Multimodal emotion recognition based on the fusion of vision, EEG, ECG, and EMG signals

Original Scientific Paper

Shripad Bhatlawande
Dept. of E&TC, VIT,
Pune, India
shripad.bhatlawande@vit.edu

Sourjadip Pramanik
Dept. of E&TC, VIT,
Pune, India
sourjadip.pramanik20@vit.edu

Swati Shilaskar
Dept. of E&TC, VIT,
Pune, India
swati.shilaskar@vit.edu

Swarali Sole
Dept. of E&TC, VIT,
Pune, India
swarali.sole20@vit.edu

Abstract *– This paper presents a novel approach for emotion recognition (ER) based on Electroencephalogram (EEG), Electromyogram (EMG), Electrocardiogram (ECG), and computer vision. The proposed system includes two different models for physiological signals and facial expressions deployed in a real-time embedded system. A custom dataset for EEG, ECG, EMG, and facial expression was collected from 10 participants using an Affective Video Response System. Time, frequency, and wavelet domain-specific features were extracted and optimized, based on their Visualizations from Exploratory Data Analysis (EDA) and Principal Component Analysis (PCA). Local Binary Patterns (LBP), Local Ternary Patterns (LTP), Histogram of Oriented Gradients (HOG), and Gabor descriptors were used for differentiating facial emotions. Classification models, namely decision tree, random forest, and optimized variants thereof, were trained using these features. The optimized Random Forest model achieved an accuracy of 84%, while the optimized Decision Tree achieved 76% for the physiological signal-based model. The facial emotion recognition (FER) model attained an accuracy of 84.6%, 74.3%, 67%, and 64.5% using K-Nearest Neighbors (KNN), Random Forest, Decision Tree, and XGBoost, respectively. Performance metrics, including Area Under Curve (AUC), F1 score, and Receiver Operating Characteristic Curve (ROC), were computed to evaluate the models. The outcome of both results, i.e., the fusion of bio-signals and facial emotion analysis, is given to a voting classifier to get the final emotion. A comprehensive report is generated using the Generative Pretrained Transformer (GPT) language model based on the resultant emotion, achieving an accuracy of 87.5%. The model was implemented and deployed on a Jetson Nano. The results show its relevance to ER. It has applications in enhancing prosthetic systems and other medical fields such as psychological therapy, rehabilitation, assisting individuals with neurological disorders, mental health monitoring, and biometric security.*

Keywords: *Emotion Recognition (ER), Analysis of Mental Health, Feature Fusion, Machine Learning (ML), Computer Vision, Physiological Signals*

1. INTRODUCTION

Emotion recognition (ER) is a fascinating field that aims to identify and understand human emotions through different modalities, such as facial expressions, speech, physiological signals, and behavioral patterns. It has become a research topic in areas such as medicine, machine learning (ML), and psychology [1]. ER technology can potentially be used in prosthetic arms to improve their usability and functionality, including adjusting their sensitivity and responsiveness for prosthetic arm wearers. Facial emotion recognition (FER) finds relevance in numerous applications, including identification processes for citizenship, identification cards, social security cards, and even intrusion detection [2]. Two prominent approaches in this domain are ER using physiological signals and FER using the ML model. Many FER systems employ ML techniques to recognize accurate emotions. One such ML-based FER system is proposed in [3], which constructs a multi-layer classifier based on a carefully curated dataset of 7 individuals, employing Haar-cascade features and histogram of oriented gradients (HOG) for feature extraction while employing Support Vector Machine (SVM) as the classifier.

Emotions can be recognized using different physiological signals, namely Electroencephalogram (EEG), skin temperature, Electrocardiogram (ECG), Electromyography (EMG), blood pressure, respiration rate, heart rate, Blood Volume Pressure, and Galvanic Skin Response, but the collection and processing of these signals become hard due to some added noise. For our system to work as intended, filtering these signals is necessary. Physiological signals, namely ECG, can reflect the relationship

between changes in emotions and heartbeat. Different emotions can be recognized by extracting heart rate variability (HRV). In the field of emotion identification, EEG signals are getting more attention day by day as they accurately reflect the feelings of any person. Signals collected from peripheral nervous systems, including ECG and EMG, can also be used for the same [4]. The studies in brain-computer interface (BCI) with different techniques that recognize emotions mostly made use of EEG, as EEG responds in time and is sensitive to changes in affective states [5]. EMG-based systems for ER have the capability to identify a person's genuine emotions.

ER systems can prove useful for understanding someone's emotional state and for achieving better communication. Harnessing the power of computer vision and ML advancements, researchers and developers are now utilizing facial expression analysis to automate ER processes with remarkable precision. Multimodal emotional datasets can be used, or the fusion of two or more signals improves accuracy and provides the relationship between different bio-signals. These signals can be fused together for the unique identification of emotions. Fusion of EEG, ECG, and EMG signals can be done to achieve better performance. Among the diverse approaches in this domain, FER holds particular significance.

In this paper, a multi-model system is proposed that considers the features from both the fusion of EMG, ECG, and EEG and facial expressions. The subsequent sections delve into the process of implementing it, aiming to contribute to the development of a more robust and reliable system.

2. RELATED WORK

Researchers have done notable research in the domain of ER using bio-medical devices due to their potential applications in various areas such as mental health and human-computer interaction. Hwang et al. made use of the SEED dataset, a publicly available dataset, recorded from 15 participants. The sample rate for the EEG signal was chosen as 1000 Hz [6]. Ferdinando Hany et al. made use of the ECG signals, which were taken from the Mahnob-HCI database, containing records of 27 participants, while participants were shown images and videos [7]. A 32-channel EEG was recorded, and a sampling frequency of 256 Hz was used in [8]. A diagram and positioning of electrodes over the face for EMG and experimental setup have been proposed in [9], which helps in getting insight into montage placement. According to research, EEG signals can be acquired using silver chloride electrodes. The EEG signal has a relatively low amplitude (5–500 uV), making it difficult to capture and evaluate. As a result, an amplifier was used to amplify the signals to a desirable level, which helped in achieving better accuracy [10].

Kumar Nitin et al. collected data from participants using music and videos as external stimuli, with almost 40 trials for each participant. 32 electrodes were placed according to a 10–20 worldwide system for EEG with a sampling frequency of 512 Hz. Down-sampling was initially carried out from 512 to 128 Hz. The Noise was removed using a bandpass filter in the frequency range of 5–45 Hz. A Butterworth filter was used for the filtering signal [11]. Signals, which were collected using 16 channels PowerLab with a sampling rate of 400 Hz, were filtered by a digital notch filter at a frequency of 50 Hz [12]. HoSeung Cha et al. used a Riemann manifold for feature extraction, and a pattern recognition-based myoelectric interface was built based on Linear Discriminant Analysis (LDA) implementation. Recall, F1 score, and precision were calculated for concluding LDA adaptation conditions. The various results were successfully reflected in the user's current state between virtual and real [13]. Time domain characteristics such as First and Second differences, Root mean square, Line length, Signal power, and Total Wavelet Energy, Frequency domain characteristics such as Dominant frequency, Total Wavelet Energy, and entropy-based characteristics such as spectral entropy, Shannon entropy, and sample entropy were employed, as well as various classifier techniques including SVM, ANN, and Naïve Bayes were used for classification in [14]. Silvio Barra performed feature extraction using a* peak detection method. The ECG signals were characterized by detecting repeating peaks consisting of Q, R, and S waves [15]. An analysis was performed by the authors using Normalization, mean, and standard deviation of the original signal for all 11 channels [16].

Many researchers have utilized a variety of validation and searching techniques to optimize the classification process. Different classification techniques were used and reviewed in [17]. In [18], the Bio Vid Emo DB dataset was used to validate the proposed method. The experimental conditions included a classifier, namely SVM using the RBF kernel, which obtained a 79.51% maximum accuracy in differentiating positive and negative emotions. P. Sarkar et al. fed emotion identification weights into a neural network with fully connected layers that were trained to categorize emotions. Databases like Dreamer, Seed, and Swell were used to test the results, with the maximum being an improvement over the 96.9% accuracy [19]. Min Chen et al. [20] used a minimal quantity of multimodal labeled data; the proposed LLEC first trains the neural network model. The unlabeled data is then automatically labeled and added to the training set, utilizing improved hybrid label-less learning to boost model detection accuracy even more. Wu et al. [21] developed a prototype of a wearable emotion-detection headband using EEG. The temporal window of 0.5 to 4 sec is suggested, with a short delay of <1 sec in between the 5 bands of the signal. The prototype included an EEG- measuring front end in the form of a headband for acquiring and pre-amplifying EEG data. Ante Topic et al. [22] built a model showing that the holographic feature map technique clearly outperforms topographic feature maps.

A successful expression recognition system has the potential to have robust features to effectively recognize the face's appearance [23]. Alghamdi et al. gave a comparison between various technologies that can be used for facial recognition. Types of detection algorithms like SVM, the Viola-Jones (VJ) algorithm, the Kanade Lucas Tomasi algorithm, the AdaBoost algorithm, the hybrid face detection algorithm, and the Elman Neural Network have been introduced [24]. Li et al. discussed the future development direction and potential application prospects of FER. Techniques like Principal Component Analysis (PCA) and LDA have been discussed. For classification, SVM, Ada-boost, small samples, and neural network techniques were compared. Some deep learning-based techniques were also reviewed [25]. He et al. proposed Laplacianfaces, which are based on the Laplacian Eigenmaps approach. Laplacianfaces is an algorithm utilizing the Laplacian Eigenmaps approach for face recognition by constructing a graph using the similarity of the face images, which is then used to compute the eigenfaces [26]. The paper [27] focuses on conducting a comparative analysis of various convolutional neural network (CNN) architectures for face recognition. The authors emphasize the significance of face recognition in numerous applications and highlight the growing use of CNN in this field. It provides an overview of popular CNN architectures, including AlexNet, VGGNet, GoogLeNet, and ResNet, explaining their design principles and features relevant to face recognition. The authors conducted experiments by training and evaluating the selected CNN architectures on the dataset, employing evaluation metrics like accuracy, precision, recall, and F1 score to assess performance.

