SN Computer Science (2023) 4:363
https://doi.org/10.1007/s42979-022-01633-9

ORIGINAL RESEARCH

Emotion Recognition System via Facial Expressions and Speech Using Machine Learning and Deep Learning Techniques

Aayushi Chaudhari[1] · Chintan Bhatt[2] · Thanh Thi Nguyen[3] · Nisarg Patel[4] · Kirtan Chavda[5] · Kalind Sarda[6]

Received: 19 August 2022 / Accepted: 22 December 2022

Abstract

Patients in hospitals frequently exhibit psychological issues such as sadness, pessimism, eccentricity, and anxiety. However, hospitals normally lack tools and facilities to continuously monitor the psychological health of patients. It is desirable to identify depression in patients so that it can be managed by instantly providing better therapy. This can be possible by advances in machine learning for image processing with notable applications in the domain of emotion recognition using facial expressions. In this paper, we have proposed two different methods, i.e. facial expression detection and voice analysis, to predict emotions. For facial expression recognition, we have used two approaches, one is the use of Gabor filters for feature extraction with support vector machine for classification and another is using convolutional neural network (CNN). For voice analysis, we extracted mel-frequency cepstral coefficients from speech data and, based on those features, predicted the emotions of the speech using a CNN model. Experimental results show that our proposed emotion recognition methods obtained high accuracy and thus could be potentially deployed to real-world applications.

Keywords Facial emotion · Speech · Expressions · Deep learning · Machine learning · SVM · CNN

Introduction

It is significantly important to understand the human psychology in today's era. Studies have shown that an expression can be useful for reading the human mind. When patients have neuropsychiatric illnesses that impede expressions of emotion and recognition, facial expressions

This article is part of the topical collection "Enabling Innovative Computational Intelligence Technologies for IOT" guest edited by Omer Rana, Rajiv Misra, Alexander Pfeiffer, Luigi Troiano and Nishtha Kesswani.

✉ Chintan Bhatt
chintan.bhatt@sot.pdpu.ac.in

Aayushi Chaudhari
aayushichaudhari.ce@charusat.ac.in

Thanh Thi Nguyen
thanh.nguyen@deakin.edu.au

Nisarg Patel
imnisarg3@gmail.com

Kirtan Chavda
kirtanc25@gmail.com

Kalind Sarda
kalind.sarda@gmail.com

[1] U & P U. Patel Department of Computer Engineering, Chandubhai S. Patel Institute of Technology, Charotar University of Science And Technology (CHARUSAT), Changa, Gujarat 388421, India

[2] Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, Knowledge Corridor, Gandhinagar, Gujarat 382007, India

[3] School of Information Technology, Faculty of Science, Engineering and Built Environment, Deakin University, Victoria 3216, Australia

[4] Bishop University, 2600 College St, Sherbrooke, QC J1M 1Z7, Canada

[5] Crest Data Systems Private Limited, Gandhinagar Highway, Makarba, Ahmedabad, Gujarat 380051, India

[6] Programmer Analyst, Meditab Software Pvt Ltd, Kalasagar Mall, Ghatlodiya, Ahmedabad, Gujarat 380061, India

Published online: 28 April 2023

may play a critical role in the assessment of those conditions [1]. Psychologists have developed seven different approaches for studying human behaviour: self-observation, observation, experimental, survey, clinical, genetic, and testing procedures. Every method is useful in different scenarios of human life. Facial expression detection is a complex field, and it has significant applications in many areas, including data-driven animation and human–computer interaction [2]. If we consider a situation that we should identify the psychology of the patient, we can use an introspection method, i.e. self-observation, where the patient will report the pain or any disturbances observed within their body. Another observation method can also be used in this scenario where the doctors can observe the movements of the patient using cameras and sensors or automated medical devices. These tools can continuously monitor the medical parameters of the patient to identify the mental health of the patient, but they will not be aware of being observed at any moment. This would lead to the observation of the natural behaviour of the patient. Clinical method can also be used by collecting or reviewing the past history of the patient's disease and based on that further treatment can be done [3]. As per current trends and technology, we can move forward to improving the observation method for human psychology identification using facial expressions as well as speech of human [4]. Facial expression in humans plays the most important role in identifying the psychic condition of the human mind and can be used for non-verbal communications. The primary goal of this study was to use an automated system to observe facial expressions and human speech to determine the state of the human mind. This will include identifying the universal facial emotions like joy, sadness, happiness, fear of humans from images and video streams based on their expressions and speech. Without requiring any manual input, the feature extraction can be set up to meet real-time requirements [6].

The human face is the most exposed portion of the body, so computer vision systems (cameras) can capture an image or video stream to assess the human's emotion, which can be used to understand the human's mental state [7]. These systems can detect the shape of eyebrows and lips to determine the expressions of the human beings and identify their mental health. In this paper, we have used images as well as human speech to identify the emotion and mental health of humans [8]. A long-standing and difficult issue in the field of artificial intelligence is the ability to recognize emotions from speech. On emotion recognition, numerous important studies have been conducted [5]. A quick survey of the current systems for identifying human emotion is given in [9]. The main difficulties in emotion recognition include selecting the emotion recognition corpora (speech database), identifying various speech features, and selecting the right classification model [10].

