IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 25, 2021, accepted January 18, 2022, date of publication January 26, 2022, date of current version February 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3146729

Emotion Recognition With Audio, Video, EEG, and EMG: A Dataset and Baseline Approaches

JIN CHEN[1], (Member, IEEE), TONY RO[2], AND ZHIGANG ZHU[1,3], (Senior Member, IEEE)
[1]Computer Science Department, The City College of New York (CUNY), New York, NY 10031, USA
[2]Programs in Psychology, Biology, and Cognitive Neuroscience, The Graduate Center, CUNY, New York, NY 10016, USA
[3]Doctoral Program in Computer Science, The Graduate Center, CUNY, New York, NY 10016, USA

Corresponding author: Jin Chen (jchen025@citymail.cuny.edu)

This work was supported in part by the NSF Emerging Frontiers in Research and Innovation Program (EFRI) under Award 1137172, in part by the NSF Division of Behavioral and Cognitive Sciences (BCS) under Award 1358893 and Award 1755477, in part by the Air Force Office of Scientific Research (AFOSR) under Award FA9550-21-1-0082, and in part by the Office of the Director of National Intelligence (ODNI) via the Intelligence Community Center for Academic Excellence (IC CAE) at Rutgers University under Grant HHM402-19-1-0003 and Grant HHM402-18-1-0007.

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the University Integrated Institutional Review Board at the City University of New York under IRB File #299876, and performed in line with Mobility Skill Acquisition and Learning through Alternative and Multimodal Perception for Visually Impaired People.

ABSTRACT This paper describes a new posed multimodal emotional dataset and compares human emotion classification based on four different modalities - audio, video, electromyography (EMG), and electroencephalography (EEG). The results are reported with several baseline approaches using various feature extraction techniques and machine-learning algorithms. First, we collected a dataset from 11 human subjects expressing six basic emotions and one neutral emotion. We then extracted features from each modality using principal component analysis, autoencoder, convolution network, and mel-frequency cepstral coefficient (MFCC), some unique to individual modalities. A number of baseline models have been applied to compare the classification performance in emotion recognition, including k-nearest neighbors (KNN), support vector machines (SVM), random forest, multilayer perceptron (MLP), long short-term memory (LSTM) model, and convolutional neural network (CNN). Our results show that bootstrapping the biosensor signals (i.e., EMG and EEG) can greatly increase emotion classification performance by reducing noise. In contrast, the best classification results were obtained by a traditional KNN, whereas audio and image sequences of human emotions could be better classified using LSTM.

INDEX TERMS Emotion recognition, data collection, electroencephalography, electromyography.

I. INTRODUCTION

In daily life, emotions are abound and there are countless reasons for determining someone's emotional state, including for better communication and work efficiency. In the product development process, product features and design can be determined to be more suitable for users by analyzing their emotional states during their user experience. In medical care, caregivers can provide better care to patients if their emotional states in different situations are known. Emotion recognition has been an important interdisciplinary research topic in various fields, including psychology, neuroscience, and artificial intelligence. Many emotion classification studies use deep learning methods in combination with state-of-the-art statistics to optimize the accuracy of emotion detection, and attempt to integrate multiple modalities for better accuracy.

With increasing attention on emotion recognition, which will be detailed in the Related Work section, many emotional datasets have been collected, including both non-physiological signals (e.g., facial expressions and speech) and physiological signals (e.g., electroencephalogram (EEG), electromyogram (EMG), and electrooculogram (EOG)).

The two major categories that emotion datasets usually fall into are posed and spontaneous expressions [1]. Posed expressions are more intense and less ambiguous, where the test subjects receive instructions to act or perform an emotion. Spontaneous expressions contain more valuable information on natural expressions, but are more difficult to evaluate than posed expressions, and the results rely on the subject's self-report, which may introduce potential differences between the actual and reported emotions experienced [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

Deliberate behavior is often exaggerated and may fail to generalize to real-world behavior [3]. Posed emotions were widely used in 22 dynamic facial expression datasets [4]. Unlike these posed facial expression datasets, spontaneous expressions are more widely used in physiological emotion datasets, in which music or video clips are often used to elicit subjects' emotions. For example, some datasets collected subjects' self-reports on valence and arousal levels on a continuous scale, and these values can be used to categorize emotion [5], [11], [13]. However, there could be errors in the subjects' self-assessment reports, and the same emotions from different subjects could have different valence and arousal levels. Posed emotions are more systematically controlled and contain direct correspondence between the collected data and its associated emotions. Therefore, databases with deliberately posed emotions are typically more reliable for obtaining multimodal data and providing higher accuracy in emotion recognition [6].

Although there are many emotional datasets, multimodal emotional datasets, particularly those that include physiological data collected from posed emotions, are still deficient. In this study, we collected a new posed multimodal emotion dataset called PME4[1] to study emotion recognition from both non-physiological (audio and video) and physiological signals (EEG and EMG). The video consists of image sequences of actors producing facial expressions and audio speech while uttering a generic sentence, ''The sky is green.''. EEG signals reflect brain activity and EMG signals reflect facial muscle movements during these utterances. Each modality has its unique contribution to emotion recognition and might have various impacts at different emotion processing stages (i.e., pre-speech, during-speech, and post-speech). Hence, instead of aiming to improve existing emotion recognition methods, our goal is to provide a new posed multimodal emotional dataset to the research community with different feature extractors and machine learning models across the four modalities to classify emotional expressions.

The key contributions of this work include: 1) A new posed multimodal emotion dataset with four modalities (PME4): audio, video, EEG, and EMG; 2) A thorough comparison and analysis of a set of state-of-the-art data pre-processing and feature extraction techniques for each modality (including bootstrapping, principal component analysis, convolutional autoencoder, and/or mel frequency cepstral coefficients); 3) Comparisons of a few baseline machine learning methods (KNN, SVM, Random Forest, MLP, CNN, LSTM) to classify emotions with optimal features; and 4) A comprehensive survey of emotion recognition in terms of datasets, features, and recognition methods.

The remainder of this paper is structured as followed. First, the existing emotion datasets, feature extraction techniques, and emotion classification methods are described in Section 2. Next, we detail the data collection process for the PME4 dataset in Section 3, followed by a discussion of the feature extraction for each modality and classification method in Section 4. The results and analysis of emotion recognition for each modality are provided in Section 5. Finally, we provide the conclusions and discuss limitations in Section 6.

[1]The PME4 Dataset can be accessed and downloaded for research purposes at https://doi.org/10.6084/m9.figshare.18737924, and the code is available at https://github.com/jinchen1036/PME4-Emotion-Recognition

II. RELATED WORK

A. EMOTION DATASETS

With the increase in human-computer interactions, more emotional databases are being developed to better classify emotions, especially physiological signals. Some popular emotion databases are listed in Table 1. The table lists these datasets with several important aspects: the number of subjects, emotion states, elicitation, data types, feature extraction methods, and classifiers. Here, we mainly focus on the datasets; the last two aspects will be discussed more thoroughly later in this paper.

CK [7] and CK+ [8] are the most widely used facial expression datasets collected by Kanade *et al.* [7], [8]. CK+ [8] consisted of both frontal views and 30-degree views of 123 subjects' facial expressions from instructions to perform different expressions, including *anger*, *contempt*, *disgust*, *fear*, *joy*, *surprise*, and *sadness*. They used active appearance models to track the subjects' face shape across the image sequences, and then extracted the similarity-normalized shape (SPTS) and canonical appearance (CAPP) features and classified the action units (AUs) and emotions using the linear support vector machine (SVM). The linear SVM obtained 94.5% accuracy in AU detection and 83.33% accuracy in emotion detection when using both SPTS and CAPP, which is better than using individual features.

Brain electrical activity measured using electroencephalography (EEG) has recently become an interesting area for detecting internal emotional states [9], [10]. *Valence* and *arousal* are commonly used to characterize emotions. Both the DEAP [11] and DECAF [13] datasets collected multimodal physiological signals elicited by music videos and/or affective movie clips. Emotion state was determined by subjects' self-evaluated arousal and valence scores, where valence is associated with the level of happiness and arousal is associated with the level of calmness [11]. The SEED dataset [12] also used video clips to elicit emotions. However, unlike DEAP and DECAF, each video clip is associated with one of the three emotional states: *positive, negative, and neutral*. The participants in the SEED dataset were extroverts with stable moods based on the Eysenck Personality Questionnaire.

Similarly, the MPED [14] dataset also used video clips to elicit target emotion states. The 28 video clips and their corresponding target emotions were selected based on the participants' self-scoring on three psychological questionnaires on approximately 1500 video clips and evaluated using the k-means algorithm. All four datasets consisted of 32 to 62

TABLE 1. Summary of some emotion databases.