Chowdary et al. discussed various deep-learning approaches used for FER. This includes CNN, Recurrent Neural Networks (RNN), or a combination of both. The authors described the architecture and configuration of the neural network used in their study. The authors also explained preprocessing techniques applied to the images, such as normalization or data augmentation, to enhance the performance of the model. The training process of the deep learning model, including details on the optimization algorithm, loss function, and hyperparameter tuning has been presented in [28]. In [29], two databases named Bosphorus Face Database (consisting of 4666 face images of 105 subjects) and the University of Milano Bicocca (UMB) Face Database (consisting of 1473 face images of 143 subjects) were used. Multiple ML models were tested on them. Enhancement attacks like blurring, sharpening, histogram equalization, and median filtering; geometric attacks like rotation and cropping resizing; and noise attacks like Gaussian attacks, speckle attacks, and poison attacks were implemented. Koonsanit et al. created a framework for ER by capturing the facial images of users. Facial expressions were categorized into seven categories, i.e., Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Classifiers like SVM, Logistics Regression (LR), K-nearest neighbor (KNN), and Multi-Layer Perceptron (MLP) were used [30]. Liu et al. collected a custom dataset for EEG signals from 16 participants. Higuchi Fractal Dimension (FD) Spectrum was used for analyzing non-linear properties of EEG signal [31]. Ergin et al. used the Empirical mode decomposition (EMD) method for EEG signals, to obtain Intrinsic Mode Functions (IMFs). Previously used methods have been described in Table 1.

Numerous solutions have been put forth for ER through the utilization of physiological signals. However, these methods have limitations, which call for further research to improve their effectiveness. Through using a multi-modular system, a more robust and reliable system can be obtained. The contemporary ER system includes the use of music, games, and videos as external stimuli, with the assembly including a complex module for analysis where the limitations include poor performance in participant-independent ER. This results in reduced robustness and increases the chance of misdiagnosis or biased decisions. Also, the use of deep learning techniques results in complex models requiring more resources to deploy in real-time.

Table 1. Review of previous technologies used for ER

Ref No.	Year	Signal	Dataset	Algorithm	Performance
[6]	2020	EEG	SEED	Input signal -> Band-pass filtering -> Short-Time Fourier Transform (STFT) -> Applying single-task DNN, multi-task DNN and adversarial DNN -> Model performance evaluation	Accuracy- 75.31%
[9]	2014	EMG, ECG, GSR	Custom	Data acquisition of EMG, EEG and GSR signals using audio-visual stimuli -> Filtering using notch filter and average filter -> Applying Higher order statistics (HOS) for feature extraction -> Classification using KNN	Accuracy- 69%
[31]	2014	EEG	Custom, DEAP	Data acquisition for EEG signal from 16 subjects -> Higuchi Fractal Dimension Spectrum -> Classification using SVM -> Validation on DEAP dataset	Accuracy- 85.38%
[32]	2019	EEG	Custom	Data acquisition for EEG signal from 25 subjects -> Filtering of data -> Empirical Mode Decomposition -> Selection of intrinsic mode functions -> Classification using SVM	Accuracy- 84.3%

Thus, a need for a lightweight system using ML arises. Even though the result of emotional identification is significant, further study is required to uncover the elements that may simply and efficiently recognize emotional patterns. The importance of a model that considers the various physiological signals together and makes decisions accordingly is taken into consideration. As a result, to increase the accuracy and other performance parameters of the final result, a fusion of EEG, ECG, EMG, and facial expressions was undertaken. An Affective Response System (ARS), which is based on stimuli from sudden news, audio, video, speech, and sudden actions, was used to provide better results. The physiological system proposes a better variable for choosing the Butterworth filter, window slicing techniques, Exploratory Data Analysis (EDA), and PCA. The model based on facial expressions leverages several feature extraction techniques to provide a comprehensive analysis of emotional states. Decision Tree, Random Forest, and Optimized Random Forest were selected as ML classifiers. The model was tested for various performance parameters using the f1 score, precision, accuracy, sensitivity, and Receiver Operating Characteristic Curve (ROC) curve. By incorporating additional modalities, the proposed system potentially enhances the accuracy and reliability of ER systems, leading to more robust and reliable results.

3. METHODOLOGY

This paper presents an approach for ER by fusing physiological signals, including EEG, ECG, EMG, and facial expressions. The proposed system is organized into two distinct sections, with one dedicated to physiological signals and the other to facial expressions. In the physiological domain, the system integrates signals from the EMG, EEG, and ECG, extracting relevant features. Simultaneously, facial emotions were captured by analyzing facial cues from images. Emotions can be broadly categorized as positive, namely surprise and happiness, and negative, namely sadness, anger, and fear. This paper focuses on recognizing three emotions, fear, neutrality, and surprise. The data utilized for this study was acquired from a sample of 10 healthy participants. The collected data was subjected to preprocessing to remove any artifacts, followed by feature extraction. These features were carefully fused and optimized to form a comprehensive feature fusion vector, which would enable a more robust ER model.

In the context of FER, a custom dataset was curated for the experiment. A model was trained using this dataset to classify the above-mentioned emotions. The final decision-making process involved a voting mechanism between the predicted facial emotion and the emotion inferred from physiological signals. The final result is the generation of a comprehensive medical report, facilitated by GPT-3, leveraging the combined insights from both physiological and facial expression data sources. This approach takes advantage of the strengths of both modalities, leading to more accurate and reliable ER outcomes. An overview of the proposed system can be seen in Fig. 1.

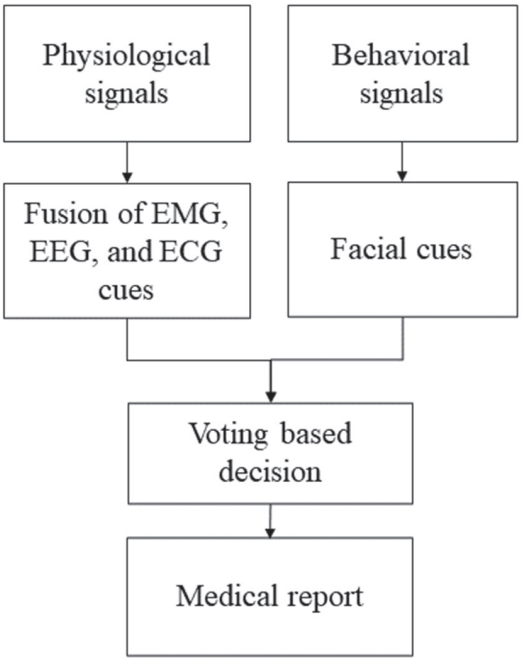


Fig. 1. Overview of the proposed multimodal ER system

3.1. IMPORTANCE OF ECG, EEG, EMG, AND FACIAL SIGNALS

There is a need to understand why vision, EEG, ECG, and EMG were selected for this experiment before delving deep into the process of multimodal ER. Fusion of vision, EEG, ECG, and EMG signals is an interdisciplinary approach that combines physiological and behavioral data to gain a comprehensive understanding of human emotions. Together, these signals provide valuable insights into how fear, surprise, and neutral emotions are represented and expressed.

3.1.1. VISION

Visual information is one of the most prominent and informative source for understanding emotions. Facial expressions, body language, and eye movements are crucial indicators of emotional states. Fear and surprise caused widened eyes, raised eyebrows, and open mouths. Neutral emotions typically result in neutral facial expressions. In neutral, there wasn't any noticeable change in the facial expressions. The analysis of these visual cues provides insights into the intensity and type of emotion being experienced.

3.1.2. EEG

EEG measures electrical activity in the brain and is particularly useful for studying cognitive and emotional processes. Different brain regions and frequency bands are associated with various emotions and cogni-

tive functions. Fear and surprise lead to increased activity in the amygdala and other brain regions associated with emotional processing, while neutral emotions have a more balanced brain activity pattern. EEG can capture these differences in neural activation during different emotional states.

3.1.3. ECG

The heart rate and heart rate variability are closely linked to emotional responses. ECG measures the electrical activity of the heart and reflects changes in autonomic nervous system activity during emotions. Fear and surprise typically lead to an increase in heart rate and reduced heart rate variability due to activation of the sympathetic nervous system ("fight or flight" response). In contrast, neutral emotions often result in relatively stable heart rate patterns. ECG can quantify these physiological changes.

3.1.4. EMG

EMG measures electrical activity in skeletal muscles, which provides information about facial expressions and emotional responses involving muscle contractions. Fear and surprise lead to increased muscle tension in the facial muscles, such as in the brows and around the mouth. Neutral emotions result in a more relaxed facial muscle state. EMG data can capture these subtle muscle activity changes.

3.2. MODEL BASED ON THE FUSION OF PHYSIOLOGICAL SIGNALS

The model architecture was divided into six main parts: data acquisition, pre-processing, feature extraction, feature selection, feature fusion, and classification. Each part plays a crucial role in the overall process of physiological ER.

3.2.1. DATA ACQUISITION

A well-designed setup was implemented to collect the physiological signals of the participants accurately. The setup for the physiological system differs depending on the type of biomedical devices used to detect the physiological signals, including EEG, EMG, and ECG.

A custom dataset for all three ECG, EMG, and EEG signals for ER was collected. The emotions that were selected for the proposed model were fear, neutrality, and surprise. The channels were selected considering the effect of emotion on the lobes of the brain, the electrical activity of muscles, and the heart. In the case of 4-channel EMG, the frontalis and masseter muscles were used, which are described in Fig. 2(a) with the ground placed at the forehead. The ECG setup included a 2-channel electrode placed on the arm, as shown in Fig. 2(b). AgCl electrodes were placed at positions namely F3-F4, T3-T4, P3-P4, and O1-O2 for a 4-channel EEG, which can be seen in Fig. 2(c). Setup for each of the devices was carried out separately.

