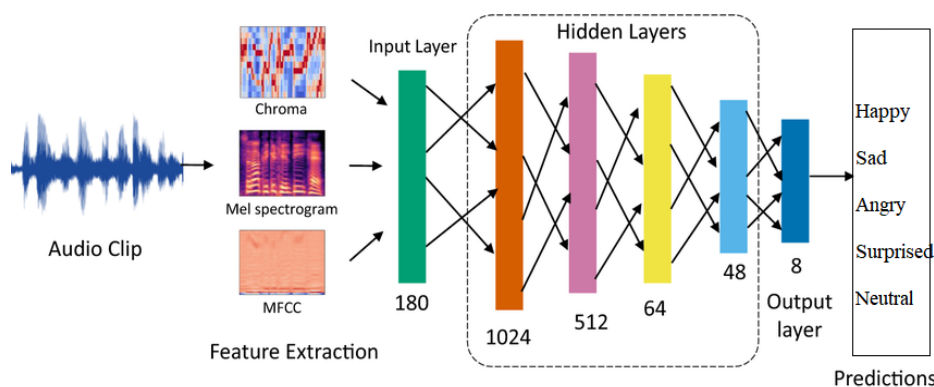


Figure 4. Speech emotion using CNN

4. RESULTS

In order to accomplish this task, we have trained a CNN algorithm using the RAVDESS Sound Dataset to identify emotions in spoken language and we have utilized the emotion facial expression pictures dataset to identify facial expressions. Using the same algorithm, we have trained both the speech dataset and the face photo dataset using CNN versions. You can find the face dataset which contains the photographs of emotion faces shown in Figure 5 in the 'Dataset' folder. We are showing dataset of audio files in Figure 6 and this dataset saved inside speech emotion dataset folder.

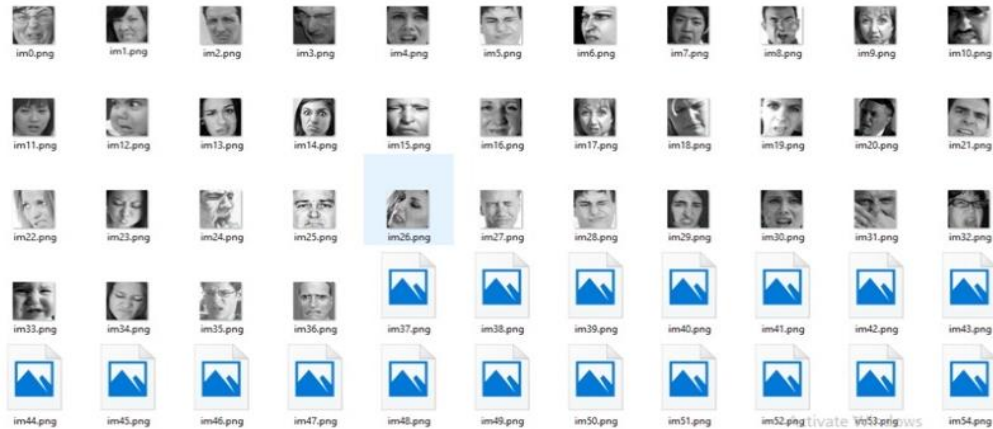


Figure 5. Emotion faces

Actor_01	12-09-2021 11:43	File folder
Actor_02	12-09-2021 10:33	File folder
Actor_03	12-09-2021 10:33	File folder
Actor_04	12-09-2021 10:33	File folder
Actor_05	12-09-2021 10:33	File folder
Actor_06	12-09-2021 10:33	File folder
Actor_07	12-09-2021 10:33	File folder
Actor_08	12-09-2021 10:33	File folder
Actor_09	12-09-2021 10:33	File folder
Actor_10	12-09-2021 10:33	File folder
Actor_11	12-09-2021 10:33	File folder
Actor_12	12-09-2021 10:33	File folder
Actor_13	12-09-2021 10:33	File folder
Actor_14	12-09-2021 10:33	File folder
Actor_15	12-09-2021 10:33	File folder
Actor_16	12-09-2021 10:33	File folder
Actor_17	12-09-2021 10:33	File folder
Actor_18	12-09-2021 10:33	File folder
Actor_19	12-09-2021 10:33	File folder
Actor_20	12-09-2021 10:33	File folder
Actor_21	12-09-2021 10:33	File folder
Actor_22	12-09-2021 10:33	File folder
Actor_23	12-09-2021 10:33	File folder
Actor_24	12-09-2021 10:33	File folder

Figure 6. Emotion audio datasets

Observed here are 24 separate star recordings of human speech, spanning 8 distinct emotional states: neutral, tranquil, delighted, depressing, upset, afraid, disgusted, and astonished. To see audio information, just go to any folder in the above display. Every wave file is associated with a number that is split by the '-' symbol; three digits represent the id, one represents the gender, and the third places a value between one and eight that represents the emotion. The data set loadig and pre-processing data shown in Figures 7 and 8.