Dataset	Number of Subjects	Emotional States	Elicitation	Collected Data Types	Feature Extraction Methods	Classifiers
CK+	123	anger, contempt, disgust, fear, joy, surprise, sadness	Instruction to perform expression	Image sequences with FACS encoded	SPTS and CAPP	Linear SVM
SEED	15	positive, neutral, negative	15 emotion-specified movie clips with self-assessment	EEG, EOG, frontal face videos	STFT, differential entropy	DBN, SVM
DEAP	32	arousal, valence, liking	40 music videos with self-assessment	EEG, EMG, EOG, GSR, RSP, frontal face video	spectral power, Fisher's linear discriminant	gaussian naive bayes classifier
DECAF	30	arousal, valence	40 music videos and 36 movie clips with self-assessment	MEG, NIR facial videos, hEOG, ECG, tEMG	spectral power and DCT	Linear SVM
MPED	23	joy, funny, disgust, anger, fear, sad, neutrality	28 movie clips with self-assessment	EEG, ECG, GSR, RSP	PSD, STFT, HHS, Hjorth, HOC	SVM, KNN, LSTM, A-LSTM
PME4 (our)	11	angry, fear, disgust, sadness, happiness, surprise, natural	Instruction to perform expression	Audio, frontal face video, EMG, EEG	PCA, PSD, MFCC, autoencoder, Pre-trained CNN model	KNN, SVM, random forest, MLP, LSTM, CNN

A-LSTM = a novel attention-long short-term memory, CAPP = canonical appearance, DBN = deep belief network, DCT = discrete cosine transform, ECG = electrocardiogram, EEG = electroencephalogram, EMG = electromyogram, EOG = electrooculogram, FACS = facial action coding system, hEOG = horizontal electrooculogram, HHS = hilbert-huang spectrum, HOC = higher order crossings, GSR = galvanic skin response, KNN = k-nearest neighbor, LDA = linear discriminant analysis, LSTM = long short-term memory, MEG = magnetoencephalography, MFCC = mel frequency cepstral coefficient, PCA = principal component analysis, PSD = power spectral density, RSP = respiration, SPTS = similarity normalized shape, STFT = short-time fourier transform, SVM = support vector machine, tEMG = trapezius-electromyogram, PME4 = posed multimodal emotion with 4 modalities

EEG channels or 306 MEG channels along with other physical peripheral physiological signals, as shown in Table 1.

Multiple studies have attempted to determine emotional states based on EEG signals. Lan *et al.* [46] proposed the use of an autoencoder in combination with the K-mean cluster algorithm to automatically learn meaningful frequency features from the power spectral density of EEG signals. Zhang and Lu [12] applied a critical EEG channel selection method based on the weight distribution of a trained Deep Belief Network (DBN) model with differential entropy features from five different frequency bands of the EEG signals. This method achieved a similar accuracy (82.88% to 86.65%) with fewer EEG channels (range of 4 to 12) compared with 86.08% accuracy using all 62 EEG channels on the SEED dataset when classifying the three emotional states.

Lan *et al.* [47] used domain adaptation techniques on the SEED and DEAP datasets to reduce inter-subject variances between subjects and technical differences between datasets. The reported accuracies were 72.47% for SEED and 48.93% for DEAP using maximum independence domain adaptation (MIDA) with differential entropy features. Soroush *et al.* [48] proposed an angle space reconstruction method to obtain geometrical features from the EEG phase space. The reported classification accuracy of the four valence-arousal spaces was 91.37% using statistically significant features with nonlinear features extracted from the estimated differential angle and vector length from the angle space.

Time-frequency analysis is also widely used in EEG signal processing. In [49], the multivariate synchrosqueezing transform (MSST) method based on continuous wavelet transform was used to obtain features that stem from multichannel dependency in addition to mono-channel features. The joint instantaneous frequency and bandwidth estimate the multivariate bandwidth for all channels to partition the time-frequency domain. This method achieved an accuracy of 86.93% for classifying eight emotional states in DEAP. Abadi *et al.* [13] used the discrete cosine transform (DCT) feature to obtain the spatio-temporal patterns of DECAF's MEG data. They reported accuracies of 62% and 59% in determining arousal and valence levels using a linear SVM classifier.

Song *et al.* [14] proposed a novel attention-long short-term memory (A-LSTM) model to extract more discriminative features by capturing the information of interest from different sequences using residual connections. The model also uses a 1×1 convolution kernel to avoid the interaction of different channels. It achieved an accuracy of 76.06% in classifying seven emotions from the MPED dataset with higher-order crossing features.

Although there are many physiological emotion databases with multiple modalities, most of them used *spontaneous* expressions and relied on subjects' self-report arousal and valence levels. Many problems can arise when trying to match the collected data with corresponding emotions, such as inaccurate self-report values, differences between various subjects' report values for the same emotion, and multiple emotions being elicited simultaneously [52].

Unlike other datasets, the PME4 dataset collected *posed* emotions, where the subjects were asked to express their emotions. All subjects were either acting students or had acting experience, which helped to minimize variance in the data, as actors were trained to express the exact emotions based on the instruction. Most likely, the subjects also experienced these emotions through embodied cognition, thus providing more comprehensive matches between the collected data and their associated emotions. PME4 is a comprehensive dataset that consists of *synchronized* physiological signals and non-physiological signals and can be used to compare emotions expressed by subjects. This is in contrast to previous studies that typically measured physiological signals of viewers' EEG activity in response to different stimuli intended to elicit different emotions. As subjects were required to switch between emotions within a short time, their physiological

signals might not immediately reflect the instructed emotion compared to non-physiological signals. This may be especially true for EEG signals reflecting brain activity, where an emotional aftereffect was found in our previous study [41]. Moreover, each emotion period for all four modalities in our PME4 dataset can be separated into three stages: pre-speech, during-speech, and post-speech. Each stage could lead to different classification performances in the four modalities, especially the EEG signals during the speech stages. The research community could find new insights by analyzing EEG activity at different stages of various emotional states. In addition, PME4 also consists of data from five data collection time blocks for each subject with nearly evenly distributed sample sizes for each emotional state, allowing researchers to analyze how different time slots could impact subjects' emotions. Finally, integrated multimodalities can improve the inference of emotional states [15], [16].

B. FEATURE EXTRACTION TECHNIQUES

Extracting meaningful features from raw data is a critical step for emotion recognition, as classifiers cannot achieve optimal performance with noisy and/or uninformative data. Each modality contains different information; thus, we need to use different feature extraction methods.

Feature transformation techniques are used to reduce data dimensions by transforming the data into a feature space. *Principal component analysis (PCA)* uses orthogonal transformation to remove data redundancy by finding the projection matrix to map the original high dimensional feature space onto a low-dimensional component subspace. The first component contains the most significant variance among the original features than the second, and so on [19]. It has been applied to image and EEG pattern classifications [9], [44], [45].

Speech signals contain significant information that can be used to identify and understand the emotions of speakers; however, these signals often contain ''uninformative'' information, such as background noise and acoustic variability across speakers. Various feature extraction techniques are available for obtaining meaningful audio features by eliminating noise, such as the *mel frequency cepstral coefficient (MFCC), perceptual linear prediction coefficient (PLPC), linear predictive cepstral coefficient (LPCC), linear predictive coder analysis (LPC),* etc. [21].

MFCC is widely used in speech recognition systems because it uses a linear cepstrum to represent an audio signal that is close to the human auditory system [11], [22]. It extracts frequency-domain features, which perform better than time-domain features [23]. Extracting MFCCs includes the following key steps: noise removal with a hamming window, time-domain to frequency-domain conversion with FFT, Mel log power computation with a bank of filters, and MFCC computation with discrete cosine transformation [24], [25]. In addition to speech signals, MFCCs can be used to extract EMG signals from several facial muscles [26]. Studies that applied MFCC to EMG data for classification have suggested that a large time frame is needed to extract a better representation of EMG features [27].

Autoencoder is a well-known sophisticated feature extractor that contains two major parts: encoder and decoder. The encoder efficiently extracts meaningful features from the data and the decoder reconstructs the original data from the features extracted by the encoder. Multiple studies have used autoencoders to extract high-dimensional EEG information [9], [46], [50]. *Convolutional autoencoder* is also widely applied in obtaining salient feature vectors from image data [28], [29]. It uses the convolution layers to extract the significant features of the input while preserving the relationship between the pixels and extracted features. Convolutional networks outperform the capturing of valuable spatial correlation features of the image, and with the deeper network, they can capture deep features [30]. The autoencoder is trained directly in an end-to-end manner without applying regularization to ensure that no features are lost between the layers.

Pre-trained *CNN models* are usually better at retrieving meaningful generic features, particularly from images. The VGG neural network (VGG16, VGG19) [53] has been widely used in image classification and to extract image data features for emotion recognition [54], [55]. Even though the VGG neural network is pre-trained for object classification of various objects rather than human images, because ImageNet contains vast data samples, the convolution filters have been trained to extract the key features of the images of faces. It was used to extract image features for emotion classification.

C. EMOTION RECOGNITION METHODS

Many traditional classifiers have been used in emotion recognition, such as *Support Vector Machines (SVM)* [11], [13], [16], [31], [42], [44], [49], *K-Nearest Neighbors (KNN)* [12], [14], [31], [32], [42], *Random Forest (RF)* [49], [54], [57] and *Multi-layer Perceptron (MLP)* [10], [48], [54]. KNN is an intuitive and straightforward supervised method that uses a voting scheme to determine the sample class based on majority voting from K nearest training samples. A SVM utilizes a radial basis function kernel to improve the performance of high-dimensional data. Random Forest is an ensemble learning method consisting of multiple independent decision trees. MLP uses nonlinear classifiers and a backpropagation algorithm to update the network weights. For our experiment, these classifiers were used as the baseline for emotion classification.

Convolution Neural Networks (CNNs) are commonly applied in areas related to the analysis of visual images, such as object detection, image recognition and classification, and facial recognition. A CNN contains convolution layers that extract the input's significant features while preserving the relationship between the 2D spatial domain and extracted features, and has been used for emotion recognition [53], [55]. We can also convert temporal data into two dimensions of time-frequency data and then use CNN to find the relationship between the time domain and the frequency/spatial domain

TABLE 2. Number of trials per each subject emotion.