Literature Review

Review on Emotion Recognition Using Images and Videos

Based on facial expressions, Tarnowski et al. performed the identification of seven fundamental human emotional states: neutral, happiness, surprise, anger, fear, sadness, and disgust [7]. Here, the authors employed the 30 frame-per-second Microsoft Kinect for 3D face modelling. This model is built on 121 points on the face that are grouped on distinctive facial features like the corners of the mouth, the nose, the cheekbones, eyebrows, and so on. A matrix is used to hold the spatial coordinates of the points. The experiments were conducted on people whose ages were between 26 and 50. All users' classification accuracy was 96% when using k-nearest neighbour method, 90% when using multilayer perceptron (MLP) with random division of data, and 73% when using MLP with subject-dependent division of data [7, 11].

Jie and Yongsheng proposed the multi-view facial expression light weight network (MVFE-LightNet), which considers non-frontal facial features for multi-view facial expression recognition [12]. multi-task cascaded convolutional neural networks (MTCNN) were applied first for facial detection and then normalization and data augmentation were applied to it. For experiments, the Radboud Faces Database and the Binghamton University 3D Facial Expression (BU-3DFE) datasets were used. For the Radboud and BU-3DFE, the proposed model's accuracy was 95.6% and 88.7%, respectively [12]. Chen et al. proposed another approach for facial expression recognition based on facial motion detection [13]. That method improved the facial emotion recognition (FER) framework by introducing novelty in the face movement detection. As a result, the authors here incorporated prior domain knowledge by including the typical differences between neutral expressions and expressive faces to make the learning of facial motion masks easier [14]. The proposed method showed great performance when experimenting with the Extended Cohn–Kanade dataset (CK+) dataset and the wild dataset, namely, AffectNet [13, 15].

Meng et al. [16] presented a method for identifying facial expressions in videos. Facial expressions can be used to categorize seven basic emotions, including joy, sad, fear, disgust, surprise, and neutral. It considers only visual views, excluding audio from the videos. The proposed method is named Frame Attention Networks that contain a combination of self-attention and relational attention models [16]. On the other hand, for individuals with behavioural

issues, Verma et al. proposed a wearable system based on an Arduino platform for detecting emotions [17]. The authors used a galvanic skin response (GSR) sensor and a pulse sensor to record body parameters, identify the wearable's emotional state, and display it using a light indicator mounted on the wearable. This allows the patient's various emotions to be represented by the indicator's colour, and thus other people can be aware of the patient's mood.

With the use of a statistical technique, the authors in [18] analysed several fundamental issues in face expression analysis. They employed statistical research to identify the locations and attributes most useful for facial expression recognition. The usefulness of the landmark-based strategy for categorizing an expression was investigated using a standard database in two different scenarios: (a) when a face with a neutral expression is provided, and (b) when there is no a priori information about the face. In another work, Wang et al. proposed a mobile application that uses the FER models on a device and works out real-time results [19]. Likewise, Yang et al. introduced a method known as "De-expression residue learning" to extract information from the expressive component of a facial expression in order to recognize the expressions [20].

Review on Emotion Recognition Using Speech

Reviewing Datasets (Speech Corpora)

A good emotion recognition system should have access to a context-rich emotional speech database (corpora). The choice of a suitable speech database is important for designing and developing the system. Three types of corpora are typically used [26, 27], which are as follows:

1. Elicited emotional speech databases: Speech corpora that have been elicited to be emotional are known as elicited emotional speech databases. Although this type of database has the benefit of being quite similar to a natural database, there are some drawbacks as well.
2. Actor-based speech databases: These speech datasets are gathered from skilled and experienced professionals. These forms of data are relatively simple to gather, and the corpora contain a wide range of emotions.
3. Natural speech databases: Data from the real world are used to develop such a database. Such information is natural and very helpful for identifying actual emotions. The issue is that it comprises background noise and that not all emotions may be present.

Depending on the goal of the study, an emotional speech database may be utilized. A large number of both male and female speakers of actual and realistic emotional speech must make up the corpus for an emotion recognition system to be effective [10].

Reviewing Speech Features

A speech's structural elements can be used to distinguish between various emotional assertions. Different components of a speech signal serve to depict the vocal tract and hearing system characteristics of a human [28]. Different prosodic and acoustic elements from a speech signal must be extracted to develop an emotion recognition system. Pitch, amplitude, formants, and spectral properties are some significant characteristics of the voice signal [29].

1. The vocal cord trembling is what gives the signal its pitch. A change in frequency can be used to gauge the pitch. The fundamental frequency or pitch frequency is the number of vibrations in one unit of time, while the period of pitch is the interval of time between two consecutive vibrations of the vocal cords [30].
2. The amplitude of the signal at different intervals serves as a representation of a sound's loudness (energy), as heard by the human ear.
3. The distinct frequency peaks in the speech signal are known as formants. For the formant extraction, a spectral root group delay function technique might be used [29].
4. The source signal excites the vocal tract, which then produces speech signals. System features or spectrum characteristics are terms used to describe features that are measured from the vocal tract system.