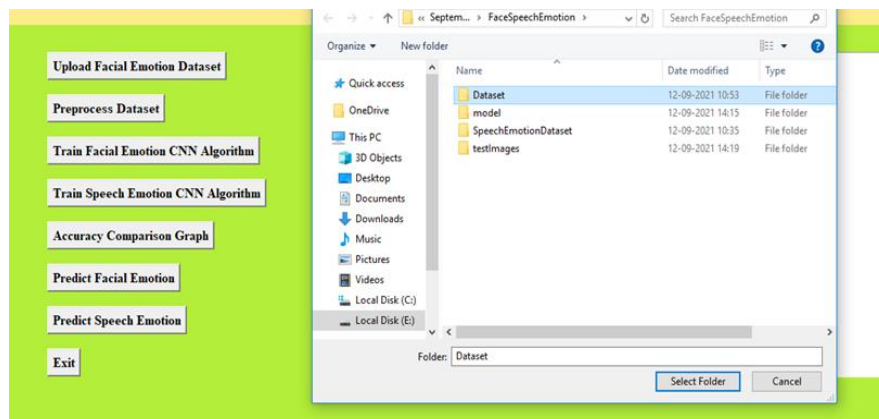


Figure 7. Data set loading

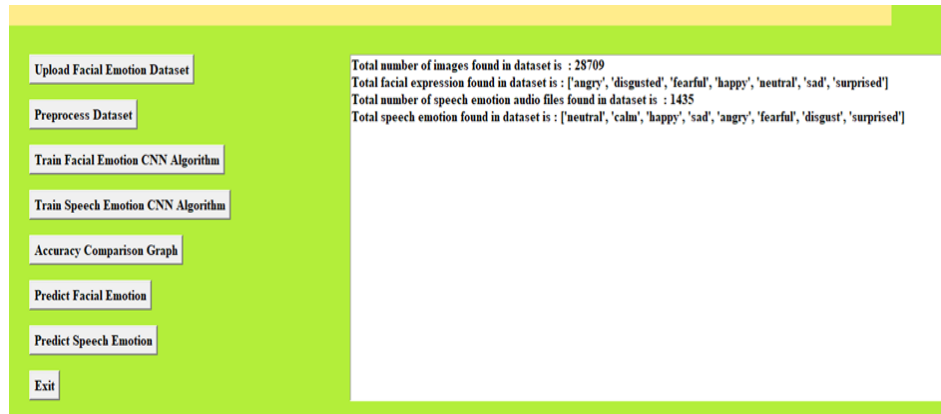


Figure 8. Pre-processing data

Figure 9 explains about the algorithm results of CNN process. The graph shown in Figure 10 explains that both algorithms shown in Figure 9 achieved a precision of 1 and both formulae achieved a loss value of 0, with the x-axis representing epoch and the y-axis representing precision and loss values respectively. The accuracy of facial expressions (green line) and vocalizations (blue line) are shown in the graph above. To post a face photograph and see the results shown in Figures 11 to 13, click the "Predict Facial Emotion" button right now.

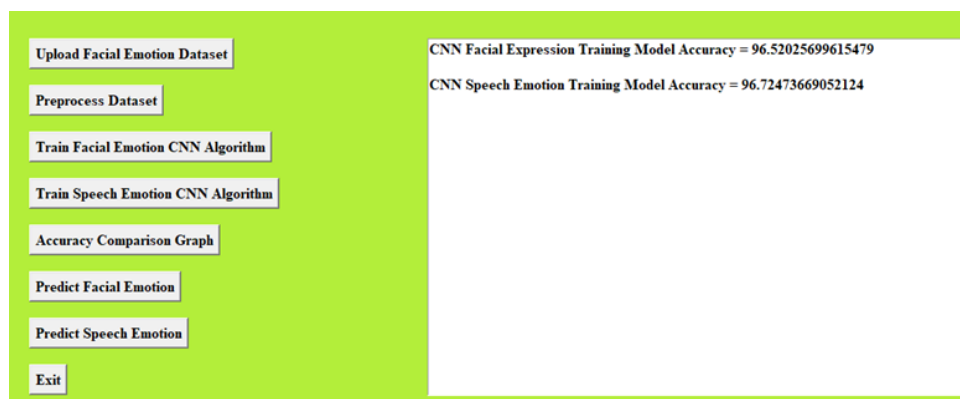


Figure 9. Algorithm results (CNN)