Subject \ Emotion	anger	disgust	fear	happy	sad	surprise	neutral
1	50	50	50	50	50	50	50
2	49	50	50	50	50	50	50
3	50	50	50	50	50	50	50
4	49	50	50	50	50	50	50
5	50	50	50	50	50	50	50
6	52	52	51	50	51	51	52
7	50	49	50	50	50	50	50
8	50	49	49	50	50	49	50
9	50	50	50	50	50	50	50
10	50	50	50	50	50	50	50
11	47	46	47	46	47	46	47

to determine the corresponding emotion. We used CNN as a baseline model with strong local spatial learning ability through its convolutional layers.

Long Short-Term Memory (LSTM) [33] is a special type of recurrent neural network with feedback connections that can process a sequence of data [34]. It overcomes the vanishing gradient problem of the traditional RNN [35] and has shown state-of-the-art performance on time-series data, including emotion recognition [14], [41], [56]. The RNN unit has neurons representing the temporal dependency of a data sequence and has a vanishing and exploding gradient problem with a long or unstable data sequence. LSTM solves this problem by integrating more memory gates to allow the network to learn long sequences of temporal data. The LSTM unit consists of a memory cell and three gates: the input, output, and forget gates. With the additional gates, these units can forget the previous states and update the current states as new information is provided. The input gate controls the effect of the input signal on the state of the memory cell. The output gate is responsible for the change in the hidden state based on the memory cell. The forget gate controls the effect of the previous hidden state. In Song *et al.* [14], LSTM outperformed other traditional classifiers in classifying emotions. We used LSTM as a baseline model to deal with all four modalities that are inherently time sequences.

III. DATA COLLECTION

According to psychologist Ekman [17], the six basic emotions are *anger*, *fear*, *disgust*, *sadness*, *happiness*, and *surprise*. Public emotion databases typically categorize five to eight emotions. This study focuses on recognizing the six basic human emotions in [17] plus a *neutral* emotion for a total of seven emotions. Data were collected from 11 human subjects (five female and six male individuals) who were students in acting, after informed consent was obtained. This study was approved by The Institutional Review Board of the City University of New York.

To enhance the accuracy of the collected posed emotions, all subjects had some acting experiences. Data collection took approximately four months, and each test *session* for each subject lasted for approximately two hours. The entire test

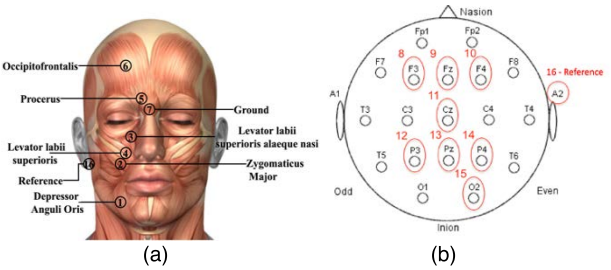


FIGURE 1. (a) EMG sensors' position on the face; (b) EEG sensors' position on the scalp.

session was divided into five *blocks*, with ten *trials* of each emotion presented in random order in each block. The subjects were allowed to take an optional break between blocks. We included a large number of repetitions of each emotion for each subject in the dataset to minimize the effects of variability and noise. Each trial was five seconds long, with one of the seven emotion labels presented on a monitor placed 57 cm in front of the subject. Subjects were required to utter the generic sentence "The sky is green" while mimicking the facial expression and experience indicated by the presented emotion label. The spoken sentence was chosen because of its neutral content, thereby minimizing interference with any emotion the subject was trying to experience and express. Each emotion label was displayed for 4 seconds, and a one-second break was provided between each emotion. Overall, the longest time for subjects to finish speaking the sentence was approximately 3 seconds.

Multiple issues can arise during data acquisition, such as electrodes becoming loose, interruptions from external sources, large head movements that prevent faces from being fully captured in the images, etc. After removing these error trials, 3829 trials remained across all four modalities, the details of which are shown in Table 2.

A. AUDIO AND VIDEO

During the test session, the subjects' facial emotional expressions were video recorded with a Logitech V-UCR45 USB webcam camera attached to a MacBook Pro 15" Retina Display Late 2013, and their voices were recorded using the laptop microphone. The laptop was placed in front of the subject to ensure adequate quality of the acquired video and audio. The audio signals were recorded at a 44.1kHz sampling rate and the video frames with a resolution of 960 $\times$ 720 pixels at 10 FPS.

B. EEG AND EMG

The EMG and EEG signals were acquired using gold-plated surface electrodes connected to Grass amplifiers. The EMG data were bandpass filtered online between 50 and 1000 Hz, whereas the EEG data were bandpass filtered online between.1 and 100 Hz. We used a 5kHz sampling rate and all electrode impedances were below 10 kΩ at the beginning of the experiment.

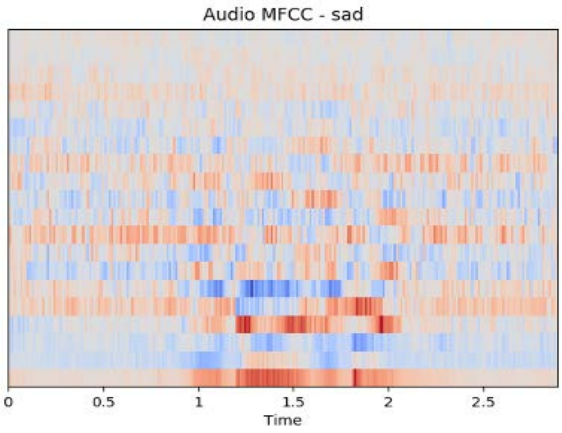


FIGURE 2. MFCC features of a sample trial with emotion sad with a 20ms window, where the horizontal axis is the time in seconds and the vertical axis is the 20 MFCCs.

The six muscles chosen for recording EMG activity were the depressor anguli oris, zygomaticus major, levator labii superioris alaeque nasi, levator labii superioris, procerus, and occipitofrontalis (Fig. 1 (a)), which are the major muscles involved in speech and associate facial actions units (AUs) during facial emotion recognition [18]. Note that the electrodes covered only half of the face; therefore, we could better use the video data for facial expression recognition. EEG data were collected through eight surface electrodes placed on the scalp: F3, Fz, F4, Cz, P3, Pz, P4, and O2 (Fig. 1 (b)). In total, we used 16 electrodes: six for EMG, eight for EEG, one ground channel that was placed on the nasion, and two references, placed on the left and right mastoids. All data were referenced online to the left mastoid and re-referenced offline to the average of the left and right mastoids.

IV. METHODS

The dataset contains both EEG and EMG signals together with the corresponding audio-video data of the 11 subjects. Although our sample size is small, it is on the same order as many neuroscience experiments and the results are still statistically significant for emotion recognition using such a dataset.

A. AUDIO

Non-speech interval signals were meaningless and contained noise that could affect the features used for emotion classification. To minimize noise, we focused only on the speech interval of audio data. We used a CNN-based audio segmentation method [51] to extract the speech intervals for each trial. After manual checking and fixing the extraction results, the speech intervals for subjects to speak the generic sentence "The sky is green" were different, ranging from 0.75 seconds to 3 seconds. Resampling speech segments with a uniform length causes multiple problems. For example, a high-frequency signal could alias a low-frequency signal, which would eventually provide invalid information when conducting feature extraction. Therefore, instead of resampling, we used a 3-second speech duration interval. To extract the 3 seconds *during-speech stage*, we started from the center of the speech interval that was automatically detected using the CNN-based audio segmentation method [51], and then evenly expanded 1.5 seconds before the center location and 1.5 seconds after the *during-speech stage*.

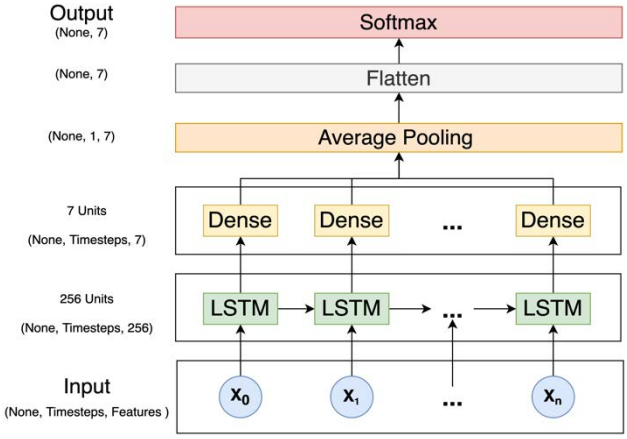


FIGURE 3. The structure of a single LSTM model for audio or image features. $X_0, X_1, \ldots, X_n$ are time interval data input to LSTM cells, for internal 0, 1, . . . , n; where n is the number of the timesteps.

To extract audio features, each 3-second during-speech audio sample was first normalized, and then a Hamming window was applied to remove the noise. Last, the 20 most significant mel-frequency cepstral coefficients (MFCCs) were extracted separately from each time *interval* of the Hamming window with 20 filter banks between 300 Hz and 3700 Hz, which are the parameters used to extract audio features by MFCC for emotion analysis in Dahake's work [36]. These extracted features formed a sequence vector that embedded both frequency (20 MFCCs) and time (number of Hamming window intervals within the 3-second data) information.

We tried two different Hamming window sizes, 20ms intervals with 10ms offsets and 100ms intervals with 50ms offsets, to compare the influence of the window size on the feature extraction performance. In total, for each trial of 3-second speech duration, we have 299 × 20 MFCC features for the 299 20ms-window and 59 × 20 MFCC features for the 59 100ms-window.