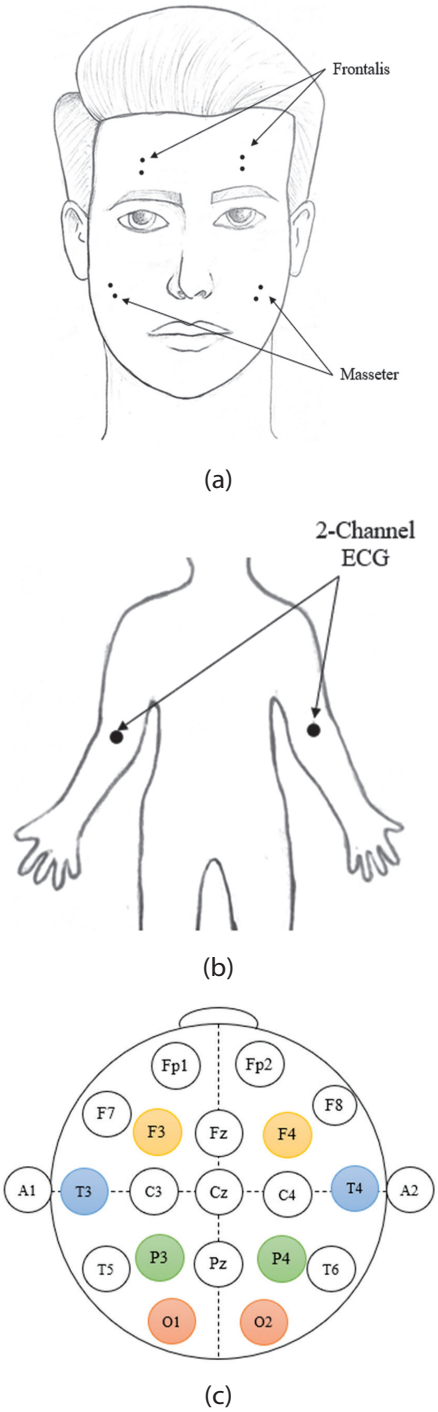


Fig. 2. Montages for (a) EMG, (b) ECG, and (c) EEG used in the System

The data acquisition setup included subject preparation and environmental setup. The chosen subjects were between the ages of 19 and 21, were healthy, and had no prior medical history. They were prepared in a manner to minimize any human artifacts. By ethical guidelines, all participants willingly signed consent forms before their involvement in the experiment.

Participants were given experimental guidance at the start of the experiment and were asked to sit in a laboratory environment. An affective response system (ARS) was used for detecting and analyzing emotion in response to audio and video stimuli. Simultaneously, a video of each

participant was recorded, which was used for testing the facial model based on the physiological emotions. Different video clips corresponding to different emotions were shown to 50 people. Based on their voting, video clips with more than 90% votes were finalized for the experiment. For each of the emotions, 3 clips were chosen, bringing the total to 9 clips with a length of 2 minutes. The clips chosen were carefully selected to induce the desired emotion. The data were recorded from 10 participants. While recording the data, each participant was asked to relax for one minute at the start of the experiment, then a signal was collected for two minutes while a video clip was shown to the participant. In this period of 2 minutes, the signal was recorded for 1 minute and 30 seconds, depending on the time for which emotion was induced. The subject was provided to rest in between each session for one minute, as described in Fig. 3. This process was repeated for each participant. For each emotion, every participant underwent three separate test sessions for signal acquisition, following an identical procedure that was repeated for ECG, EMG, and EEG measurements. Data were collected separately for the different signals. After collecting the signal readings, the data was exported as Excel sheets and formatted accordingly.

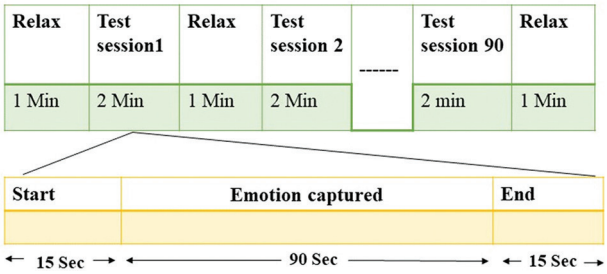


Fig. 3. Experimental setup for physiological signal acquisition of EEG, ECG, and EMG

3.2.2. DATA PRE-PROCESSING

In the process of bioelectric data collection, noise may be added due to the participant's body movements, eye blinking, or other sources. Notch and Butterworth filters were used to remove these artifacts. A default-frequency notch filter was maintained during the collection of data to remove artifacts from the signals. A notch filter was used to eliminate 50Hz supply frequency noise. The range of this band-stop filter was 49–51 Hz. As shown in Fig. 3, a window of 90 seconds was selected to capture the signals generated by the subject when subjected to a particular emotion. The window size was based on noticing the frequency and changes in the signal during the emotion. It was chosen as 5 seconds for EEG, EMG, and ECG, respectively. The culmination of all the windows was added to the final Excel file, along with the other necessary steps such as imputing missing values and clearing out outliers. All the collected data was formatted according to their specific requirements, making it organized for feature extraction. The process flow for signal acquisition, preprocessing, feature extraction, and classification is described in Fig. 4, with the end result being an accurate classifier that recognizes emotions.

3.2.3. FEATURE EXTRACTION

Feature extraction extracts the different properties of signals. Different feature extraction techniques, domains, and methods exist for bio-signal processing. For this experiment, time, frequency, and wavelet domain features were considered, as shown in Table 2.

3.2.3.1. TIME DOMAIN FEATURES

Signal properties that pertain to the temporal dimension are designated as time-domain features. These features are helpful in analyzing the temporal characteristics of signals. Root Mean Square, Integral, Variance, Mean Absolute Value, Amplitude Change, Difference absolute standard deviation value, Average Wavelength, Willison amplitude, Zero-Crossing, and Myopulse percentage rate were all taken into account.

3.2.3.2. FREQUENCY DOMAIN FEATURES

In this domain, the signals are analyzed with respect to frequency rather than time. They provide insights into the spectral content of a signal. For the frequency domain, features, namely Mean Power, Mean Frequency, Frequency ratio, and total power were selected.

3.2.3.3. WAVELET DOMAIN FEATURES

The analysis incorporates techniques that examine the signal simultaneously in both time and frequency domains, offering a comprehensive and nuanced understanding of the signal characteristics. Wavelet decomposition up to four levels was utilized to extract features from the signal. The chosen wavelet features include Mean, Standard Deviation, Energy, and Entropy of the coefficients.

In addition, Kurtosis, Max-Min, and H2-H1 were computed for each bio-electrical signal. The combination of these features with all the domain features resulted in a total of 35 features. These features were extracted from each channel of ECG, EMG, and EEG, whose details can be found in Table 2.

Table 2. Selection of features according to different domains on a single channel for EEG, EMG, and ECG

Domain	Features Extracted
Time domain	Variance, Difference absolute standard deviation value, Willison amplitude, Zero Crossing and Myopulse percentage rate, Root Mean Square, Integral, Mean Absolute Value, Wavelength, Average Amplitude Change
Frequency domain	Mean Power, Mean Frequency, Frequency ratio, Total Power
Wavelet domain	Mean, Kurtosis, Standard Deviation, Energy and Entropy of the coefficients

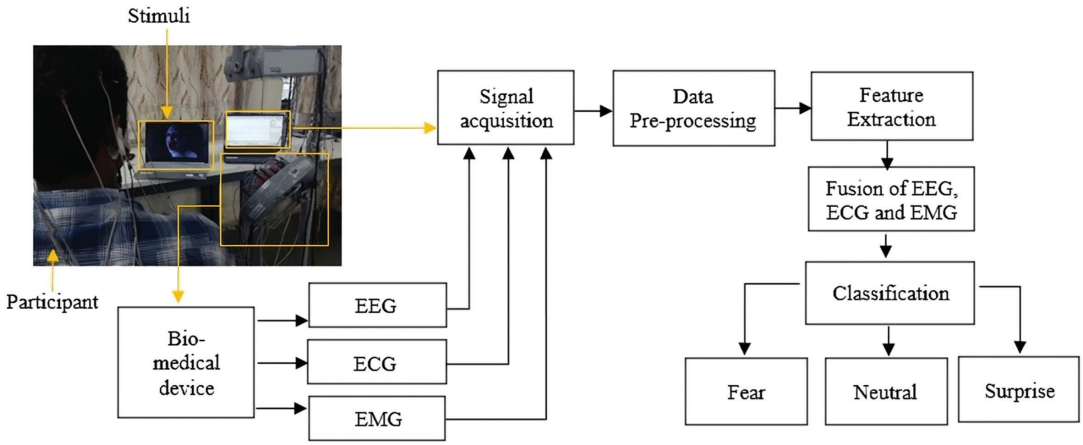


Fig. 4. Subject placement and laboratory setup for implemented workflow

3.2.4. FEATURE SELECTION AND FUSION

After extracting 35 features, for EMG (4 channel), the size of the feature vector was (280, 140) for a single participant, and for EEG (4 channel EEG + 1 channel ECG), it was (1070, 175). EDA was performed on these extracted features to categorize them according to their importance in classifying emotions. 11 less important features based on visual inspection of mean and variance were removed. The final size of the feature vector becomes (280, 129) for EMG and (1070, 164) for EEG.

As the sampling rates for EEG and EMG were different, the number of data points generated was also different. The number of data points generated for EEG and ECG was higher than for EMG. As a result, up-sampling (ups) and down-sampling (ds) techniques were used to prevent a biased model, as shown in Fig. 5. One such synthetic sampling technique is the Synthetic Minority Oversampling Technique (SMOTE), whose formula is given in (1). The SMOTE process involves identifying the S closest neighbors to every sample in the minority class and then using those neighbors to generate synthetic samples. This is achieved by computing the dissimilarity between a sample from the minority class and one of its S nearest neighbors and multiplying it by a random number between 0 and 1 to generate a new synthetic sample with a slightly different feature set. This preserves important information about the minority class and helps to improve model performance.