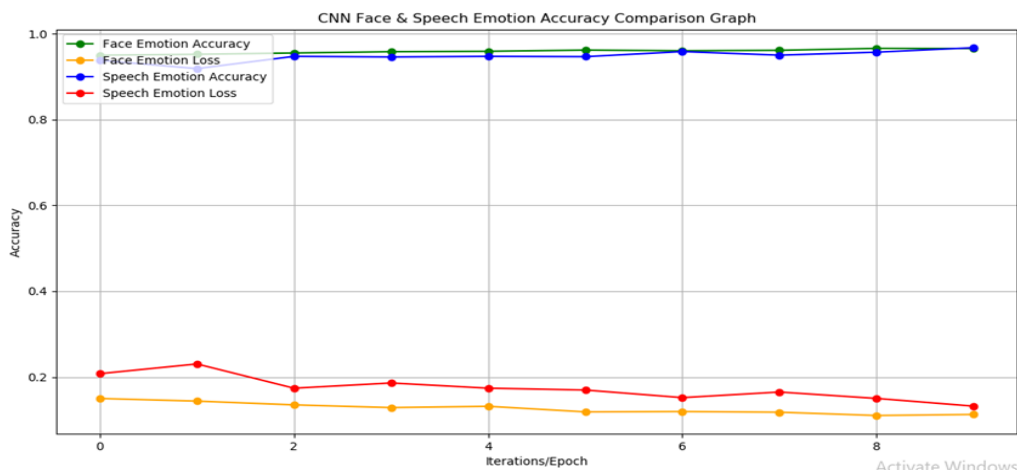


Figure 10. Emotion loss and accuracy

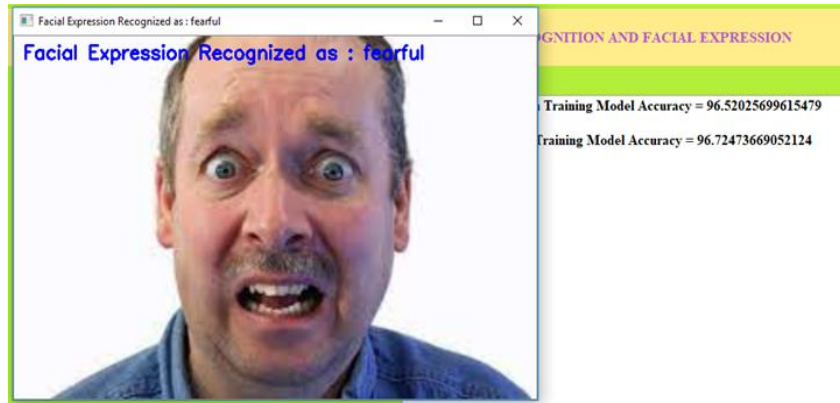


Figure 11. Test image results

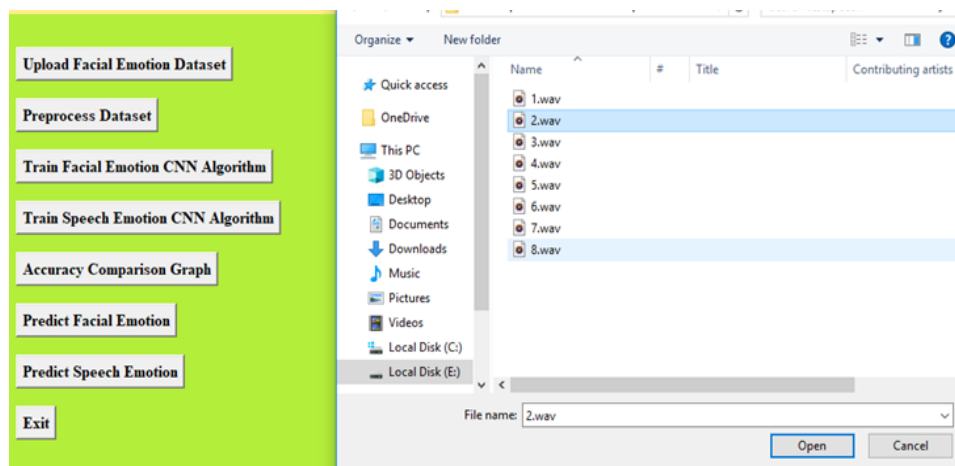


Figure 12. Audio test files



Figure 13. Speech emotion test

5. CONCLUSION

One promising avenue for advancement in emotional computing is the combination of facial and vocal analysis with deep learning's CNNs for emotion identification. By combining helpful emotional cues from speech signals with the robustness of CNNs for spatial function extraction, this method enables the detailed analysis of face emotions. The combination of both visual and auditory modalities improves the design's capacity to capture the nuance and complexity of human emotions. This makes the industrialized