As MFCC features contain time information, we used LSTM for the analysis, because the LSTM architecture is optimal for time-series prediction. As the speech duration for each trial varied, some portions of the speech interval were likely to contain noise. Therefore, instead of obtaining the last output state of the LSTM, we connected the output state of each LSTM cell (over time) to a fully connected layer (with dense cells) and then averaged the output of the dense cells to obtain the final prediction through softmax, as illustrated in Fig. 3. We also applied an ensemble learning approach to the LSTM model by training 30 simple LSTM models. Each LSTM model had the same structure, as shown in Fig. 3. After

all 30 models were trained with the same training dataset, we took the average of the output from the softmax layer of each model. The average result is the final probability of each emotion class, and the emotion with the highest probability is the final prediction of the input data.

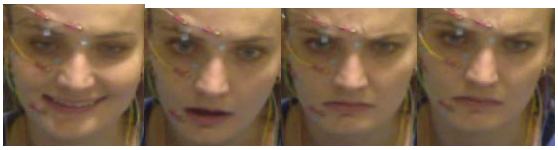


FIGURE 4. Cropped faces from the video sequence of subject 3.

To evaluate the effectiveness of the LSTM approach, we compared it with baseline models. These include the KNN with K equal to 10 and using the Euclidean distance formula to determine the nearest neighbor, Support Vector Machines (SVMs) with Gaussian kernels, Random Forest (RF) with 100 estimators and a maximum depth of 7 to reduce overfitting, and a multilayer perceptron (MLP) with 512 hidden nodes. The input trials to these baseline models are concatenated into one dimension with a size equal to the number of timesteps multiplied by the number of features.

To compare the performance of different classifiers, we used K-fold cross-validation, where K was equal to five in our current implementation. We randomly split each subject's emotion data samples into five subsets evenly. The split method ensures randomness within the training and testing datasets and maintains sufficient samples for each emotion and subject during the training process. Most importantly, this helps minimize information leaks and provides more accurate results for model performance. Each subset contained either 765 or 766 samples, and the classifiers were trained using four subsets and tested on the remaining subset.

B. IMAGE

As some subjects did not provide consent to release the original image data, we applied multiple feature extraction methods to obtain meaningful features from the original images to be released to the public and to train the machine learning models rather than using the original images.

Similar to audio signal processing, we focused on the image sequence during each trial's utterance interval (the during-speech stage) because it contains the most emotional expression. However, there were large variations in the lengths of the during-speech stages for the different trials, and the average during-speech interval was 1.3 seconds. To equate the speech interval for all trials, we extracted an image sequence of 16 screenshots per trial, evenly sampled from the central 1.5 seconds during-speech window at 10FPS.

Before extracting the image features, we cropped the face area on each frame as other regions did not contain any emotional information. We used the open-source MTCNN [37] and Python face recognition library [38] built with DLib [39] for face detection and extraction. As the existing face detection networks do not always provide 100% accuracy for detecting the correct face region, manual correction was also applied to fix any errors in the extracted images. Note that we used electrode paste and transparent tape to affix the surface electrodes to the face to collect EMG data (see Fig. 4). This minimized occlusion of facial expressions. The cropped face images varied in size and were converted to 224×224 pixels to input to the pre-trained networks.

We applied four different feature extraction techniques to the extracted facial images: PCA, a convolutional autoencoder, and two pre-trained networks (VGG16 and VGG19) [53].

Each image contains three color channels; however, the color does not have a significant influence on emotion recognition. Before applying PCA, we converted all images into grayscale and normalized the grayscale values to [0, 1]. All images in the training set were used to calculate the PCA transform matrix and were applied to the result matrix on both the training and testing sets to obtain their corresponding PCA components.

Convolutional networks should be more powerful than PCA for obtaining significant visual features for larger image sizes. The convolution autoencoder considers both the encoder (shown in Fig. 5) and decoder (a mirror of the encoder sharing the same parameters and replacing the convolutional layers with transposed convolutional layers). A feature vector of 2048 elements was obtained from the output of the encoder's final layer (dense layer) for each image.

Images were also passed into VGG16 and VGG19 with pre-trained weights on ImageNet to extract the features. We used the output from the last max pooling layer with size $7 \times 7 \times 512$ as the image feature because the remaining layers were originally used for classification. We also tried to use In-ceptionV3 [58] and ResNet50 [59], but the extraction feature size was too large for our classifiers, so the results are not reported.

To ensure a fair evaluation of the extracted features, we used the same 5-fold cross-validation technique as in the audio process. As LSTM is more powerful in dealing with temporal features, for each of the four extracted features, LSTM (Fig. 3) performance was compared with the baseline methods described in the audio section.

C. EEG

The EEG data were recorded from scalp electrodes and reflect activity from a large number of neuron potentials. Because the EEG signal also contains noise, extracting meaningful information from the EEG signals can be challenging.

Scalp EEG signals are typically unstable and noisy; therefore, we applied several noise reduction techniques when preprocessing the EEG data. First, the EEG data were converted into voltage values and then re-referenced to the right mastoid. Because neural activity measured with non-invasive EEG electrodes is more robust in the 0.1 to 30 Hz frequency band range, a bandpass Butterworth filter from 0.1 Hz to 30 Hz was applied to eliminate noise and less meaningful

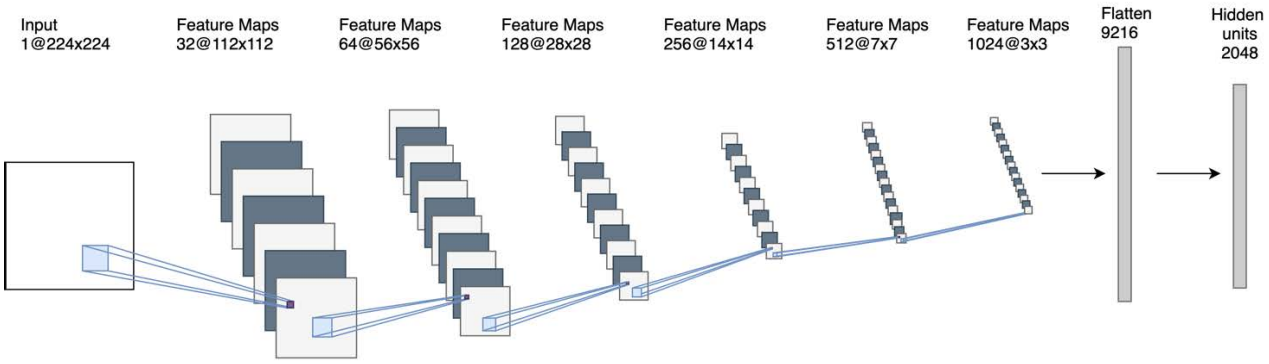


FIGURE 5. Encoder structure of the CNN autoencoder for cropped face image, each convolution layer with 3 × 3 convolutional kernel.

parts of the signals. After filtering, the data were downsampled from 5kHz to 1kHz for the sake of data dimension reduction in the later steps while maintaining a high fidelity of the neural signals.

EEG recordings often include various artifacts such as, blinks and facial movements. Because subjects in this study made facial expressions to convey different emotions, it was essential to minimize the influence of these extraneous, non-neural signals in the EEG data. Therefore, we applied an automatic artifact detection method that removes the impact of muscular activity in the EEG data. To remove the EMG effect on the EEG signals, we used the AAR plug-in for EEGLAB [40] and applied this removal process before band-pass filtering.

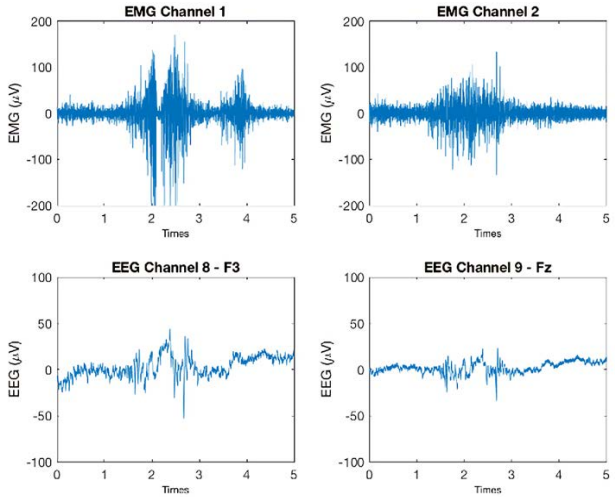


FIGURE 6. Filtered EMG and EEG data of subject 4's sample trial with the emotion anger.

To further minimize noise, we applied a *bootstrapping* method to extract EEG features, averaging over multiple trials of the same emotion for the same subject to obtain a more stable EEG signal. We processed the data of all 11 subjects using the bootstrapping method shown in Fig. 7. First, we extracted each emotional state from each subject, where the number of trials per emotion ranged from 46 to 51 (see Table 2). Second, we randomly split each subject's emotion trials evenly into training and testing subsets, with each subset consisting of 23 to 26 trials. After splitting the subsets for each subject's emotion, we randomly selected 20 trials from each subset and averaged these trials to obtain a new sample. The last step was repeated 400 times for the training set and 100 times for the testing set for each subject and emotion.

The bootstrapping method results in $400 \times 11 \times 7$ training samples (400 random bootstrapping sampling, 11 subjects and 7 emotions) and $100 \times 11 \times 7$ testing samples and overcomes the issue of limited trials available for training the model. Each of the eight EEG channels contained a sequence vector of 5×1000 elements (five seconds, 1000 sampling points per second at 1kHz) and was input into the 1D CNN model and LSTM model. Similar to image processing, we applied PCA feature extraction methods for each EEG channel and tested their performance with the baseline models. We obtained 50 PCA components for each EEG channel, accounting for over 97% of the energy spectrum. We also applied the 5-fold cross-validation technique to validate the performance of this method, where we repeated the bootstrapped process five times with different training and testing subsets for each subject emotion.