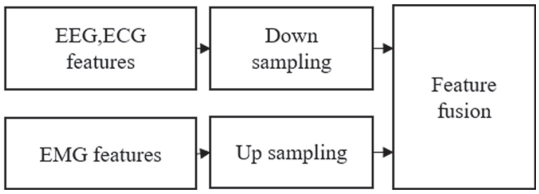


Fig. 5. EEG, ECG, EMG feature data with synthetic sampling

For a given sample from the minority class, denoted as w, SMOTE generates a synthetic y as:

$$y = w + \lambda^* (neighbor\_i\text{-}x) \quad (1)$$

where λ is a random number between 0 and 1, and $neighbor\_i$ is one of the nearest neighbors to w, where s is the number of neighbors to consider. The resultant matrix was formed after using the smote function on the EMG dataset.

Algorithm 1: Fusion of EEG, ECG and EMG features

Input: Subject-wise dataset on EMG, EEG and ECG
Output: Fusion Dataset for Model Training
Initialization
1: fv(EMG,EEG,ECG)←No. of feature vector
2: Nn←Nearest Neighbours of a Sample
3: kN(class samples)←Minority Samples
4: N←No. of Samples
5: Compute $N1$ ←Number of Classes
6: if fv(EMG)<fv(EEG,ECG) then
7: fvo ← Split fv(EMG) into $N1$ frame
8: end if
9: for i==class do:
10: if kN≥Nn then
11: pre set ←$kN(i)$:N(Classes-i)
12: smote (y) ← Ups(pre set)
13: emg ($y1$) ←y+fvo
14: if $y1$<fv(EEG,ECG) then
15: $N2$ ←$fv(y1)$
16: $y2$←ds[N(EEG,ECG)]
17: end if
18: end if
19: end for
20: Final model= $y1$+$y2$
21: return Final model
end

The EMG dataset was first divided into training and testing sets before applying SMOTE. The subsequent technique was only applied to the training set. This training set was further split into subject-wise sets, and each test was further divided using labels assigned to the emotions.

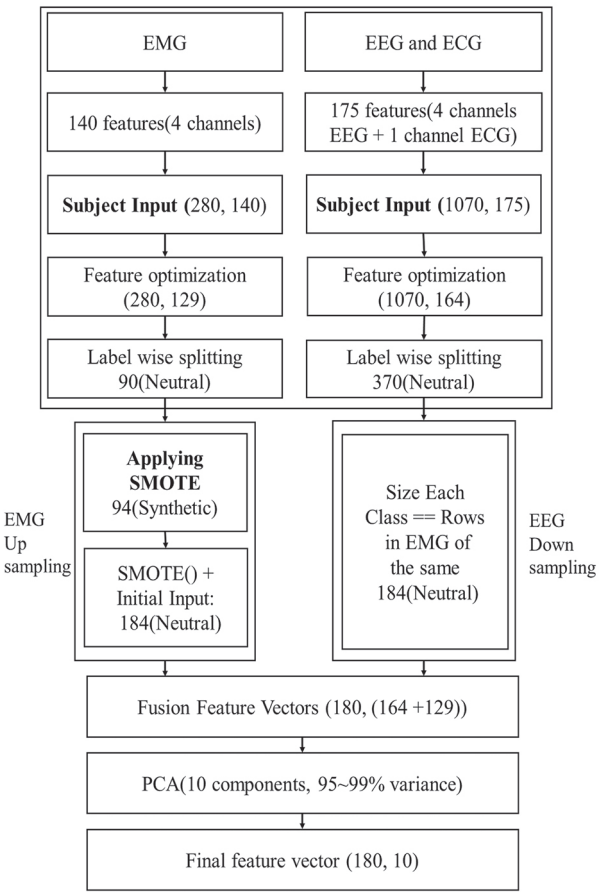


Fig. 6. Description of feature vector size at each step of the methodology

Considering a single participant, the feature vector of 280 rows was split label-wise, resulting in the ratio for neutral, fear, and surprise being 90:96:94. This helped to separate out the data for each subject's emotions. These were then randomly selected with a ratio of S: Majority class samples among the chosen emotion label vs. the rest of the labels, resulting in class unbalancing among the EMG labels for every subject.

The value of neighbor_i was observed to be 7. Hence, the initial minority input for each individual label S was chosen to be greater or equal to 7, i.e., 10. The resultant ratio was 10:96, which was stored in a separate set. The Smote technique was subsequently applied to this dataset. The process occurred recursively for each subject's label, resulting in a unique ratio of 96:96.

The original feature vector was then appended to these synthetically generated sets, which resulted in a ratio of 184: 192: 190. This provided the final resultant training set for each subject as (566, 129) for the EMG.

As the feature vector of the final resultant training set of EMG (566, 129) is less than the total size of the feature vector of EEG and ECG (1070, 164), a down-sampling technique was used. Random down-sampling was applied to the EEG dataset for each subject on separate labels to equalize the rows. This method randomly chooses a subset of samples from the majority class to match the size of the minority class. This created separate datasets with the same rows and ratios as in the EMG set. These separate data frames were then appended together to form the final resultant training set for each subject as (566, 164) for the EEG. The resultant training set for EEG, ECG, and EMG was merged to produce a fusion model for the recognition of emotions. This fusion model was further optimized for classification. The complete process of change in the feature vector size can be visualized in Fig. 6.

3.2.5. FEATURE OPTIMIZATION AND CLASSIFICATION

Feature optimization is crucial in ER as it aims to enhance the accuracy and efficiency of classification models by selecting the most relevant and significant features from the dataset. Feature optimization streamlines the classification process, resulting in improved performance and a better understanding of emotional patterns by reducing dimensionality and focusing on essential attributes.

3.2.5.1. FUSION DATASET

The resulting training sets for EEG, ECG, and EMG are amalgamated to form a fusion model capable of ER. The fusion set consisted of (3898, 292) samples. Feature vector tuning techniques were employed to further optimize this fusion set.

A dimensionality reduction technique, PCA, was applied to the fused multi-dimensional dataset to reduce computational complexity and enhance classification accuracy. It was used to identify and extract the most significant features from the resultant fusion dataset into a new set of variables. Here, a set of 10 principal components was chosen, and the final optimized feature matrix was obtained, which was of size (3898,10).

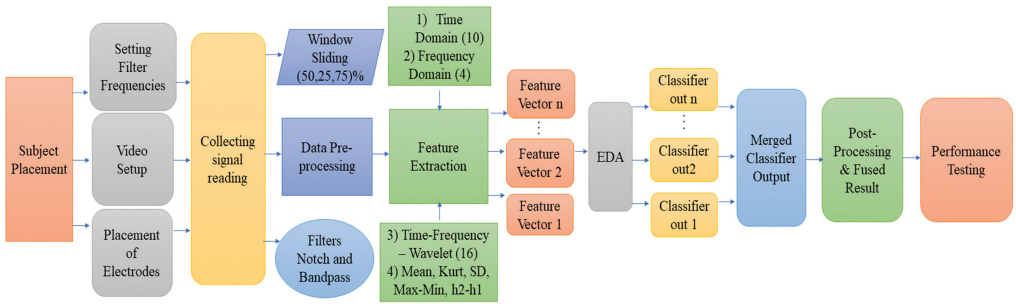


Fig. 7. Schematic of the model representing systematic steps from subject placement to model testing

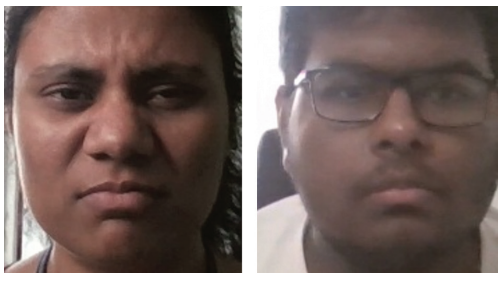
The model testing was done in two stages. First, the individual datasets of EEG, EMG, and ECG were used and optimized. Secondly, the fusion model was used. The classifiers used were decision tree classifier, random forest classifier, and Optimized random forest. After performing classification, the outputs of the classifiers were merged, followed by post-processing optimization. Different performance parameters were calculated, including accuracy, F1-score, precision, recall, ROC, and Area Under Curve (AUC). The complete process for physiological signal-based ER is illustrated in Fig. 7.

3.3. MODEL BASED ON FACIAL EXPRESSIONS

The facial model architecture was implemented in four main steps: facial data acquisition, image pre-processing, feature optimization, and classification. Each part plays a crucial role in the overall process of physiological ER.

3.3.1. FACIAL DATA ACQUISITION

A custom dataset was created for the face detection process. Sample images are illustrated in Fig. 8. The data acquisition includes participants exposed to a variety of stimuli and scenarios carefully designed to elicit specific emotions. A total of 10 participants were chosen, and video clips for various emotions were shown to them. As participants responded to these stimuli, their facial expressions were captured by the camera, resulting in a diverse dataset of facial images depicting different emotional states. This dataset consisted of face images captured against a uniform background. Firstly, normal images capturing various facial expressions were obtained. These images were then cropped manually to isolate the face region for more precise facial feature analysis. A total of 4,000 facial images were collected. These facial images were then categorized into four emotions, namely fear, surprise, neutral, and "other emotion". To enhance the visibility of facial cues during data acquisition, appropriate lighting conditions, facial angles, distances, and orientations were maintained. This ensured the proper capture of facial cues during each emotion.



(a) (b)

Fig. 8. Images for (a) fear and (b) neutral in the dataset

3.3.2. DATA PRE-PROCESSING

The collected images were resized to 64x64 and converted from RGB scale to grayscale for the facial dataset.