system a powerful tool for fields as diverse as mental wellness monitoring and HCI because to its improved accuracy and generalizability across different datasets. The combination of facial and voice recognition in deep learning not only advances emotion discovery but also paves the path for better understanding and context-aware technology, which will allow expert systems to be more seamlessly integrated into many parts of our life.




REFERENCES

- [1] K. M. Kudiri, A. M. Said, and M. Y. Nayan, "Emotion detection through speech and facial expressions," *2014 International Conference on Computer Assisted System in Health*, Kuala Lumpur, Malaysia, 2014, pp. 26-31, doi: 10.1109/CASH.2014.22.
- [2] N. K. Gondhi and E. N. Kour, "A comparative analysis on various face recognition techniques," *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2017, pp. 8-13, doi: 10.1109/ICCONS.2017.8250626.
- [3] D. Ghimire, S. Jeong, S. Yoon, J. Choi, and J. Lee, "Facial expression recognition based on region specific appearance and geometric features," *2015 Tenth International Conference on Digital Information Management (ICDIM)*, Jeju, Korea (South), 2015, pp. 142-147, doi: 10.1109/ICDIM.2015.7381857.
- [4] V. Hosur and A. Desai, "Facial emotion detection using convolutional neural networks," *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Mysuru, India, 2022, pp. 1-4, doi: 10.1109/MysuruCon55714.2022.9972510.
- [5] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, 2017, pp. 2278-2288, doi: 10.1109/CVPRW.2017.282.
- [6] C. Wang, J. Zeng, S. Shan, and X. Chen, "Multi-task learning of emotion recognition and facial action unit detection with adaptively weights sharing network," *2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, 2019, pp. 56-60, doi: 10.1109/ICIP.2019.8802914.
- [7] Y. Kim and J. Kim, "Human-like emotion recognition: multi-label learning from noisy labeled audio-visual expressive speech," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 5104-5108, doi: 10.1109/ICASSP.2018.8462011.
- [8] Y. Tatebe, D. Deguchi, Y. Kawanishi, I. Ide, H. Murase, and U. Sakai, "Pedestrian detection from sparse point-cloud using 3DCNN," *2018 International Workshop on Advanced Image Technology (IWAIT)*, Chiang Mai, Thailand, 2018, pp. 1-4, doi: 10.1109/IWAIT.2018.8369680.
- [9] X. Xu *et al.*, "Survey on discriminative feature selection for speech emotion recognition," *The 9th International Symposium on Chinese Spoken Language Processing*, Singapore, 2014, pp. 345-349, doi: 10.1109/ISCSLP.2014.6936641.
- [10] F. Eyben *et al.*, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190-202, 2016, doi: 10.1109/TAFFC.2015.2457417.
- [11] P. Tosidis, N. Passalis, and A. Tefas, "Active vision control policies for face recognition using deep reinforcement learning," *2022 30th European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, 2022, pp. 1087-1091, doi: 10.23919/EUSIPCO55093.2022.9909691.
- [12] Z. Meng, S. Han, P. Liu, and Y. Tong, "Improving speech related facial action unit recognition by audiovisual information fusion," *IEEE Transactions on Cybernetics*, vol. 49, no. 9, pp. 3293-3306, Sept. 2019, doi: 10.1109/TCYB.2018.2840090.
- [13] S. L. Happy, A. George, and A. Routray, "A real time facial expression classification system using local binary patterns," *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, Kharagpur, India, 2012, pp. 1-5, doi: 10.1109/IHCI.2012.6481802.
- [14] H. Dwivedi, "Cryptocurrency sentiment analysis using bidirectional transformation," *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, Trichy, India, 2023, pp. 140-142, doi: 10.1109/ICSMDI57622.2023.00032.
- [15] K. M. Kudiri, A. M. Said, and M. Y. Nayan, "Emotion detection through speech and facial expressions," *2014 International Conference on Computer Assisted System in Health*, Kuala Lumpur, Malaysia, 2014, pp. 26-31, doi: 10.1109/CASH.2014.22.
- [16] L. Z. Ruiz, R. P. V. Alomia, A. D. Q. Dantis, M. J. S. San Diego, C. F. Tindugan, and K. K. D. Serrano, "Human emotion detection through facial expressions for commercial analysis," *2017 IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, Manila, Philippines, 2017, pp. 1-6, doi: 10.1109/HNICEM.2017.8269512.
- [17] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: a review," *IEEE Access*, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.
- [18] S. Begaj, A. O. Topal, and M. Ali, "Emotion recognition based on facial expressions using convolutional neural network (CNN)," *2020 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA)*, Tirana, Albania, 2020, pp. 58-63, doi: 10.1109/CoNTESA50436.2020.9302866.
- [19] K. Muhammad, S. Khan, J. D. Ser, and V. H. C. d. Albuquerque, "Deep learning for multigrade brain tumor classification in smart healthcare systems: a prospective survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 507-522, Feb. 2021, doi: 10.1109/TNNLS.2020.2995800.
- [20] N. Song, H. Yang, and P. Wu, "A gesture-to-emotional speech conversion by combining gesture recognition and facial expression recognition," *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, Beijing, China, 2018, pp. 1-6, doi: 10.1109/ACIIAsia.2018.8470350.
- [21] H. Razalli and M. H. Alkawaz, "Real-time face tracking application with embedded facial age range estimation algorithm," *2019 IEEE 9th International Conference on System Engineering and Technology (ICSET)*, Shah Alam, Malaysia, 2019, pp. 471-476, doi: 10.1109/ICSEngT.2019.8906420.
- [22] Kartali, M. Roglić, M. Barjaktarović, M. Đurić-Jovičić, and M. M. Janković, "Real-time algorithms for facial emotion recognition: a comparison of different approaches," *2018 14th Symposium on Neural Networks and Applications (NEUREL)*, Belgrade, Serbia, 2018, pp. 1-4, doi: 10.1109/NEUREL.2018.8587011.
- [23] Harimi, A. Shahzadi, and A. Ahmadyfard, "Recognition of emotion using non-linear dynamics of speech," *7th International Symposium on Telecommunications (IST'2014)*, Tehran, Iran, 2014, pp. 446-451, doi: 10.1109/ISTEL.2014.7000745.
- [24] S. Kim and H. Kim, "Deep explanation model for facial expression recognition through facial action coding unit," *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Kyoto, Japan, 2019, pp. 1-4, doi: 10.1109/BIGCOMP.2019.8679370.