According to Palo et al. [28], the most prevalent spectral features employed by various emotion recognition algorithms are linear prediction coefficients (LPCs), mel-frequency cepstral coefficients (MFCCs), and linear prediction cepstral coefficients (LPCC). Based on the limits of human hearing, mel-frequency cepstral coefficients are calculated. Two different types of filters are applied in the MFCC approach. Some filters have linear spacings at frequencies below 1 kHz, while others have logarithmic spacings at frequencies above 1 kHz [28].

Reviewing Speech Classifiers

A classification system is a method for assigning each speech to the appropriate emotion category in accordance with the features of speech that are retrieved. Different classifiers are available for recognizing emotions. There is no general guideline for picking a reliable classifier. Most of the time, the classifier chosen is based on precedent. Each speech sample's feature (a feature vector) is provided to classifiers as input along with a linear combination of the real

weight vector. A suitable training technique is then used to alter this weight vector. The model's output is then generated using an activation function, which translates each input to a certain class of emotions. Both linear and non-linear activation functions are possible.

Function classifiers can be divided into two categories: linear and non-linear classifiers, depending on the type of activation. If the feature vectors can be linearly separated, the linear classifier will classify them effectively.

As most feature vectors in real-world scenarios cannot be separated linearly, a nonlinear classifier is preferable [27]. The support vector machine (SVM), Gaussian mixture model (GMM), multi-layer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), K-nearest neighbours (KNN), and hidden Markov model are just a few of the nonlinear classifiers available for emotion recognition.

In this paper, we have proposed individual models for detecting emotions from images and speech, and based on that we can identify whether the person's state of mind is stable or depressed. It is very much necessary to identify the patient's state of mind to provide them with appropriate treatment and counsel. We have experimented our model on image and audio, in which one patient is communicating with a psychiatrist about his problems and was about to commit suicide, which proves that the patient is depressed. For emotion detection using facial expressions, we used CNN and SVM. For voice analysis, we extracted MFCC features from speech and then, based on those features, predicted the emotion of the speech using CNN.

Experimental Datasets

Datasets for Emotion Recognition from Images

In this research, we used three different datasets, as follows:

1. JAFFE (Japanese Female Facial Expression): The JAFFE dataset includes 213 pictures of various facial expressions made by 10 different Japanese female individuals. Each participant was asked to make 7 facial expressions (6 basic and 1 neutral), and 60 annotators rated each facial emotion on average in terms of its semantic content [21].
2. Kaggle face expression recognition dataset: Kaggle dataset for face emotion recognition showcases faces in grayscale, measured by 48×48 pixels. Because the faces are automatically registered, each image has a face that is roughly centred and takes up about the same amount of space. Each graphic represents a face expression in one of seven categories (0 = angry, 1 = disgust, 2 = fear, 3 = happy, 4 = sad, 5 = surprise, 6 = neutral). Approximately, 36K photos make up the dataset.
3. EMOTions In Context (EMOTIC): The EMOTIC dataset is a collection of images of people taken in the real world and labelled with the emotions they were showing at the time. The captions for the photos include a thorough list of 26 different emotion kinds as well as the three continuous dimensions valence, arousal, and dominance [22].

Datasets for Emotion Recognition from Audio

The following audio datasets are used in our experiments.

1. Ryerson audio-visual database of emotional speech and song (RAVDESS): This is a comprehensive dataset that includes voice and music, audio, and video (24.8 GB). The RAVDESS contains 1440 files resulting from 60 trials multiplied by 24 actors. In the RAVDESS dataset, actors were asked to perform two lexically related phrases with neutral North American accents. The various emotions that can be expressed through speech include calmness, joy, sadness, anger, fear, surprise, and contempt. For each expression, there is a neutral expression and two emotional intensity levels (strong and normal) [23].
2. Toronto emotional speech set (TESS): This dataset consists of 2800 data points (audio files), containing recordings of the set acting out each of the seven emotions and a set of 200 target phrases said by two actresses (aged 26 and 64) in the carrier phrase "Say the word" (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). The dataset is built up so that each of the two actresses, with the emotions they portray, is housed in its own folder, which also contains the audio files for all 200 target words.
3. Crowd sourced emotional multimodal actors dataset (CREMA-D): There were 7442 pieces of original videos from 91 performers included in the CREMA-D data collection. There were 48 men and 43 women among the actors who appeared in these clips, ranging in age from 20 to 74, and represented various races and ethnicities (African, American, Asian, Caucasian, Hispanic, and unspecified). The performers read a selection of 12 sentences. The phrases were delivered using four different emotion levels (low, medium, high, and unspecified) and six different emotions (angry, disgust, fear, joyful, neutral, and sad) [24].
4. Surrey audio-visual expressed emotion (SAVEE): The SAVEE database is a necessity for developing an automatic emotion recognition system. The collection includes 480 British English words spoken by four male actors as they were acting out seven different moods.

The sentences were from the normal TIMIT corpus and were phonetically balanced for each mood [25].