To enhance the system's generalization capability, further preprocessing was performed on the images, including normalization, standardization, and histogram equalization. Fig. 9 depicts the output of the pre-processing stage. This pre-processed dataset was subsequently partitioned into training, testing, and validation sets, which were then forwarded for feature extraction.

(a) (b)

Fig. 9. Pre-processed images

3.3.3. FACIAL FEATURES

Facial features represent the different properties and characteristics of facial images. Different techniques were used for extracting the required facial key points from the dataset. To select the features from the face, different techniques were studied that focused on extracting texture and shape information. Gabor filters, Histograms of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Local Ternary Patterns (LTP) filters were selected and applied to the pre-processed images to extract facial landmarks. The selection of these features was based on their effectiveness in capturing multiple aspects of facial expressions.

The pre-processed dataset images were passed through the feature extraction filters, and the resulting features were used to train a classification model. In the presented method, techniques for feature extraction, namely Gabor, LBP, LTP, and HOG, were used. These techniques were designed to capture different aspects of facial expressions, such as texture, shape, and spatial relationships between facial features. The method of feature extraction is illustrated in Fig. 10.

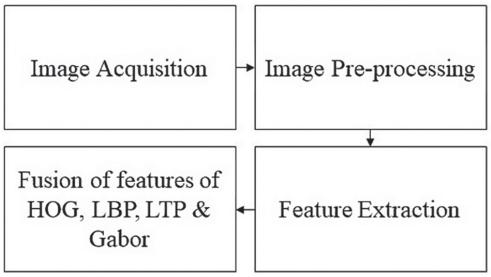


Fig. 10. Facial feature extraction

HOG was used to detect edges and oriented features in images. It detected the presence of facial expressions such as smiles, frowns, and raised eyebrows. This was achieved by detecting the changes in orientation

and gradient magnitude of the facial features. The applied HOG filter with a cell size of 16*16 pixels on the pre-processed image obtained a vector size of (1,700). LBP captures the texture information of the image and detects facial expressions such as wrinkles, dimples, and other small facial features. These features were extracted by comparing the neighboring pixels' values and their intensities. LTP, a variant of LBP, captures the texture information of an image in a more robust way. It helped to detect facial expressions that involve subtle changes in texture, such as those associated with emotions such as surprise and fear. Gabor features, derived from Gabor filters, were used to detect features at different scales and orientations. It detected the presence of facial expressions such as wrinkles, creases, and other fine details that are associated with emotions such as fear and surprise. The features mapped from the GABOR came around to (1,200).

The resulting feature vectors from all four feature extraction methods were concatenated to form the final feature vector that represents the facial expression. The final feature vector had a size of (21445, 964) and was then optimized for better performance using PCA and K-Means.

3.3.4. FEATURE OPTIMIZATION AND CLASSIFICATION

The optimization of the facial features involved two steps, i.e., clustering and feature selection. K-Means was used for clustering the extracted data features. The number of clusters K is a hyperparameter, which was chosen as 18, based on the elbow plot. This technique helped to group the similar facial emotion feature set together, which was then split into six histogram bins for creating feature columns. These feature columns were further optimized using PCA to select important features and reduce their dimensionality, making them easy for further processing. Features representing 95% of the total variance contributed by 13 principal components were chosen to create a lower-dimensional representation of the data that still retains most of the important information. The results were then used as input for a classifier for better classification performance. The implementation stages for the facial model are illustrated in Fig. 11.

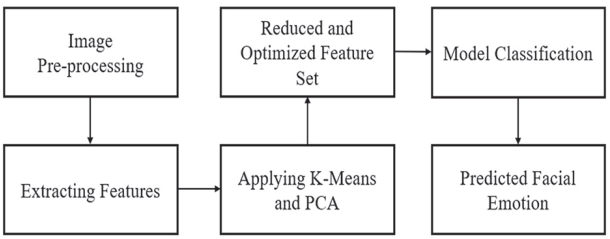


Fig. 11. FER process

Classification techniques like KNN, XGBoost, Random Forest, and decision trees are employed on the optimized feature vector. Decision trees recursively split the input space to form a tree-like structure for class prediction. KNN, which used k-nearest neighbors was initialized with a value of 59. For Random Forest, multiple decision tree predictions were combined to yield a more robust outcome, and it was initialized with 100 features. Similarly, XGBoost, an advanced gradient boosting algorithm, utilized 100 decision trees in each iteration, with the best one being 73. These techniques were then applied to the facial dataset for classification based on the provided labels. The models were trained on the extracted features, and their performance was subsequently evaluated.

3.4. VOTING BASED DECISION AND IMPLEMENTATION ON JETSON NANO

The facial and bio-electric fusion models were carefully selected with the consideration that the system needed to be deployed in real-time, aiming for a balance between performance and time requirements. In the video capture process, where subjects displayed specific emotions, the video stream was segmented into individual frames.

Algorithm 2: Voting based Decision on the Candidate Emotions

Input: Video input from camera & Raw input from the Datasheet
Output: Voting based Decision on the final output
Initialization
$f(x) \leftarrow$ features for face emotion, $f(y) \leftarrow$ features for bio-emotion
$r \leftarrow$ ROI (Face)
$fe(m) \leftarrow$ Facial Emotion Trained Model
$be(t) \leftarrow$ Fused Bio-Emotion Trained Model
1: for frame $f(x,y)$ in video input
2: if frame then
3: $f(y) \leftarrow$ Feature Extraction (Signal Input)
4. Pred_bio_emotion $\leftarrow be(f(y))$
5: ROI $\leftarrow$ Haar Cascade(Frame)
6: if $face=True$ then
7: $Op \leftarrow Feature\ Extraction$ (ROI)
8: $Pred\_f \leftarrow fe$(Op)
9: end if
10: end if
11: end for
12: Count Number of $Pred\_f$ & $Pred\_b$
13: final decision = $Max\ vote(emotion)$
14: return final decision
end

Accurate real-time ER relies on the identification and interpretation of human facial expressions, making the localization of the face within each frame crucial to identifying the Region of Interest (ROI). To accomplish this, the Haar Cascade Localization technique was employed for face detection. Once the face was successfully recognized through the HAAR cascade, the captured image underwent pre-processing, includ-

ing feature extraction and filtering. This pre-processed data was then fed into a classification model, allowing the real-time prediction of emotions based on this comprehensive facial analysis. A review of facial model implementation is given in Fig. 12.

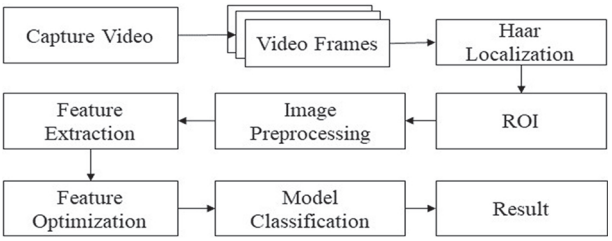


Fig. 12. Complete implementation of facial model from data acquisition to model classification

Jetson Nano was selected as the embedded development board as it is suitable for handling real-time, scalable ML models. Here, a Logitech C270 camera was interfaced with the Jetson Nano Board to capture the real-time video. The Jetson Nano was configured accordingly, and then a bio-datasheet was stored in it. The developed system was a stand-alone system in which the video captured by the camera was processed to predict facial emotion, which was fed to the final voting mechanism. Simultaneously, the bio-emotion predicted by the bio-electric model was also integrated into the system. Depending on the frequency of emotions, for every 2-second duration of video input, the final output emotion was given as described in Algorithm 2. This resulted in five voting classifier decisions for a duration of 10 seconds of video input.

This array of outputs was further given to the GPT language model to generate a medical report.

3.5. CONCLUSIVE REPORT GENERATION USING GPT

GPT is a state-of-the-art natural language processing technology developed by OpenAI. It is a type of artificial neural network that has gained significant attention in the fields of AI and ML due to its remarkable ability to generate human-like text. GPT models are pre-trained on massive amounts of text data from the internet and then fine-tuned for specific tasks, making them highly versatile for various natural language understanding and generation tasks. The core of GPT technology is the Transformer architecture, which is designed to handle sequential data efficiently. Transformers use a mechanism called attention to process and generate sequences of text. GPT models are accessed and utilized through an Application Programming Interface (API) provided by OpenAI, allowing for a wide range of natural language processing tasks.

The final output from the voting classifier based on the predictions from the physiological and vision classifiers is given to the GPT using this API. The use of GPT here is to synthesize these diverse emotional cues into a comprehensive conclusion. For a 10-second video, the voting classifier produces a decision every 2 seconds, resulting in 5 sequential emotions. Using this array of emotions, GPT generates a comprehensive medical report. This report succinctly encapsulates the array of emotions exhibited by the subject throughout the 10-second timeframe. This approach allows for a holistic understanding of the subject's emotional state, bridging the gap between physiological and visual indicators to provide valuable insights.

4. RESULTS AND DISCUSSION

The major part of the proposed study focuses on offering a comprehensive and systematic approach to ER, effectively combining physiological signals and facial expressions to achieve more accurate and nuanced emotion estimation. Utilizing GPT, a medical conclusion derived from these emotions yields promising implications for medical applications.

4.1. CHANGES IN PHYSIOLOGICAL SIGNALS AND FACIAL EXPRESSIONS

In this experiment, several alterations were observed while participants were expressing a particular emotion. These variations in traits provided valuable insights into the unique patterns exhibited in physiological and facial signals across different subjects.

4.1.1. VISION

During emotions like fear and surprise, there is an observable increase in electrical activity in the muscles responsible for controlling eye and brow movements, reflected in the EMG signal. The amplitude and voltage of the EMG signal are elevated as the electrical impulses to these muscles intensify. This heightened electrical activity corresponds to widened eyes, raised eyebrows, and a more open mouth, all of which are indicative of the emotional response.

4.1.2. EEG

Fear and surprise lead to increased electrical activity in specific brain regions, particularly the amygdala. This heightened electrical activity is reflected as an increase in the amplitude of the EEG signal. The amygdala's role in emotional processing and arousal results in these elevated electrical signals. Conversely, during neutral emotions, the EEG signal shows a more balanced amplitude across various brain regions, reflecting a state of relative emotional calm.