We also applied the autoencoder method stated in [46] and the K-mean cluster algorithm to extract ten features ($K = 10$) for each EEG channel. The input of the autoencoder was a raw periodogram from 0.1 to 30 Hz with a resolution of 0.2 Hz, resulting in a vector size of 155 for each channel. Two types of features result from this method. The first was the features directly extracted from the hidden layer of the autoencoder, which contained 100 features for each of the eight channels. The second was the ten features from the 10 cluster groups based on the similarity of the hidden layer weights for each channel, where each feature was the average value of each cluster group's hidden node values.

The LSTM models were used to evaluate the features extracted from these methods. Unlike the LSTM in Fig. 3, we used the last output state of the LSTM layer and then connected it to the softmax layer to obtain the emotion class

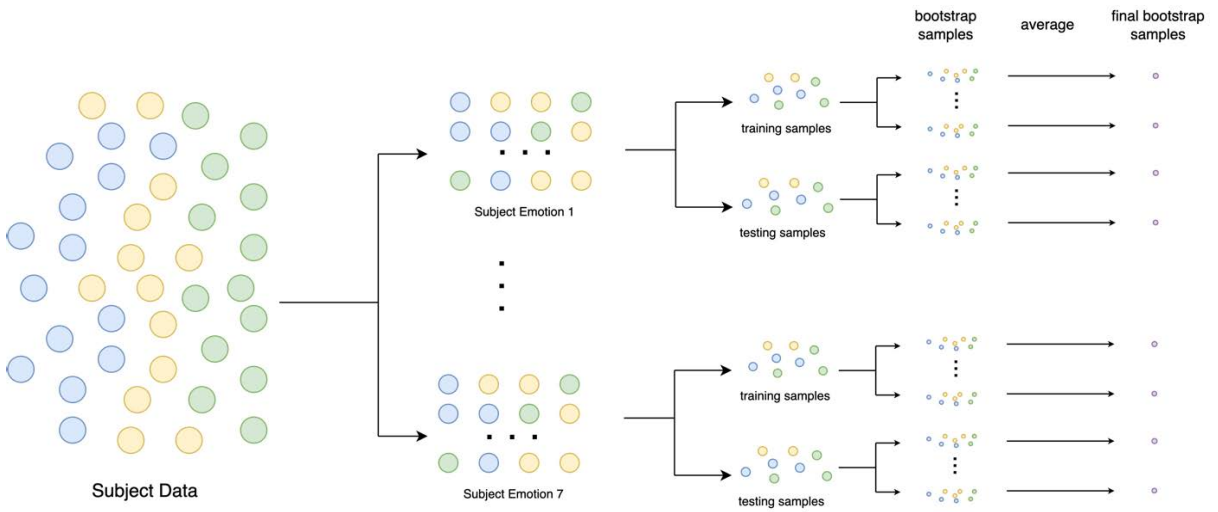


FIGURE 7. Bootstrapping process for EEG and EMG data of a single subject.

for the EEG and EMG data because we took the entire 5-second interval for the EEG (and EMG) data as one feature vector. The eight channels of the EEG data were also treated as eight timesteps for the LSTM model. In addition to LSTM, we applied the CNN model to both EMG and EEG features. The CNN model for the EEG PCA features is listed in Table 3. The architecture was similar for the other input sizes. It contains 1D convolution layer with a kernel size of 4 and stride of 2, an average pooling layer, and a dropout layer with a dropout rate of 0.2 after the second convolutional layer. The four baseline models discussed in the audio sections were also applied for comparison with the LSTM and CNN results.

D. EMG

Like EEG, EMG data are also noisy. Thus, we applied a Butterworth filter from 20Hz to 500Hz to EMG data to minimize noise within the signals. The EMG data were processed using the same bootstrapping methods as the EEG data.

We also used MFCC for EMG feature extraction because EMG signals are often used with audio data for analysis. Norali's work [27] suggested that longer frame sizes better represent EMG data. Based on [27], frame sizes of 2000ms, 4000ms, 5000ms and 10000ms result in higher accuracies. However, each of our trials is only 5-seconds long, so we used a 2000ms window size. We first applied a window size of 2000ms on a single trial to extract its MFCC feature with a 1000ms overlap. For the second method, a bootstrapping method (Fig. 7) similar to the EEG was used, with the same window size applied to the average 20 trials of the EMG for the MFCC features. For the third method, instead of averaging the 20 trials, we concatenated 20 trials to form a "super-trial" of a longer time that allows for a larger frame size and used the same window size to obtain the MFCC features.

We extracted 12 MFCCs for each EMG channel for each time step and concatenated all the channel coefficients into one feature vector. LSTM and CNN were used to classify the emotion using the EMG feature vectors with an architecture similar to that used for the EEG data. The input size for the LSTM and CNN was two dimensions: 12×6 coefficients and n time steps, based on the window size and method. We concatenated all coefficients together into a one-dimensional feature vector for the four baseline models.

TABLE 3. CNN model architecture for EEG PCA data.

Layers	Output Size	Parameters
Input	(None, 50, 8)	
Convolution	(None, 24, 16)	4 conv, stride 2
Convolution	(None, 11, 32)	4 conv, stride 2
Pooling	(None, 5, 32)	2 avg pool
Dropout	(None, 5, 32)	0.2 rate
Convolution	(None, 1, 64)	4 conv, stride 2
Flatten	(None, 64)	
SoftMax	(None, 7)	

V. RESULTS AND DISCUSSIONS

Table 5 summarizes the emotion recognition results for the four sensory modalities with various feature extraction methods and classifiers (four baseline models and three deep learning methods). We discuss the details of each modality below.

A. AUDIO

Table 5 shows that the best classification model for audio data with MFCC coefficients was the ensemble method with 30 LSTM models for both window sizes. For the 20ms window, we obtained an average accuracy of 71.32% and an average accuracy of 69.60% for the 100ms window. This performance is better than that of all the other baseline models, which have accuracies between 42.7% and 56.57%. To statistically validate that the LSTM model performs better than other baseline models, we compared the LSTM with the most accurate baseline model (SVM) using the t-test and obtained highly significant p-values of 1.24e-7 and 3.99e-7 and t-values of 17.34 and 14.92 for the 20ms and 100ms

TABLE 4. Subject's emotion recognition accuracies (mean ± std (%)) for audio MFCC features on 20ms window using ensemble LSTM network.

	ANGER	Disgust	Fear	Happy	Sad	Surprise	Neutral	All Emotions
Subject 1	57.4±19.1	66.1±9.9	50.9±15.0	43.1±18.9	59.4±18.1	66.4±17.7	67.5±7.0	58.2±6.5
Subject 2	57.1±14.1	53.0±16.7	48.9±18.1	70.0±11.6	56.7±14.3	77.1±13.7	45.9±9.1	57.8±3.1
Subject 3	72.7±28.1	89.3±10.8	85.0±7.1	83.6±12.7	80.6±13.5	91.3±8.1	90.7±8.3	84.0±2.6
Subject 4	39.3±9.1	34.0±8.3	23.9±4.6	35.1±9.1	49.0±11.0	49.4±17.9	78.0±9.8	43.4±3.2
Subject 5	69.6±10.5	85.3±8.8	82.5±11.9	78.8±9.5	77.8±22.3	65.0±22.7	80.2±8.1	77.1±7.3
Subject 6	93.8±5.3	39.6±15.1	48.9±9.4	96.5±4.4	91.2±8.5	52.6±11.0	72.3±16.7	70.6±4.9
Subject 7	74.1±18.2	63.0±13.2	58.1±26.7	72.9±11.8	69.8±23.3	71.0±18.1	82.1±11.6	68.5±4.3
Subject 8	83.5±16.0	79.8±8.5	91.0±9.2	74.1±10.2	65.4±4.0	66.0±7.2	68.9±15.5	75.1±3.7
Subject 9	82.9±18.4	88.5±11.8	70.1±18.0	88.1±7.6	54.6±16.4	84.2±4.4	88.7±11.5	79.0±5.3
Subject 10	95.3±5.8	82.0±7.2	87.9±8.0	81.4±12.7	88.7±7.6	76.1±7.1	94.5±7.8	86.1±4.8
Subject 11	84.6±14.7	72.3±12.0	68.9±15.7	90.4±8.4	71.8±12.8	76.7±15.6	85.2±16.2	78.2±2.4
All Subjects	72.8±2.7	68.1±5.9	63.8±4.2	73.1±1.5	70.1±4.2	70.2±2.3	77.5±3.6	71.3±1.6

windows, respectively, confirming that LSTM outperformed the baseline models. The difference between the 20ms window and 100ms window was not significant ($t = 1.7299$, $p = 0.1219$), suggesting that the difference in the window size of the audio data does not have much influence on the MFCC characteristics of the audio data. The details of the t-test results comparing the performance of the LSTM model with all the other models for all four modalities are shown in Table 6.

The model can be confused between emotions that result in similar tones. As shown in Fig. 8, emotion "fear" with the lowest classification accuracy was misclassified as disgust, surprise, and sadness approximately 8% to 10% of the time. One reason could be that some subjects have over-exaggerated voices that make the models unable to find clear boundaries to classify these trials into the correct emotion category. This can be further seen in Table 4, where the model has a huge difference in predicting each subject's emotions, ranging from 43.4% to 86.1%. Subject 4 had the worst accuracy, especially for the emotion "fear" that gets misclassified with other emotions except "anger" for over 10% of the time. Each subject expresses the same emotion in various tones and talking speeds, thus increasing the difficulty of training a general model to adopt all varieties.