Models for Emotion Recognition Using Images

In the proposed models, we extracted the facial features and audio speech individually and processed them further to verify the accurate emotion within a person. For person detection within the image, we employed the well-established Viola Jones algorithm. This algorithm's crucial feature is its ability to detect objects more quickly. It uses Haar-based feature filters, and due to this, it does not use multiplications of images. The face inside the frame will be detected and cropped. At each stage, a different part of face is detected. Main parts such as the lips, nose, and eyes of the face are detected at different levels, then a window will be formed around the face, and if the facial part is not detected at any particular level then that window will be rejected.

The SVM Approach

Processing the Detected Faces

For texture analysis in image processing, we employed the Dennis Gabor-inspired Gabor filter, a linear filter that effectively determines whether the image contains any particular frequency content in a particular direction, in a confined region around the point or region of analysis. We employed a bank of 40 Gabor filters in this system. This filter is used to do the feature extraction from the frames. The hyperparameters are set as follows: theta is in the range (0, π, π/8), lambda is in the range (0, $\pi\times$ 6/4, π/4), gamma is equal to 0.5, and phi is equal to 0.0.

Dimensionality Reduction

Attributes present in our data are known as dimensionality of our data. If this dimensionality increases within the system, it can make the system complex. So, to reduce the storage space and time, dimensionality reduction can be done by various methods. A model's complexity can be decreased through the dimensionality reduction, which can also prevent overfitting. Feature selection and feature extraction are the two fundamental divisions of dimensionality reduction. In contrast to feature extraction, which uses data from the feature set to create a new feature subspace, feature selection allows us to select a subset of the original features. We used principal component analysis (PCA) to reduce the 80 features we received through the Gabor filter into 30 features. PCA consists of three different steps: scaling the vectors, formation of the covariance matrix, and calculating eigenvalues and vector.

Classification Using SVM

After the dimensional reduction using the PCA, we will apply the SVM model for classification purposes (see Fig. 1). The classification will be performed for two different classes, angry and happy. We used the three hyperparameters which were stated in the reference paper, i.e. kernel-'rbf', gamma-'scale', tol-'0.001'. By using this setup, we obtained a test accuracy of around 61%. For training the model, we used three different datasets: (1) Japanese Female Facial Expression (JAFFE) dataset, (2) Kaggle dataset for face emotion recognition, and (3) EMOTions in Context (EMOTIC). We used 80:20 for the train:test split.

The CNN Approach

Face recognition entails the extraction of various facial features and their classification to distinguish one person's face from another. Because faces are intricate, multidimensional visual stimuli and because the face recognition rate depends on differences in position, expression, occlusion, resolution, and illumination, it is a difficult task. In the presence of significant degree differences in human face photos, most face recognition algorithms now in use perform poorly. Therefore, to enhance the performance of face recognition

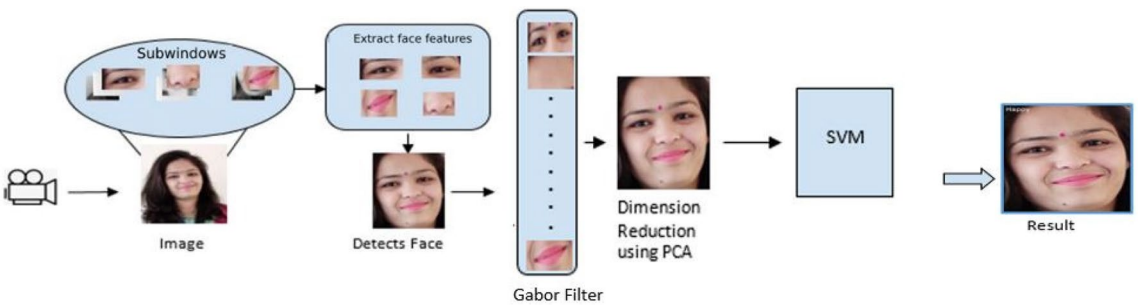


Fig. 1 Architecture for emotion identification from images using SVM

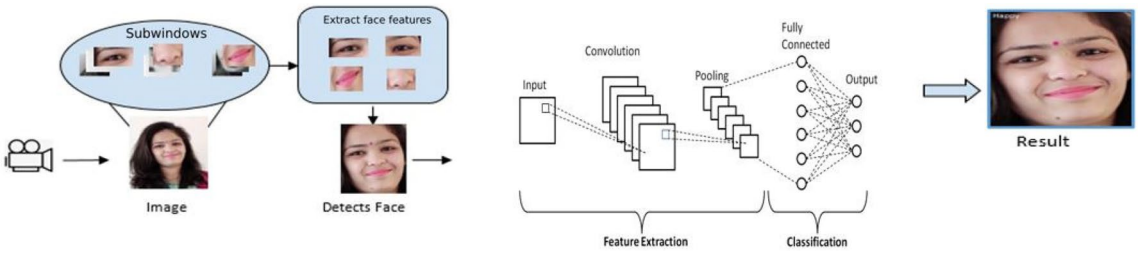


Fig. 2 CNN-based approach for emotion recognition from images

with the aforementioned modifications, we present a deep learning-based method that uses CNN models for categorization (see Fig. 2).