4.1.3. ECG

During fear and surprise, there is an increase in the electrical signal's amplitude, which directly corresponds to a rise in heart rate. The sympathetic nervous system's activation during these emotions leads to the generation of stronger electrical impulses in the heart,

causing an elevated heart rate. Simultaneously, heart rate variability may decrease as the electrical signals governing heartbeat rhythms become more regular.

4.1.4. EMG

During fear and surprise, there is an increase in the amplitude and voltage of the EMG signal in facial muscles, such as those controlling eyebrow and mouth movements. These changes reflect heightened muscle tension in response to emotional arousal. In contrast, during neutral emotions, the EMG signal exhibits a more relaxed state with lower amplitude and voltage, indicating a lack of pronounced muscle contractions. In the case of EMG, during fear and surprise, there is an increase in the amplitude and voltage of the electrical signals detected in facial muscles. These changes are a direct result of heightened muscle tension and increased electrical impulses in the muscles controlling facial expressions, such as those around the eyebrows and mouth. These electrical signals are indicative of the muscle contractions associated with the emotional response.

4.2. EVALUATION OF CLASSIFIERS

EDA was performed on both the facial and physiological features to visualize and identify the important features. These features were further optimized and then provided to the classifier. The performance of the classifier was estimated using different parameters, including accuracy, recall, precision, F1 score, and ROC.

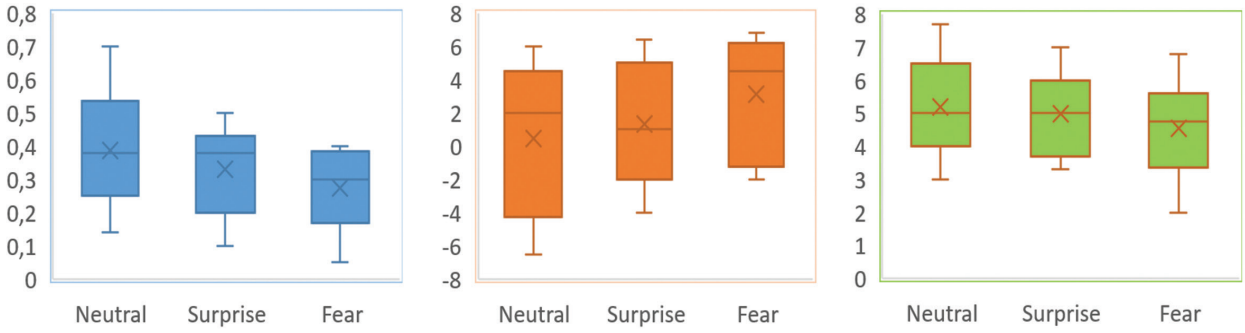


Fig. 13. Data analysis of some of the important features (a) Myopulse percentage rate (channel – 2), (b) Kurtosis (channel – 3), and (c) Mean frequency (channel – 1) considered in the model for different channels of device

4.2.1. PERFORMANCE OF BIO-ELECTRIC CLASSIFIER

EDA performed on the bioelectric features produced different variations depending on the emotions chosen. This helped to separate out the important features among the set of features and remove the redundant features. Features from the wavelet domain, namely mean and kurtosis showed distinct categorization in the minimum and maximum values of their boxplots. The myopulse percentage rate is an important feature in the time domain, showing variations in its EDA visualization aiding in the categorization of the different emotions captured by the EMG biomedical device. These important features were depicted through their boxplots in Fig. 13. Changes in performance parameters were observed using low frequencies (LF) and high frequencies (HF).

The frequency range was first chosen depending on the highest performance parameters and then changed according to the distribution of mean band amplitude over the EEG frequency bands. For EMG, when 2 and 10 were selected for LF and HF, respectively, the accuracy was around 63%. In the case of EEG and ECG, when 5 and 20 were selected for LF and HF, respectively, the accuracy was around 59%. Among all of these above classifiers, different classifiers for different scenarios have achieved better accuracy. The performance of different classifiers for each EEG, EMG, and ECG has been visualized. Accuracy was boosted up to 71% and 62% in the cases of EMG and EEG-ECG, respectively, by changing the notch filter frequency. These changes in the performance parameters have been visualized in Fig. 14 and Fig. 15.

Fig. 14. Parameter comparison of classifiers on the EMG dataset

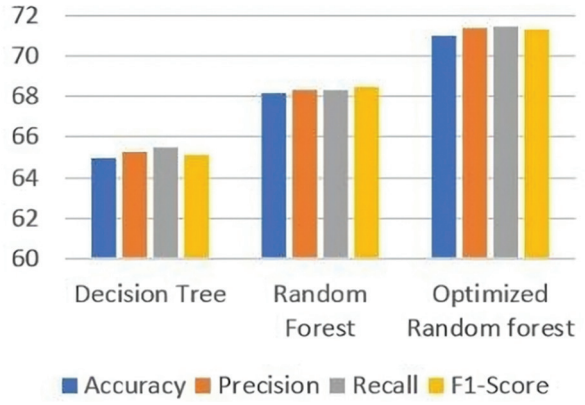


Fig. 15. Comparison of parameters of classifiers on EEG and ECG dataset

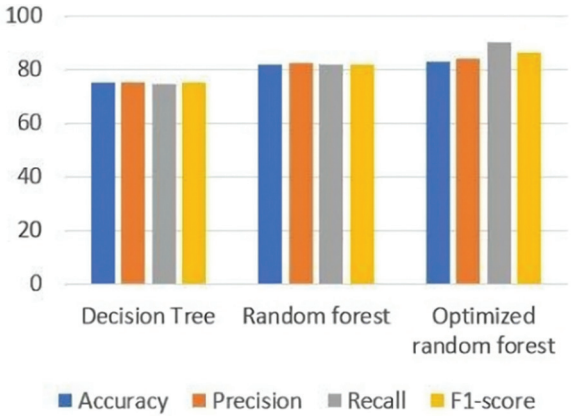


Fig. 16. Parameters comparison for models on class balanced fusion dataset of EEG, EMG, ECG with smote technique

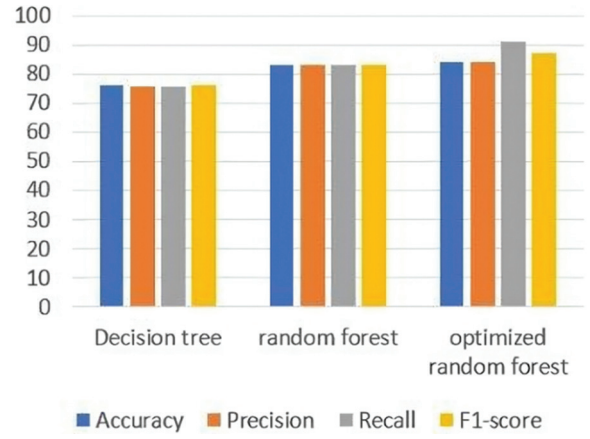


Fig. 17. Comparison of parameters for models on optimized fusion dataset after PCA

The resultant data set was obtained using fusion techniques. Further optimization was carried out by choosing 10 linearly uncorrelated variables and ordering them in terms of their importance in explaining 99% of the variance in the original data. These chosen principal components helped in identifying the key features or variables that have the most significant impact on the outcome variable, which aids in the final classification. After optimization, the accuracy was increased up to 84% in the case of EMG and 76% for EEG and ECG, as can be seen in Fig. 16 and Fig. 17.

The results highlight the importance of considering multiple performance metrics, as different metrics provide different insights into the performance of the model. Here, the parameters were calculated based on the Weighted-average which considers the relative importance of different performance parameters. The Weighted average precision was calculated based on precision per class and takes into account the number of samples of each class in the data, providing the highest result compared to other weights.

The fusion model showed a boost in classifier performance compared to individual datasets. The 84% accuracy of the optimized random forest model suggests that it was able to detect positive emotions effectively, which is important for clinical applications. The decision tree model achieved an accuracy of 76%. The precision of the models was 85% and 76.4%, respectively, providing the ability to correctly identify a particular emotion (e.g., neutral) among all the emotions it predicts. The recall of the models was compared, and the optimized random forest achieved the highest recall of 91%, showing its ability to identify particular emotions (e.g., fear) among all the samples that actually represent that emotion. This ensures that almost no positive cases are missed, which can improve patient outcomes. Both recall and precision are important in clinical settings and for prosthetic patients, as they provide insights into the model's ability to accurately detect positive and negative cases of a particular medical condition. It is important to strike a balance between precision and recall. Accounting for this F1 score, a harmonic mean of precision and recall, which provides a balance between the two metrics, was calculated. The models achieved an F1 score of 88% and 79%, respectively.

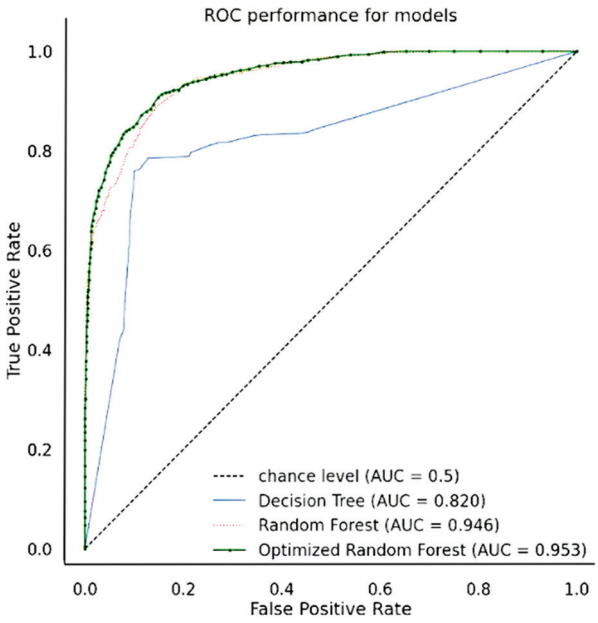


Fig. 18. ROC curve and AUC curve for neutral vs. all for the highest-achieving models

A multiclass ROC curve is visualized in Fig. 18. The multiclass ROC curve extends the binary ROC curve to multiple classes by considering each class as a positive class (neutral) and all other classes as the negative class (surprise & fear), plotting the ROC curve for each class, and then averaging them. The curve illustrates the trade-off between the true positive rate and false positive rate for each class, providing a comparison of model performance across different classes.