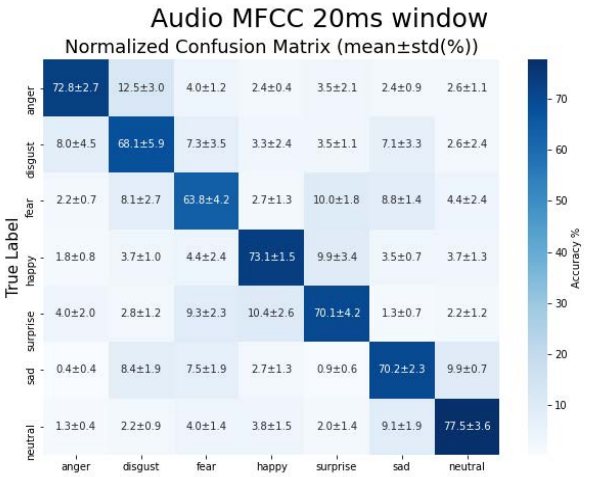


FIGURE 8. Confusion matrix of the ensemble LSTM network for audio MFCC features on 20ms window for individual emotion categories.

B. IMAGE

We have 224 × 224 pixel values for each image, but PCA significantly reduces the number of values while preserving data variability in reconstructing the image. For example, 50 PCA coefficients accounted for an average of 83.37% of the data variance with 0.08% standard deviations. PCA was mainly used for dimension reduction, but the CNN autoencoder seemed to be better at extracting meaningful spatial features as it outperformed PCA by approximately 9%, as shown in Table 5 in both MLP and LSTM. However, MLP and LSTM have similar performances with PCA or autoencoder features ($p = 0.58$ and $p = 0.21$). These two feature extraction methods might not preserve the temporal characteristics of the image sequence as trained with a small dataset; therefore, LSTM loses its advantage.

VGG19 features were similar to VGG16 features, with no statistical difference between the LSTM model results of these two features ($t = -0.93$; $p = 0.38$). The VGG16 feature with the LSTM model achieved a mean accuracy of 67.20%, which was over 10% better than the autoencoder features and 20% better than the PCA features. Features extracted from the pre-trained models (VGG16 and VGG19) are significantly different from those extracted from the PCA and autoencoder trained with our dataset (with p-values between 3.14e-6 and 8.70e-7). This is because the pre-trained models were trained with over 10k images that can learn richer image features. With more meaningful features, LSTM outperformed MLP by nearly 20%, as shown in Table 5.

Similar to audio, the model also has different performances in recognizing each subject's emotion, ranging from 46.3% to 89.4%. Subjects with lower accuracies either showed less-exaggerated facial expressions or ones that were over-exaggerated. This made it difficult to distinguish between emotions with similar facial expressions, such as fear and surprise, sad and neutral, with a misclassification rate of approximately 15%.

C. EEG

Without using the bootstrap method, all the models fail to recognize the emotion with EEG data. They all have accuracies near the random guess rate (14.3%) shown in Table 5, where the best accuracy is only 20% with SVM.

With the bootstrap method applied, there was an accuracy increase of 15% - 20%. Surprisingly, the baseline models either outperformed or had very similar accuracies to those of the deep learning models. The KNN model achieved an

TABLE 5. Emotion recognition accuracies of all four modalities.

	Feature Extraction Method	KNN	SVM	Random Forest	MLP	LSTM (1 model)	LSTM (30 models)	CNN
Audio	MFCC (20ms window size)	52.38% (2.27%)	53.76% (2.51%)	42.7% (1.58%)	48.93% (2.89%)	64.09% (2.20%)	**71.32% (1.57%)**	-
	MFCC (100ms window size)	52.17% (3.26%)	56.57% (1.95%)	45.30% (2.26%)	54.19% (2.48%)	64.68% (1.44%)	**69.60% (1.58%)**	-
Image	PCA (50 Components)	31.31% (1.22%)	45.60% (1.45%)	36.09% (1.16%)	**48.37% (1.35%)**	47.22% (2.62%)	-	-
	Autoencoder	35.00% (1.86%)	50.07% (2.28%)	47.06% (0.89%)	**57.95% (2.34%)**	56.25% (1.33%)	-	-
	VGG 16	50.25% (1.04%)	15.15% (1.51%)	55.24% (1.66%)	45.83% (15.57%)	**67.20% (1.71%)**	-	-
	VGG 19	51.11% (0.89%)	64.01% (1.81%)	54.53% (1.95%)	45.52% (11.53%)	**66.94% (1.12%)**	-	-
EEG	Single EEG Trial (50 PCA Components)	18.63% (1.07%)	**20.00% (1.14%)**	16.74% (1.01%)	17.97% (1.92%)	15.50% (1.16%)	-	15.82% (1.23%)
	Bootstrap Average 20 EEG Trials (50 PCA Components)	**39.70% (2.60%)**	37.50% (5.74%)	29.39% (2.48%)	30.85% (3.54%)	30.40% (2.76%)	-	21.63% (2.61%)
	Bootstrap Average 20 EEG Trials (Filter)	**36.99% (5.07%)**	35.29% (2.93%)	28.82% (2.54%)	23.65% (2.34%)	24.26% (4.73%)	-	30.83% (1.93%)
	Bootstrap Average 20 EEG Trials (Periodogram + Autoencoder)	**22.70% (1.75%)**	18.32% (2.37%)	18.40% (1.79%)	17.53% (1.50%)	18.78% (2.51%)	-	19.20% (3.83%)
	Bootstrap Average 20 EEG Trials (Periodogram + Autoencoder + Cluster)	**21.09% (3.32%)**	18.90% (2.05%)	20.01% (3.12%)	19.76% (1.74%)	18.92% (2.87%)	-	-
EMG	Single EMG Trial (MFCC with 2000ms window size)	**27.40% (1.73%)**	27.21% (1.04%)	24.97% (1.39%)	26.46% (1.33%)	24.03% (0.70%)	-	14.36% (0.76%)
	Bootstrap Concatenate 20 EMG Trials (MFCC with 2000ms window size)	36.08% (0.97%)	**36.40% (4.33%)**	22.66% (1.39%)	29.74% (2.41%)	30.58% (3.94%)	-	18.79% (3.25%)
	Bootstrap Average 20 EMG Trials (MFCC with 2000ms window size)	**37.09% (3.43%)**	23.42% (2.04%)	32.29% (2.79%)	22.49% (2.07%)	28.07% (1.97%)	-	19.14% (1.82%)

average accuracy of 39.70% with PCA features and was significantly better than that of the LSTM model ($p = 5.9e\text{-}4$).

Our data generated two types of features using the autoencoder method [46] (Section 4.3). The first was a feature vector of size 8×100 for each trial, which was the output from the hidden layer of the autoencoder with the raw periodogram of the EEG data as the input. The second was the average value of the ten cluster groups of the first feature type, where the clusters were based on the similarity of the hidden layer weights, resulting in a vector of size 8×10. Unfortunately, neither of the extracted features performed well in emotion classification, which can be attributed to several reasons. First, our dataset contains only eight EEG channels, which is less than the 32 channels used in [46]. Second, our dataset has a distinct data collection process, where the subject emotion is posed and required to change within a short period of time (5s), which may introduce more noise in the signals in comparison to the SEED dataset, which used movie clips to elicit emotions with long trial times (60s). Third, even when we reduced noise and expanded the number of training samples through bootstrapping, there is still insufficient data to train the autoencoder to capture the generic features of the EEG data.

PCA can reduce redundant information in EEG signals and preserve meaningful information compared to the autoencoder method. In addition, PCA is faster and requires less computational power than the autoencoder or raw data, as shown in Table 7. The PCA performance was very similar to the filtered data results and outperformed the autoencoder features. With 50 PCA components, it preserved an average of 97.9% data variability with a standard deviation of less than 0.1%. PCA features do not work well with the CNN model because the CNN model can extract features from the raw data in its convolution layers. In addition, because the input EEG features were not temporal data, LSTM did not work well compared to the baseline models.

D. EMG

Performance in classifying emotions based on EMG signals was similar to that of EEG signals (Table 5). As we used a window size of 2000ms, there were only four timesteps for the single and averaged trial data, which did not provide much temporal information for the LSTM. With concatenated trials, which had 99 timesteps, the LSTM model showed a slight improvement. The data used in Norali's paper [27] was 50 seconds in duration, but our data were only 5 seconds per trial, and even with concatenated trials, signals are not consecutive, which introduces errors in extracting the MFCCs. With this limitation of a small dataset and noise within the features, the deep learning models easily overfit the training data and do not perform better than the baseline models.

The bootstrap method also shows an improvement in denoising the EMG data, with an approximately 10% accuracy increase using KNN for both the concatenated and average methods. Moreover, the p-value of the KNN accuracy between the concatenate and average methods is 0.54, proving that these two methods have similar performance to the KNN model. This could be due to the similarity within the EMG signal characteristics of the same subject emotion. The average method smooths out the noise, and the concatenated method emphasizes the common characteristics.

E. DISCUSSION

We used general feature extraction techniques to obtain features from the four modalities; however, when we used more advanced and fine-tuned methods to extract task-specific features, the models better recognized the emotions, as shown in another of our studies [41]. Our dataset is relatively small; therefore, it is difficult to train the CNN autoencoder to obtain significant image features. Our previous work [41] used pre-trained models combined with an ROI net to extract features from the region of interest on the faces for emotion recognition. The accuracy of the trained LSTM model was increased by approximately 20%.