The hyperparameters are as follows: sigma is set to 1.0, and the kernel size (ksize) is set to (5,5). Theta pulls values out of the array (0, $\pi/4$, $\pi/2$, $\pi\times3/4$, π, $\pi\times 5/4$, $\pi\times3/2$, 2π), while lambda pulls values out of the array (2,3,6). We used a function called "lin-space" from the NumPy library which expands this array and gives us a bigger array with numbers from 2 to 3 having a gap of 6 in between them. Gamma is set to 1.0, phi has kernels stacked upon each other, namely, real and imaginary.

The real kernel had phi value of 0, while the imaginary kernel had phi value equal to $\pi/2$. These values of the hyperparameters were taken from the reference paper we referred to for developing the custom filter. On training the model with 50 epochs, we achieved a training accuracy of 99% and a validation accuracy of 94%.

Compared with the results obtained from the SVM approach presented in Sect. 4.1, we can signify that deep learning approaches can provide far better accuracy than traditional machine learning approaches when applied to recognize emotions from images.

Emotion Recognition Using Speech

The phrase "mel-frequency cepstrum" (MFC) refers to a representation of a sound's short-term power spectrum that is employed in sound processing and is based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Numerous coefficients known as mel-frequency cepstral coefficients (MFCCs) make up an MFC. They were produced using a nonlinear "spectrum-of-a-spectrum", which is the cepstral representation of an audio clip. The mel-frequency cepstrum, in contrast to the standard cepstrum, uses frequency bands that are uniformly spaced on the mel scale, which more closely approximate the response of the human hearing system. With the use of this frequency warping, sound can be represented more accurately during audio compression.

Python is used to process and extract the features from audio files using the Librosa package. It offers the components required to build music data retrieval systems. We were capable of obtaining MFCC features using the Librosa library.

MFCCs are frequently used in speaker and automatic voice recognition [31]. A signal's MFCCs are a small group of features (often 10–20) that succinctly represent the general contours of a spectral envelope. It displays the amplitude changes over a predetermined period. Time is represented on the x-axis (horizontally), while the signal's amplitude can be measured on the y-axis (vertically). The CNN receives these generated spectrograms as inputs. Figure 3 presents the proposed approach for emotion recognition from speech using CNN. The CNN architecture includes six convolutional layers, a dense layer and, a softmax layer. The CNN model uses ReLU as a nonlinear activation function. To prevent the overfitting problem, dropout layers are inserted after fully connected layers. Figure 4 presents the speech frequencies in Mel spectrogram and MFCC spectrogram forms.

Figure 5 showcases the sample signal frequency representations of happy and angry voices. It can be seen that there is a difference between happy and angry emotions in terms of signal frequency representations. Figure 6 represents the confusion matrix including the combination of RAVDESS, SAVEE, TESS, and CREMA-D datasets which helps in understanding human emotion accurately and getting the statistical classification of various emotions.

Figure 7 shows the graphical presentation of accuracy for various combinations of datasets along with and without augmentation. Using the combination of all four datasets,

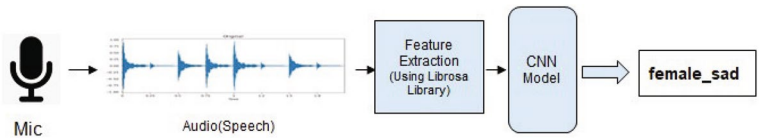
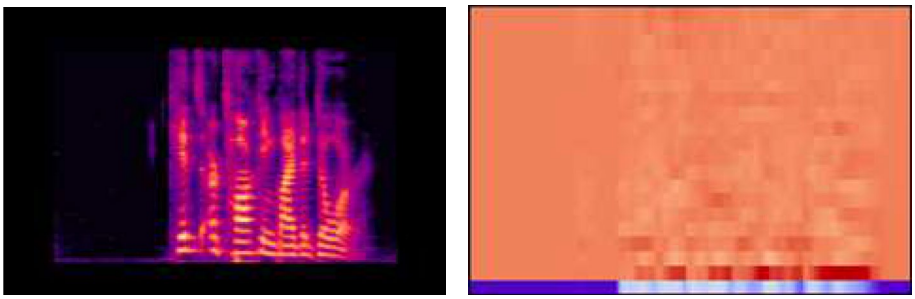
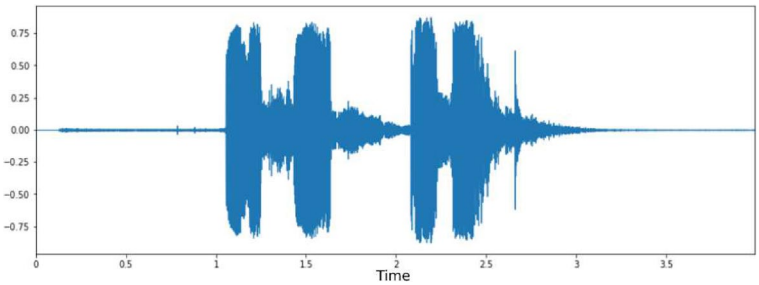