The optimized random forest classifier achieved the highest Area Under Curve of 0.95, demonstrating the model's capability to differentiate between the positive and negative classes. This model was then saved for implementation on the embedded system.

4.2.2. PERFORMANCE OF FACIAL FEATURE CLASSIFIER

Facial features were extracted from each facial image, with a specific focus on facial key points. Here, HOG, LBP, LTP, and Gabor feature techniques were chosen according to their ability to capture important facial cues. Depending on the emotion, distinct facial cues were identified. The visualization of these different facial features is illustrated in Fig. 19.

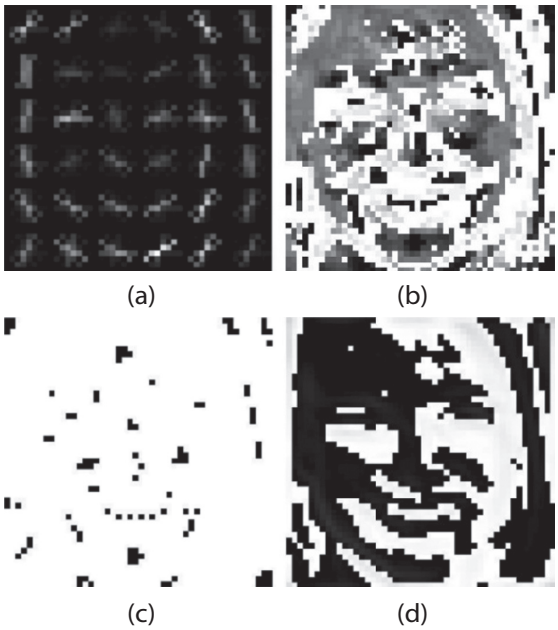


Fig. 19. Facial feature visualization on applying different feature extraction techniques (a) Hog, (b) Gabor, (c) LTP, and (d) LBP

This helped to visualize facial features and remove redundant features. Further optimization was carried out using clustering and PCA. This feature vector was then provided to the ML algorithms. Their performance metrics were calculated to provide useful insights for FER. Similar to the bio-electric model, here the parameters were calculated based on the Weighted average which considers the relative importance of different performance parameters. Fig. 20 summarizes the classification performance of each classifier for ER. KNN achieved the highest accuracy of 84.6%, followed by Random Forest with 74.3%, Decision Tree with 67%, and Xg-Boost with 64.5%. KNN also achieved the highest precision and F1-score for each emotion category, indicating its superior performance compared to the other classifiers. This model was saved and further used for embedded deployment.

Fig. 20. Classification performance of each classifier for emotion recognition

4.3. FUNCTIONAL/USABILITY EVALUATION

The models to be deployed on the embedded board were chosen based on the highest ratio of accuracy to time taken. A comparison of model accuracy and the time taken to predict the emotion after the input is given is shown in Table 3.

Model	Classifier	Accuracy	Time Taken
FER	KNN	84.6%	0.6 s
	Random forest	74.3%	0.8 s
	Decision Tree	67%	0.5 s
	Xg-Boost	64.5%	1.4 s
Bio-signal Fusion based ER	Optimized Random Forest	84%	0.3 s
	Optimized decision tree	76%	0.2 s

Table 3. Comparison of accuracy and time taken to generate prediction for bio-electric and facial models

Accordingly, for physiological signals and facial expression-based emotion prediction, optimized random forest and KNN were selected, respectively, to be used in the implemented system, which is illustrated in Fig. 21. Video recorded during participant under the ARS system was used as a basis to test the implemented system, and the facial expression video was mapped according to the physiological emotions. This helped to compare the ground-truth emotions with the physiologically and facially predicted emotions.

The EEG, ECG, and EMG fusion data was directly provided to the bio-electric classifier on the embedded system. The real-time video of the subject was captured at the same time. This helped process the video to predict the facial emotion and compare it with the predicted physiological emotion. The predicted emotions

from both sources were fed into the voting mechanism, which determined the final output based on the highest frequency of the recurring predicted emotion. Out of a total of 17 predictions made for the fusion of facial and bio-electric models, the average of 15 output predictions matched the ground truth emotion. The average accuracy for the fusion model was found to be 87.5%.

The predicted emotions were further subjected to a voting mechanism where, depending on the highest frequency of emotions in a 2-second window, the final output emotion is provided. This results in a total of five final emotion predictions during a 10-second video input.

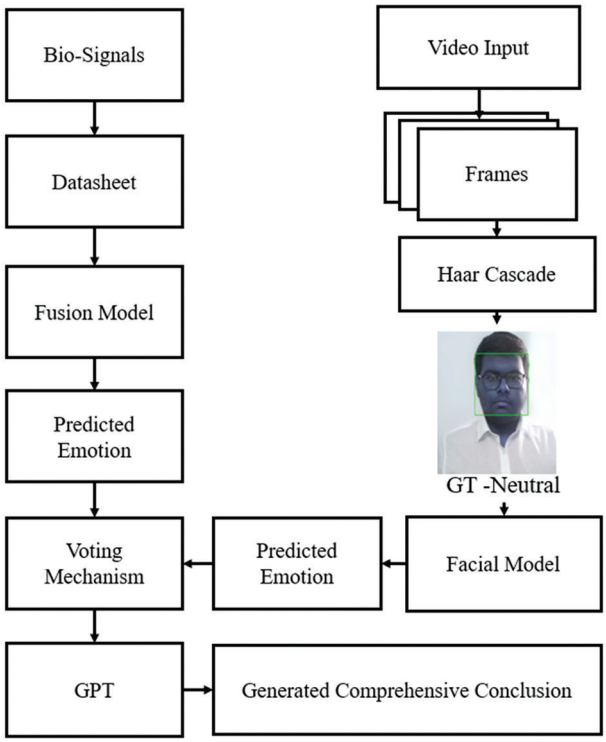


Fig. 21. Implementation of the system on a standalone embedded board

These emotions were provided to the GPT-3 model to generate a suitable, comprehensive conclusion. Its performance was compared by observing the similarity between the ground truth for the patient over the period of video input and the generated conclusion. The GPT-generated conclusion for an array of emotions, namely fear, is: "Based on the predicted outcome of the patient's emotions, it can be concluded that they experience fear, and it is important for healthcare professionals to provide emotional support and resources to help patients cope with it". This helps to provide a brief conclusion on the subject's emotions that are conveyed during the test.

4.4. SIGNIFICANCE AND COMPARISON WITH EXISTING WORK

The fusion of EEG, ECG, and EMG signals provides a holistic understanding of emotional states by capturing neural, cardiac, and facial muscle insights.

These signals collectively represent the diverse dimensions of human emotions. An integrated decision, combined with facial insights, is essential due to the intricate nature of emotions, accommodating individual variations, and enhancing accuracy. This integration adds robustness, overcoming challenges like noise, and cross-validation across signals further boosts credibility, and minimizes the risk of misclassification.

Table 4. Comparison of proposed system with existing approaches

Name	Dataset	Multi-Modularity	Fusion	Accuracy	Real-Time Implementation
Proposed System	Custom	Yes	Yes	87.5%	Yes
[6]	Available	No	No	75.31%	No
[9]	Custom	Yes	Yes	69.5%	No
[31]	Custom & Available	No	No	85.38%	No
[32]	Custom	No	No	84.3%	No

This proposed system has been compared with existing approaches in two ways, i.e., by comparing complete systems and by comparing only datasets. The effectiveness of the proposed facial model was assessed through testing on established benchmark datasets, including FER2013 [33] and CK+ [34], resulting in accuracy rates of 75% on FER and 90% on CK+. Due to the innovative nature of the fusion model incorporating EEG, EMG, and ECG signals, a dedicated testing dataset tailored to this unique approach was not available for assessment. A complete system comparison has been shown in Table 4 by using different parameters like multi-modularity, real-time nature, and system performance.

The proposed system model shows high accuracy in real-time systems compared to other existing ones. It uses a custom dataset and focuses on multi-modularity. The development of this system represents a significant advancement in the understanding and application of emotions within prosthetic systems and as general medical solutions. The integration of vision and physiological signals plays an integral part in user-machine interactions, allowing prosthetic arm users to experience an emotionally intuitive control interface.

5. CONCLUSION

The recognition of emotions can be achieved using signals such as EEG, ECG, and EMG, as well as facial expressions. The fusion of these signals and facial emotions has demonstrated a notable improvement in system performance. This study explored an ER system that uses EEG, ECG, and EMG data to identify the three emotional

states of fear, neutral, and surprise. Concurrently, facial expressions were also incorporated into the analysis. In this work, a dataset of fear, neutral, and surprise emotions for EEG, ECG, and EMG signals was meticulously collected. The subsequent steps involved the extraction of diverse domain-specific features, followed by the visualization of feature vectors. Important features were discerned and subsequently refined through the application of EDA and PCA The resultant fusion matrix was formed using fusion and class-balancing techniques on a combined dataset. These matrices were then subjected to classification employing specific classifiers. The accuracy of 84% and 76% were obtained using the optimized random forest model and the optimized decision tree, respectively, for the bio-signal-based model. Comparatively, when EEG-ECG and EMG signals were chosen individually, accuracy was lower due to the consideration of an unimodular system.