In addition to the methods discussed above, we also conducted a wavelet analysis on the EEG data, which should be more powerful than PCA and FFT in obtaining both time and frequency domain information. We used a Morlet wavelet at varying frequencies to extract the power at frequencies from 1 Hz to 40 Hz over a 5-second window. The wavelet feature vector results in a size of $8 \times 40 \times 5000$ (8 channels,

TABLE 6. t-test results of comparing LSTM model with others.

	Feature Extraction Method	KNN vs LSTM		SVM vs LSTM		RF vs LSTM		MLP vs LSTM		LSTM vs CNN	
		p-value	t-value	p-value	t-value	p-value	t-value	p-value	t-value	p-value	t-value
Audio	MFCC (20ms window size)	1.41e-7	17.07	8.47e-7	13.56	5.67e-10	34.39	7.88e-8	18.38	-	-
	MFCC (100ms window size)	3.31e-7	15.29	6.36e-6	10.39	1.87e-8	22.07	6.46e-7	14.02	-	-
Image	PCA (50 Components)	5.71e-6	10.54	2.39e-3	4.36	5.92e-5	7.66	5.82e-1	-0.57	-	-
	Autoencoder	4.05e-8	20.01	9.36e-4	5.09	2.30e-6	11.88	2.12e-1	-1.35	-	-
	VGG 16	3.63e-8	20.29	1.58e-11	53.79	2.46e-6	11.78	1.57e-2	3.05	-	-
	VGG 19	5.79e-8	19.12	6.38e-2	2.14	5.47e-6	10.59	3.99e-3	3.99	-	-
EEG	Single EEG Trial (50 PCA Components)	3.11e-3	-4.15	1.61e-4	-6.64	1.74e-1	-1.49	1.52e-1	-1.58	7.64e-1	-0.31
	Bootstrap Average 20 EEG Trials (50 PCA Components)	5.80e-4	-5.48	3.70e-2	-2.49	2.60e-4	6.17	8.20e-1	-0.22	8.50e-4	5.16
	Bootstrap Average 20 EEG Trials (Filter)	3.41e-3	-4.10	2.11e-3	-4.43	9.42e-2	-1.89	8.02e-1	0.25	2.50e-2	-2.87
	Bootstrap Average 20 EEG Trials (Periodogram + Autoencoder)	2.11e-2	-2.86	7.76e-1	0.29	7.91e-1	0.27	3.68e-1	0.95	8.41e-1	-0.20
	Bootstrap Average 20 EEG Trials (Periodogram + Autoencoder + Cluster)	3.00e-1	-1.10	9.89e-1	0.01	5.80e-1	-0.57	5.92e-1	-0.55	-	-
EMG	Single EMG Trial (MFCC with 2000ms window size)	3.76e-3	-4.03	4.55e-4	-5.69	2.12e-1	-1.35	6.83e-3	-3.61	6.65e-7	13.97
	Bootstrap Concatenate 20 EMG Trials (MFCC with 2000ms window size)	1.61e-2	-3.03	5.66e-2	-2.22	2.82e-3	4.24	6.92e-1	0.40	3.62e-4	5.89
	Bootstrap Average 20 EMG Trials (MFCC with 2000ms window size)	9.26e-4	-5.10	6.39e-3	3.66	2.45e-2	-2.76	2.39e-3	4.36	6.03e-4	5.45

40 frequencies, and 5000 sampling points per 5 seconds at 1kHz) per trial. With this large number of features per trial and limited number of trials, it is challenging to train the model to distinguish between different emotional states without overfitting. The emotion recognition accuracy for the wavelet features was 20.72% with the LSTM model. However, our preliminary studies suggest that there might be a delay in the brain to evoke an emotional state. As our data can be split into three different emotional processing stages, we will analyze the various stages of each trial in our future work. This study aims to provide different baseline methods along with a dataset for the research community to work on this interesting and challenging task.

Table 7 lists the computation time required to determine the emotion of a single trial using the best method for each modality. All experiments were simulated in Python on a MacBook Pro equipped with Intel Iris Plus Graphics 655, i7 CPU @ 2.7 GHz, 16 GB RAM, and 512 GB SSD. If we apply face tracking, we can reduce the time required for face detection in each image. Overall, our method does not require a large amount of computational power and has a large potential in many applications. For example, virtual assistant applications can use audio and image emotion detection to monitor user reactions, whereas portable EEG and EMG systems can be used for online classification of emotions based on neural and muscular signals. The application's A/B testing can incorporate these reaction data to achieve a better application design fit for user needs. EEG and EMG emotion detection could also help convey the emotions of people with various disabilities or disorders, such as cerebral palsy or vegetative states. Even though the emotion recognition accuracies for these two data types are not accurate for determining which of the seven emotions people express, we may be able to assess whether people have positive or negative reactions. One possible disadvantage of our method is that it required multiple trials for each of the emotions. However, given the variability within and between individuals, collecting responses on numerous occasions will enhance the correct interpretation of their emotions.

TABLE 7. Computation time (MS) of single trial using the best method of each modality.

	Audio (20ms MFCC)	Image (VGG16)	EEG (PCA)	EMG (2000ms MFCC)
Feature Extraction	595.8	2890.8	12.34	1428.6
Classification	178	34.3	5.87	5.29
Total	773.8	2925.1	18.21	1433.89

VI. CONCLUSION

This study provided baseline approaches for posed emotion recognition based on our new dataset, PME4, using four different modalities. We examined various feature extraction techniques (MFCC, PCA, autoencoder, and pre-trained CNN) and machine learning models (KNN, SVM, Random Forest, MLP, CNN, and LSTM) for each modality. We found that the LSTM deep learning model performs better than the traditional KNN in classifying emotions using audio and image sequences, as these data contain more abundant features of human emotion. As everyone expressed their emotions differently, we found that the general model had a large difference in determining each person's emotions, ranging from 43.4% to 86.1% accuracy. This problem can also be due to the fact that we collected posed expressions instead of spontaneous expressions, and some subjects might have overperformed or underperformed the required expressions. As there was more noise in the biosensor EEG and EMG data, bootstrapping improved data stability. With 20 bootstrapped samples, the accuracy increased by approximately 20% for the EEG and 10% for the EMG data. Our initial motivations for collecting only eight EEG channels were to focus the EEG channels on those that might be more influenced by the subject's emotional states and to minimize the difficulty and complexity of biosensor data collection. We achieved 39.70% accuracy with only eight channels of EEG data using the traditional KNN model, which has much less data than other emotional datasets and requires shorter computation time.

A. LIMITATIONS AND FUTURE WORK

There are a few limitations that we faced while performing the analyses, especially with biosensor data. Bootstrapping the EEG data increased the emotion recognition performance and focusing on the speech interval for biosensor data may also improve the performance. However, our PME4 dataset is relatively small, and because EEG and EMG data are inherently noisy, larger numbers of trials and subjects in future datasets may improve the classification based on these types of signals. Moreover, each person uses different talking speeds and tones to express the same emotion. Resampling the speech interval could introduce many problems but using the same speech duration for all subjects includes unnecessary data that affect the extracted features. Moreover, each experimental session was relatively long, which might have differentially affected the performance of each emotion at the beginning compared with the end of the experimental session. It will be interesting to analyze each experimental session as a function of time to assess whether the subjects' emotional states were expressed evenly over five testing blocks. We could apply the baseline model (LSTM for audio and images, KNN for EEG and EMG) to the four modalities in each block and compare the results to determine how time can influence the subjects' emotions. Subjects could also experience multiple emotions during the same trial, such as exhaustion towards the last block. Further analysis could be performed using the initial block as the baseline.

In the future, we plan to conduct multiple analyses. First, we plan to compare personal identification of each of the 11 subjects within each emotion or combination of both person and emotion recognition to explore the differences between people and their emotions in the four different modalities. Second, we will analyze the time course of the physiological signals (EEG and EMG) for different emotional states to determine whether there were differential delays in evoking emotion in the brain. The integration of multimodalities for emotion recognition is another line of future research. We will make our dataset available for research purposes and welcome other researchers to work on the dataset with new ideas and improved performance.

REFERENCES

[1] E. G. Krumhuber, A. Kappas, and A. S. R. Manstead, "Effects of dynamic aspects of facial expressions: A review," *Emotion Rev.*, vol. 5, no. 1, pp. 41–46, Jan. 2013.

[2] L. Nielsen and A. W. Kaszniak, "Conceptual, theoretical, and methodological issues in inferring subjective emotion experience," in *Handbook of Emotion Elicitation and Assessment*. New York, NY, USA: Oxford Univ. Press, 2007, pp. 361–375.

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[4] E. G. Krumhuber, L. Skora, D. Kuster, and L. Fou, "A review of dynamic datasets for facial expression research," *Emotion Rev.*, vol. 9, no. 3, pp. 280–292, 2017.

[5] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[6] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2005, pp. 474–477.

[7] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 46–53.

[8] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.

[9] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena, "EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation," *Sci. World J.*, vol. 2014, pp. 1–10, Sep. 2014.

[10] N. Kumar, K. Khaund, and S. M. Hazarika, "Bispectral analysis of EEG for emotion recognition," *Proc. Comput. Sci.*, vol. 84, pp. 31–35, May 2016.

[11] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, Oct./Mar. 2012.

[12] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[13] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 209–222, Jul. 2015.

[14] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "MPED: A multi-modal physiological emotion database for discrete emotion recognition," *IEEE Access*, vol. 7, pp. 12177–12191, 2019.