Fig. 3 Architecture of emotion recognition using human speech

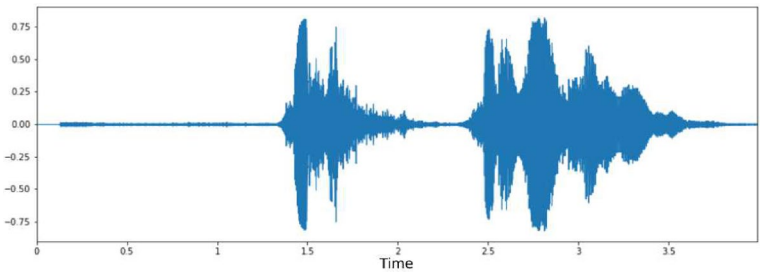


a) Mel Spectrogram b) MFCC Spectrogram

Fig. 4 Representations of speech frequencies in two different forms



a) Speech signal data of Happy voice



b) Speech signal data of Angry voice

Fig. 5 Representations of speech signals for different classes: **a** happy and **b** angry

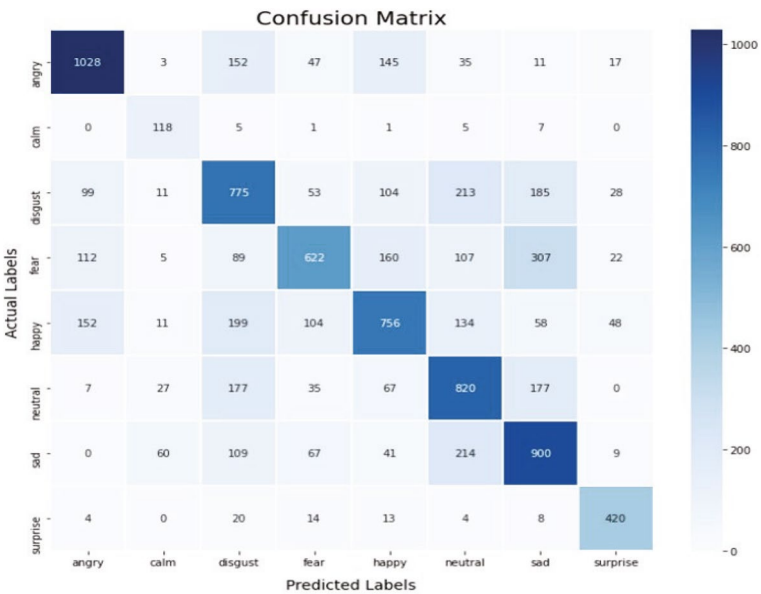


Fig. 6 RAVDESS, SAVEE, TESS and CREMA-D's confusion matrix with augmentation

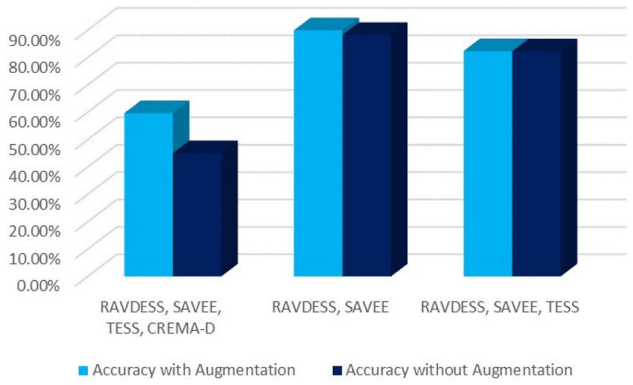


Fig. 7 Graphical representation of accuracy in the datasets

we got an accuracy of 59.62% on applying augmentation, while without augmentation, we got an accuracy of 45%. The combination of two datasets RAVDESS and SAVEE provides an accuracy of 89.87%.

Figures 8 and 9 present the confusion matrices generated by performing a test on the RAVDESS, SAVEE and TESS datasets, bifurcating the emotion categories into male and female gender voices. Figure 8 shows results obtained without data augmentation while Fig. 9 is for results with augmentation. Dealing with female voices has been challenging because of the heavy tones, which were considered under the characteristics of male voices. As a result, we attain an