- [25] Y. Fan, V. O. K. Li, and J. C. K. Lam, "Facial expression recognition with deeply-supervised attention network," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 1057-1071, 2022, doi: 10.1109/TAFFC.2020.2988264.
- [26] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 815-826, April 2019, doi: 10.1109/TASLP.2019.2898816.

BIOGRAPHIES OF AUTHORS






Ravi Gummula    is a Ph.D. scholar of electronics and communication engineering at the Dr. M.G.R. Educational and Research Institute in Chennai, Tamil Nadu, India. He obtained his M. Tech in VLSI design in 2012 from Shadan College of Engineering and Technology, and his BE in Electronics and Communication Engineering from Deccan College of Engineering and Technology in 2007. He is a Life Member in The Indian Society for Technical Education, The Institution of Electronics and Telecommunication Engineers and International Association of Engineers. He has organized several national and international seminars, workshops, and conferences. He can be contacted at email: ravi.gummula@gmail.com.



Dr. Vinothkumar Arumugam    is a professor of Electronics and Communication Engineering at the Dr. M.G.R. Educational and Research Institute in Chennai, Tamil Nadu, India. He obtained his Ph.D. in Machine Learning in 2017 and his M.Tech in applied electronics in 2010 from Dr.M.G.R. Educational and Research Institute, and his BE in Electronics and Communication Engineering from Anna University in 2008. He received a M.Sc. in real estate valuation from Annamalai University in 2016. He is a Chartered Engineer and Member of the Institution of Engineers (India), and he is recognized as an Approved Valuer and Member of the Institution of Valuers and a Member of various national and international professional societies. He can be contacted at email: dravinoth@gmail.com.



Abilasha Aranganathan    is an assistant professor of electronics and communication engineering at the Dr. M.G.R. Educational and Research Institute in Chennai, Tamil Nadu, India. She obtained her M.Tech in Nanotechnology in 2014 and her BE in Electronics and Communication Engineering in 2012 from Anna University. She has participated in several national workshops, seminars, and conferences. She has published several national and international journals and reputed publications. She is a Life Member in The Indian Society for Technical Education, The Institution of Electronics and Telecommunication Engineers and International Association of Engineers. She can be contacted at email: vabilasha90@gmail.com.