[15] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[16] W. Liu, W. L. Zheng, and B. L. Lu, "Emotion recognition using multimodal deep learning," in *Proc. Int. Conf. Neural Inf. Process.*, Oct. 2016, pp. 521–529.

[17] P. Ekman, "An argument for basic emotions," *Cognition Emotion*, vol. 6, nos. 3–4, pp. 169–200, May 1992.

[18] Z. Wang, S. Wang, and Q. Ji, "Capturing complex spatio-temporal relations among facial muscles for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3422–3429.

[19] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[20] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis—A brief tutorial," *Inst. Signal Inf. Process.*, vol. 18, no. 1998, pp. 1–8, 1998.

[21] D. Namrata, "Feature extraction methods LPC, PLP and MFCC in speech recognition," *Ijaret*, vol. 1, no. 6, pp. 1–4, 2013.

[22] A. Krishnan and M. Fernandez, "System and method for recognizing emotional state from a speech signal," U.S. Patent 8 595 005, Nov. 26, 2013.

[23] L. Xie and Z.-Q. Liu, "A comparative study of audio features for audio-to-visual conversion in Mpeg-4 compliant facial animation," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Aug. 2006, pp. 4359–4364.

[24] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," 2010, *arXiv:1003.4083*.

[25] J. Martinez, H. Perez, E. Escamilla, and M. M. Suzuki, "Speaker recognition using Mel frequency cepstral coefficients (MFCC) and vector quantization (VQ) techniques," in *Proc. CONIELECOMP, 22nd Int. Conf. Electr. Commun. Comput.*, Feb. 2012, pp. 248–251.

[26] H. Manabe and Z. Zhang, "Multi-stream HMM for EMG-based speech recognition," in *Proc. 26th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2004, pp. 4389–4392.

[27] A. N. Norali, "Human breathing classification using electromyography signal with features based on Mel-frequency cepstral coefficients," *Int. J. Integr. Eng.*, vol. 9, no. 4, pp. 85–92, 2017.

[28] N. Andrew, "Sparse autoencoder," *Lect. Notes*, vol. 72, pp. 1–19, Oct. 2011.

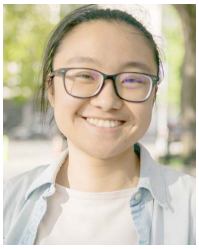
[29] X. Guo, X. Liu, E. Zhu, and J. Yin, "Deep clustering with convolutional autoencoders," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 373–382.

[30] L. A. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 262–270, 2015.

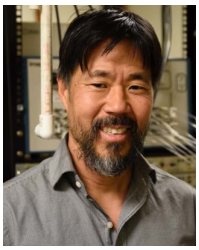
[31] R. M. Mehmood and H. J. Lee, "Emotion classification of EEG brain signal using SVM and KNN," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, Jun. 2015, pp. 1–5.

[32] M. Murugappan, "Electromyogram signal based human emotion classification using KNN and LDA," in *Proc. IEEE Int. Conf. Syst. Eng. Technol.*, Jun. 2011, pp. 106–110.

[33] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 473–479.
[34] K. Greff, R. K. Srivastava, J. Koutnìk, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.
[35] L. R. Medsker and L. Jain, "Recurrent neural networks," *Des. Appl.*, vol. 5, pp. 64–67, Dec. 2001.
[36] P. P. Dahake, K. Shaw, and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and support vector machine," in *Proc. Int. Conf. Autom. Control Dyn. Optim. Techn. (ICACDOT)*, Sep. 2016, pp. 1080–1084.
[37] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[38] P. Wagner. (2012). *Face Recognition With Python*. Tersedia Dalam. Accessed: Feb. 16, 2015. [Online]. Available: www.bytefish.de
[39] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jan. 2009.
[40] G. Gomez-Herrero, W. De Clercq, H. Anwar, O. Kara, K. Egiazarian, S. Van Huffel, and W. Van Paesschen, "Automatic removal of ocular artifacts in the EEG without an EOG reference channel," in *Proc. 7th Nordic Signal Process. Symp. (NORSIG)*, Jun. 2006, pp. 130–133.
[41] F. Abtahi, T. Ro, W. Li, and Z. Zhu, "Emotion analysis using audio/video, EMG and EEG: A dataset and comparison study," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 10–19.
[42] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 186–197, Mar. 2010.
[43] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalogr. Clin. Neurophysiol.*, vol. 29, no. 3, pp. 306–310, 1970.
[44] A. Subasi and M. Ismail Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8659–8666, Dec. 2010.
[45] F. Ye, Z. Shi, and Z. Shi, "A comparative study of PCA, LDA and kernel LDA for image classification," in *Proc. Int. Symp. Ubiquitous Virtual Reality*, Jul. 2009, pp. 51–54.
[46] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. Ḿuller-Putz, "Unsupervised feature learning for EEG-based emotion recognition," in *Proc. Int. Conf. Cyberworlds (CW)*, Sep. 2017, pp. 182–185.
[47] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, "Domain adaptation techniques for EEG-based emotion recognition: A comparative study on two public datasets," *IEEE Trans. Cogn. Devel. Syst.*, vol. 11, no. 1, pp. 85–94, Mar. 2019.
[48] M. Z. Soroush, K. Maghooli, S. K. Setarehdan, and A. M. Nasrabadi, "Emotion recognition through EEG phase space dynamics and dempster-shafer theory," *Med. Hypotheses*, vol. 127, pp. 34–45, Jun. 2019.
[49] P. Ozel, A. Akan, and B. Yilmaz, "Synchrosqueezing transform based feature extraction from EEG signals for emotional state prediction," *Biomed. Signal Process. Control*, vol. 52, pp. 152–161, Jul. 2019.
[50] B. Yan, Y. Wang, Y. Li, Y. Gong, L. Guan, and S. Yu, "An EEG signal classification method based on sparse auto-encoders and support vector machine," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Jul. 2016, pp. 1–6.
[51] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1755–1758.
[52] C. Harmon-Jones, B. Bastian, and E. Harmon-Jones, "The discrete emotions questionnaire: A new tool for measuring state self-reported emotions," *PLoS ONE*, vol. 11, no. 8, Aug. 2016, Art. no. e0159915.
[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[54] A. Rassadin, A. Gruzdev, and A. Savchenko, "Group-level emotion recognition using transfer learning from face identification," in *Proc. 19th ACM Int. Conf. Multimodal Interact. (ICMI)*, Nov. 2017, pp. 544–548.
[55] J. Sujanaa and S. Palanivel, "Real-time video based emotion recognition using convolutional neural network and transfer learning," *Indian J. Sci. Technol.*, vol. 13, no. 31, pp. 3222–3229, Aug. 2020.
[56] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion recognition based on EEG using LSTM recurrent neural network," *Emotion*, vol. 8, no. 10, pp. 355–358, 2017.
[57] D. Steyrl, R. Scherer, J. Faller, and G. R. Müller-Putz, "Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: A practical and convenient non-linear classifier," *Biomed. Eng./Biomed. Tech.*, vol. 61, no. 1, pp. 77–86, 2016.
[58] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.



JIN CHEN (Member, IEEE) received the B.S. degree in computer science and applied mathematics from The City College of New York, in 2020, where she is currently pursuing the M.S. degree in data science and engineering. She has been a member of the City College Visual Computing Laboratory, since 2017. Her research interests include computer vision, indoor localization and navigation, and assistive technology.



TONY RO received the B.A. degree in psychology from the University of California at Berkeley, in 1993, and the Ph.D. degree in neuroscience from the University of California at Davis, in 1998.

He completed the postdoctoral fellowship in cognitive neuroscience at University College London, in 1999. He was an Assistant Professor and then an Associate Professor with the Department of Psychology, Rice University. He was a Professor with the Department of Psychology, The City College of New York. He is currently a Presidential Professor in the programs in psychology and biology and the Director of the program in cognitive neuroscience with The Graduate Center, CUNY. He has published over 70 peer-reviewed research articles. He was a co-editor of a special issue of the journal *Cortex* and has published over ten review articles and book chapters. His research has been supported by NIH, NSF, and Philanthropy. His current research interests include the neural mechanisms of perception and attention. He is a member of the Society for Neuroscience, Association for Psychological Science, and Vision Sciences Society.

ZHIGANG ZHU (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees in computer science from Tsinghua University, Beijing, in 1988, 1991, and 1997, respectively.

He was an Assistant Professor, a Lecturer, and then an Associate Professor with Tsinghua University, Beijing, and a Senior Research Fellow/Visiting Professor with the University of Massachusetts Amherst. He is currently a Herbert G. Kayser Chair Professor of computer science with The City College of New York (CCNY) and The Graduate Center, CUNY, where he directs the City College Visual Computing Laboratory (CcvcL). His research has been supported by AFOSR, AFRL, ARO, DARPA, DHS, NSF, ODNI, and industry. His research interests include 3D computer vision, multimodal sensing, human–computer interaction, virtual/augmented reality, and various applications in assistive technology, robotics, surveillance, and transportation. He has published over 200 peer-reviewed technical papers in the related fields. He is a Senior Member of ACM. In May 2013, he received the President's Award for Excellence, The City College of New York, in the inaugural year of the President's Awards. His Ph.D. thesis "On Environment Modeling for Visual Navigation" was selected, in 1999, as a special award in the top 100 dissertations in China over three years, and a book based on his Ph.D. thesis was published by China Higher Education Press, in December 2001. He has been an Associate Editor of the *Machine Vision Applications* journal (Springer), since 2006. He was a Technical Editor of IEEE/ASME TRANSACTIONS ON MECHATRONICS (2010–2014).