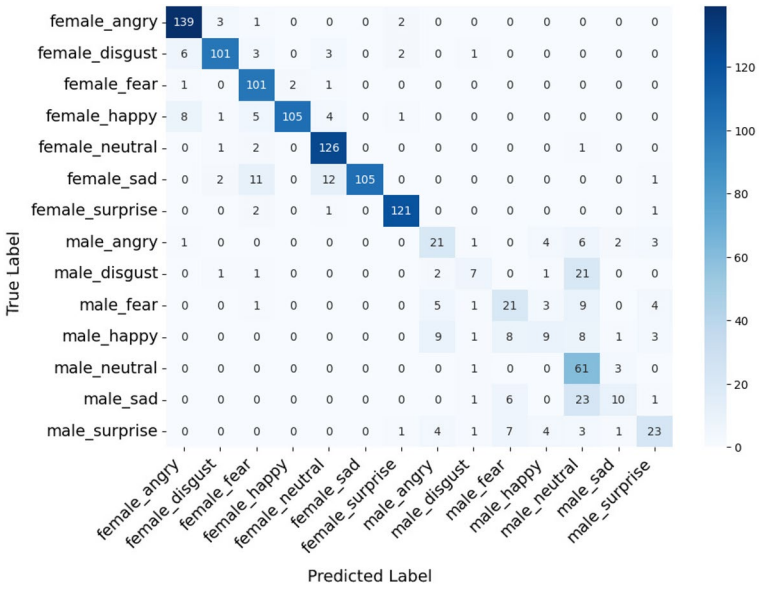


Fig. 8 RAVDESS, SAVEE, and TESS's confusion matrix without augmentation

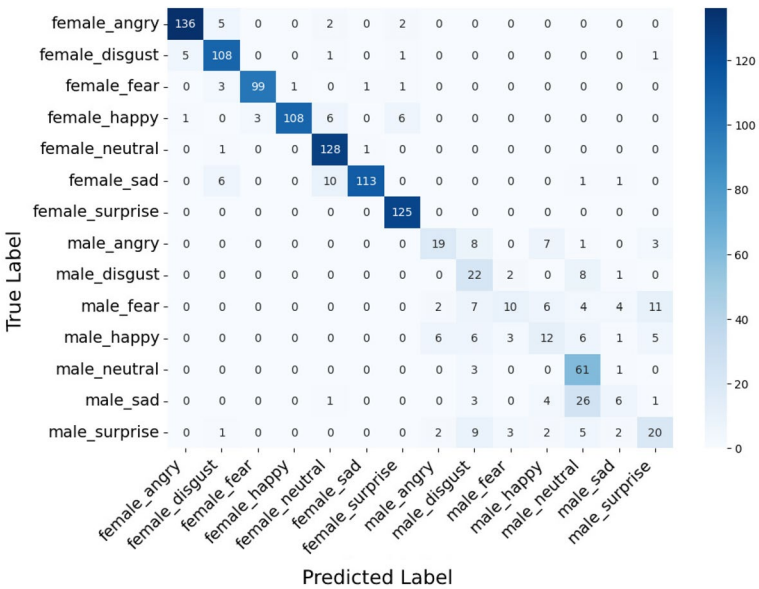


Fig. 9 RAVDESS, SAVEE, and TESS's confusion matrix with augmentation

accuracy of 82.29% without data augmentation as shown in Fig. 8 and 82.20% with data augmentation in Fig. 9.

Conclusions

In this research, we tested various deep learning and machine learning models on different image and audio datasets with six emotion categorization classes to understand how these algorithms function and to expand our understanding towards multimodal approach to emotion recognition. Using these six categories, clinicians can determine whether a patient is depressed, anxious, or sad. We have employed two alternative methods for extracting emotions from images. One of them is a machine learning technique that uses the Gabor filter and SVM to extract facial features from images and classify emotions. The CNN deep learning model combined with Gabor filters and PCA is used to recognize emotions as a second method. Additionally, using the Librosa library, we have experimented with other combinations of audio datasets to identify emotions. Using a multimodal emotion identification approach, our future study will combine the results of emotion categorization from audio and images.

Author Contributions AC data curation, investigation, writing—original draft, implementation. CB methodology, project administration, writing—review and editing. TTN writing—review and editing. NP data curation, investigation, implementation, testing. KP methodology, writing—review and editing, implementation. KS methodology, writing—review and editing, implementation.

Funding The research received no external funding.

Availability of data and materials Data is available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analysed in this study. The data can be found here: [https://www.kasrl.org/jaffe_download.html] (accessed on 12th March 2022), [https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio] (accessed on 13th April 2022), [https://www.kaggle.com/datasets/barelydedicated/savee-database] (accessed on 13th April 2022), [https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess] (accessed on 15th April 2022), and [https://www.kaggle.com/datasets/ejlok1/cremad] (accessed on 17th April 2022).

Code Availability The code is available at the following github link: https://github.com/AayushiChaudhari5694/EmotionRecognition_Image_Speech.git.

Declarations

Conflict of Interest The authors declared no conflict of interest.

Ethical Approval Not applicable.

Consent to Participate All authors have read and agreed to participate in the publication of the manuscript.

Consent for Publication All authors have read and agreed to publish the latest version of the manuscript.