A boost in performance parameters was observed for the fusion dataset. Optimization also helped in increasing this accuracy. In the case of facial emotion, features, namely LBP, LTP, HOG, and Gabor were extracted and optimized by using K-means clustering and PCA. An accuracy of 84.6%, 74.3%, 67%, and 64.5% was achieved using KNN, Random Forest, Decision tree, and Xg-Boost Classifiers, respectively. Both modalities were used to determine the optimal emotion. To consolidate the results of both physiological and facial emotion analyses, a voting classifier was employed to determine the final emotional classification. This model was deployed in a real-time ER system by interfacing a Logitech C270 camera with the Jetson Nano board.

In this context, the system was beset with several issues, such as subject independence, which provides less accuracy owing to limited data. Expanding sample sizes is crucial to enhancing the generalizability of findings. By using more participants' data, higher accuracy can be achieved. Additional model tuning can be done using a greater number of channels and features. Other techniques and classifiers can be explored to improve the current performance estimators. Other preprocessing steps could be explored for recognizing facial key points.

The emotions recognized by this model can be used to improve the functionality and usability of prosthetic devices. By enabling users to control their prosthetics using their emotions, it can provide a more intuitive and natural user experience. Its inception followed a comprehensive survey of existing prosthetic system users, addressing the need for more personalized and responsive assistive technologies. The system not only enhances the functionality and comfort of the prosthetic system but also opens doors to a wide range of medical applications. It allows for emotionally aware control, enhances the user experience, and blurs the lines between human and machine. This novel integration not only improves the functionality of prosthetic arms but also profoundly impacts users' comfort and confidence. Moreover, while its primary application is within prosthetic arms, its adaptability and versatility make it usable for general medical systems. Its potential impact spans across the spectrum of medical and assistive technologies, making it a pioneering and transformative development.

Future research could focus on improving ER accuracy by incorporating other modalities, such as voice and body language, and using larger and more diverse datasets. This system holds promise for further refinements and broader applications. Additional parameters, such as Galvanic Skin Response (GSR) sensors, voice analysis, and temperature sensors, can be integrated to enhance its accuracy and versatility. This expansion of sensor types will enable even more precise recognition and response to emotional states, further improving the quality of life for prosthetic system users. The incorporation of these sensors may lead to applications in mental health support, user experience enhancement, and medical diagnosis, marking an exciting path for future research and development. The proposed work highlights its significance in the domain of ER and suggests the potential for further exploration in related fields.

6. REFERENCES:

[1] J. Chen, T. Ro, Z. Zhu, "Emotion Recognition With Audio, Video, EEG, and EMG: A Dataset and Baseline Approaches", IEEE Access, Vol. 10, 2022, pp. 13229-13242.

[2] S. Sharma, M. Bhatt, P. Sharma, "Face recognition system using machine learning algorithm", Proceedings of the 5th International Conference on Communication and Electronics Systems, Coimbatore, India, 10-12 Jun 2020, pp. 1162-1168.

[3] S. Kim, G. H. An, S. Kang, "Fusion expression recognition system using machine learning", Proceedings of the 14th International SoC Design Conference, Seoul, Korea, 5-8 Nov 2017, pp. 266-267.

[4] A. Kumar, N. Garg, G. Kaur, "An emotion recognition based on physiological signals", International Journal of Innovative Technology and Exploring Engineering, Vol. 8, No. 9, 2019, pp. 335-341.

[5] H. E. Houssein, H. Asmaa, A. A. Abdelmgeid, "Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review", Neural Computing and Applications, Vol. 34, No. 15, 2022, pp. 12527-12557.

[6] S. Hwang, M. Ki, K. Hong, H. Byun, "Subject-independent EEG-based emotion recognition using

adversarial learning", Proceedings of the 8th BCI International Winter Conference on Brain-Computer Interface, Gangwon, Korea, 26-28 February 2020, pp. 1-4.

[7] H. Ferdinando, T. Seppänen, E. Alasaarela, "Comparing features from ECG pattern and HRV analysis for emotion recognition system", Proceedings of the IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, Chiang Mai, Thailand, 5-7 October 2016, pp. 1-6.

[8] S. Katsigiannis, N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices", IEEE Journal of Biomedical and Health Informatics, Vol. 22, No. 1, 2017, pp. 98-107.

[9] S. Jerritta, M. Murugappan, W. Khairunizam, S. Yaacob, "Emotion recognition from facial EMG signals using higher order statistics and principal component analysis", Journal of the Chinese Institute of Engineers, Vol. 37, No. 3, 2014, pp. 385-394.

[10] M. Hafsa, B. B. Hajira, B. L. Meghana, Y. A. Arpitha, H. R. Niveditha, "Human Basic Emotion Recognition from EEG Signals using IOT", International Journal of Engineering Research & Technology IETE, Vol. 8, No. 11, 2020, pp. 47-50.

[11] N. Kumar, K. Kaushikee, M. S. Hazarika, "Bispectral analysis of EEG for emotion recognition", Procedia Computer Science, Vol. 84, 2016, pp. 31-35.

[12] A. Goshvarpour, A. Abbasi, A. Goshvarpour, "An accurate emotion recognition system using ECG and GSR signals and matching pursuit method", Biomedical Journal, Vol. 40, No. 6, 2017, pp. 355-368.

[13] H. S. Cha, C. H. Im, "Performance enhancement of facial electromyogram-based facial-expression recognition for social virtual reality applications using linear discriminant analysis adaptation", Virtual Reality, Vol. 26, No. 1, 2022, pp. 385-398.

[14] M. R. Kose, M. K. Ahirwal, A. Kumar, "A new approach for emotions recognition through EOG and EMG signals", Signal, Image and Video Processing, Vol. 15, No. 8, 2021, pp. 1863-1871.

[15] B. Silvio, A. Casanova, M. Fraschini, M. Nappi, "EEG/ECG signal fusion aimed at biometric recognition", Proceedings of the 18th International Conference on Image Analysis and Processing, Genoa, Italy, 7-8 September 2015, pp. 35-42.

[16] S. Ganguly, R. Singla, "Electrode channel selection for emotion recognition based on EEG signal", Proceedings of the 5th International Conference for Convergence in Technology, Bombay, India, 29-31 March 2019, pp. 1-4.

[17] K. P. Wagh, K. Vasanth, "Electroencephalograph (EEG) based emotion recognition system: A review", Proceedings of the 6th Innovations in Electronics and Communication Engineering, Hyderabad, India, 21-22 July 2017, pp. 37-59.

[18] Z. Cheng, L. Shu, J. Xie, C. P. Chen, "A novel ECG-based real-time detection method of negative emotions in wearable applications", Proceedings of the International Conference on Security, Pattern Analysis, and Cybernetics, Shenzhen, China, 15-17 December 2017, pp. 296-301.

[19] P. Sarkar, A. Etemad, "Self-supervised learning for ECG-based emotion recognition", Proceedings of the 45th International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4-8 May 2020, pp. 3217-3221.

[20] M. Chen, Y. Hao, "Label-less learning for emotion cognition", IEEE Transactions on Neural Networks and Learning Systems, Vol. 31, No. 7, 2019, pp. 2430-2440.

[21] Y. Wu, Y. Wei, J. Tudor, "A real-time wearable emotion detection headband based on EEG measurement", Sensors and Actuators A: Physical, Vol. 263, 2017, pp. 614-621.

[22] A. Topic, M. Russo, "Emotion recognition based on EEG feature maps through deep learning network", Engineering Science and Technology, an International Journal, Vol. 24, No. 6, 2021, pp. 1442-1454.

[23] A. Majumder, L. Behera, V. Subramanian, "Automatic Facial Expression Recognition System Using Deep Network-Based Data Fusion", IEEE Transactions on Cybernetics, Vol. 48, No. 1, 2016, pp. 103–114.

[24] J. Alghamdi, R. Alharthi, R. Alghamdi, W. Alsubaie, R. Alsubaie, D. Alqahtani, R. Ramadam, L. Alqarni, R. Alshammari, "A survey on face recognition algo-

rithms", Proceedings of the 3rd International Conference on Computer Applications & Information Security, Riyadh, Saudi Arabia, 19-21 March 2020, pp. 1-5.

[25] L. Li, X. Mu, S. Li, H. Peng, "A review of face recognition technology", IEEE Access, Vol. 8, 2020, pp. 139110-139120.

[26] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using Laplacian faces", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 3, 2005, pp. 328-340.

[27] T. Ahmed, P. Das, F. Ali, and F. Mahmud, "A comparative study on convolutional neural network-based face recognition", Proceedings of the 11th International Conference on Computing, Communication and Networking Technologies, Kharagpur, India, 1-3 July 2020, pp. 1-5.

[28] M. K. Chowdary, T. N. Nguyen, D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications", Neural Computing and Applications, Vol. 35, No. 32, 2021, pp. 1-18.

[29] S. Sharma, V. Kumar, "Performance evaluation of machine learning based face recognition techniques", Wireless Personal Communications, Vol. 118, No. 4, 2021, pp. 3403-3433.

[30] K. Koonsanit, N. Nishiuchi, "Classification of user satisfaction using facial expression recognition and machine learning", Proceedings of the IEEE Region 10 Conference, Osaka, Japan, 16-19 November 2020, pp. 561-566.

[31] Y. Liu, O. Sourina, "EEG-based subject-dependent emotion recognition algorithm using fractal dimension", Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, USA, 5-8 October 2014, pp. 3166-3171.

[32] T. Ergin, M. Ozdemir, A. Akan, "Emotion recognition with multi-channel EEG signals using visual stimulus", Proceedings of the Medical Technologies Congress, Izmir, Turkey, 3-5 October 2019, pp. 1-4.

[33] L. Goodfellow et al. "Challenges in representation learning: A report on three machine learning contests", Proceedings of the 20th International Conference on Neural Information Processing, Daegu, Korea, 3-7 November 2013, pp. 117-124.

[34] L. Patrick, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, "The extended Cohn-Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops, San Francisco, CA, USA, 13-18 June 2010, pp. 94-101.