References

1. Sonawane B, Sharma P. Deep learning based approach of emotion detection and grading system. Pattern Recognit Image Anal. 2020;30(4):726–40.
2. Kim DJ. Facial expression recognition using ASM-based post-processing technique. Pattern Recognit Image Anal. 2016;26(3):576–81.
3. Muhammad K, Khan S, Kumar N, Del Ser J, Mirjalili S. Vision-based personalized wireless capsule endoscopy for smart healthcare: taxonomy, literature review, opportunities and challenges. Futur Gener Comput Syst. 2020;113:266–80.
4. Pisor AC, Gervais MM, Purzycki BG, Ross CT. Preferences and constraints: the value of economic games for studying human behaviour. R Soc Open Sci. 2020;7(6): 192090.
5. Le DN, Nguyen GN, Van Chung L, Dey N. MMAS algorithm for features selection using 1D-DWT for video-based face recognition in the online video contextual advertisement user-oriented system. J Glob Inf Manag (JGIM). 2017;25(4):103–24.
6. Panning A, Al-Hamadi AK, Niese R, Michaelis B. Facial expression recognition based on Haar-like feature detection. Pattern Recognit Image Anal. 2008;18(3):447–52.
7. Tarnowski P, Kołodziej M, Majkowski A, Rak RJ. Emotion recognition using facial expressions. Proc Comput Sci. 2017;108:1175–84.
8. Le DN, Nguyen GN, Bhateja V, Satapathy SC. Optimizing feature selection in video-based recognition using Max-Min Ant System for the online video contextual advertisement user-oriented system. J Comput Sci. 2017;21:361–70.
9. Rozaliev VL, Orlova YA. Motion and posture recognition for identifying human emotional reactions. Pattern Recognit Image Anal. 2015;25(4):710–21.
10. Basu S, Chakraborty J, Bag A, Aftabuddin M. A review on emotion recognition using speech. In: 2017 international conference on inventive communication and computational technologies (ICICCT). IEEE; 2017. p. 109–114.
11. Liu M, Shan S, Wang R, Chen X. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 1749–1756.
12. Jie S, Yongsheng Q. Multi-view facial expression recognition with multi-view facial expression light weight network. Pattern Recognit Image Anal. 2020;30(4):805–14.
13. Chen Y, Wang J, Chen S, Shi Z, Cai J. Facial motion prior networks for facial expression recognition. In: 2019 IEEE visual communications and image processing (VCIP). IEEE; 2019. p. 1–4.
14. Hibare R, Vibhute A. Feature extraction techniques in speech processing: a survey. Int J Comput Appl. 2014;107(5).
15. Meng Z, Liu P, Cai J, Han S, Tong Y. Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017). IEEE; 2017. p. 558–565.
16. Meng D, Peng X, Wang K, Qiao Y. Frame attention networks for facial expression recognition in videos. In: 2019 IEEE international conference on image processing (ICIP). IEEE; 2019. p. 3866–3870.
17. Verma A, Dogra A, Malik K, Talwar M. Emotion recognition system for patients with behavioral disorders. In: Intelligent

communication, control and devices. Singapore: Springer; 2018. p. 139–145.
18. Alugupally N, Samal A, Marx D, Bhatia S. Analysis of landmarks in recognition of face expressions. Pattern Recognit Image Anal. 2011;21(4):681–93.
19. Wang X, Huang J, Zhu J, Yang M, Yang F. Facial expression recognition with deep learning. In: Proceedings of the 10th international conference on internet multimedia computing and service. 2018. p. 1–4.
20. Yang H, Ciftci U, Yin L. Facial expression recognition by deexpression residue learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 2168–2177.
21. Kamachi M, Lyons M, Gyoba J. The Japanese Female Facial Expression (JAFFE) database. 1997. http://www.kasrl.org/jaffe.html.
22. Kosti R, Alvarez JM, Recasens A, Lapedriza A. Context based emotion recognition using emotic dataset. IEEE Trans Pattern Anal Mach Intell. 2019;42(11):2755–66.
23. Livingstone SR, Russo FA. The Ryerson audio-visual database of emotional speech and song (RAVDESS) [Data set]. In: PLoS ONE 2018;(1.0.0, Vol. 13, Number 5, p. e0196391). Zenodo. https://doi.org/10.5281/zenodo.1188976.
24. Cao H, Cooper D, Keutmann M, Gur R, Nenkova A, Verma R. CREMA-D: crowd-sourced emotional multimodal actors dataset. IEEE Trans Affect Comput. 2014;5:377–90.
25. Haq S, Jackson PJB. Multimodal emotion recognition. In: Wang W, editor. Machine audition: principles, algorithms and systems. Hershey: IGI Global Press; 2010. p. 398–423. https://doi.org/10.4018/978-1-61520-919-4.
26. El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. 2011;44(3):572–87.
27. Koolagudi SG, Rao KS. Emotion recognition from speech: a review. Int J Speech Technol. 2012;15(2):99–117.
28. Palo HK, Chandra M, Mohanty MN. Emotion recognition using MLP and GMM for Oriya language. Int J Comput Vis Robot. 2017;7(4):426–42.
29. Murthy HA, Yegnanarayana B. Formant extraction from group delay function. Speech Commun. 1991;10(3):209–21.
30. Choudhary A, Govil MC, Singh G, Awasthi LK. Workflow scheduling algorithms in cloud environment: a review, taxonomy, and challenges. In: 2016 4th international conference on parallel, distributed and grid computing (PDGC). IEEE; 2016. p. 617–624.
31. Albu F, Hagiescu D, Vladutu L, Puica MA. Neural network approaches for children's emotion recognition in intelligent learning applications. In: Proceedings of the 7th international conference on education and new learning technologies (EDULEARN15). 2015. p. 3229–3239.