



# Augmenting ECG Data with Multiple Filters for a Better Emotion Recognition System

Muhammad Anas Hasnul<sup>1</sup> · Nor Azlina Ab. Aziz<sup>1</sup> · Azlan Abd. Aziz<sup>1</sup>

Received: 24 August 2022 / Accepted: 18 December 2022 / Published online: 11 January 2023  
© King Fahd University of Petroleum & Minerals 2023

## Abstract

A physiological-based emotion recognition system (ERS) with a unimodal approach such as an electrocardiogram (ECG) is not as popular compared to a multimodal approach. However, a single modality has the advantage of lower development and computational cost. Therefore, this study focuses on a unimodal ECG-based ERS. The ECG-based ERS has the potential to become the next mass-adopted consumer application due to the wide availability of wearable and mobile ECG devices in the market. Currently, ECG-inclusive affective datasets are limited, and many of the existing datasets have small sample sizes. Hence, ECG-based ERS studies are stunted by the lack of quality data. A novel multi-filtering augmentation technique is proposed here to increase the sample size of the ECG data. This technique augments the ECG signals by cleaning the data in different ways. Three small ECG datasets labelled according to emotion state are used in this study. The benefit of the proposed augmentation techniques is measured using the classification accuracy of five machine learning algorithms; k-nearest neighbours (KNN), support vector machine, decision tree, random forest and multilayer perceptron. The results show that with the proposed technique, there is a significant improvement in performance for all the datasets and classifiers. KNN classifier improved the most with the augmented data and the reported classification accuracies of over 90%.

**Keywords** Emotion recognition · Electrocardiogram · Affective computing · Augmentation · Filter · Machine learning

## 1 Introduction

Emotion recognition system (ERS) has been popularised by the rising interest in artificial intelligence, especially towards instilling emotion in computer programs and robotics machinery [1]. This interest contributed to the growth of emotional artificial intelligence. Emotional artificial intelligence or affective computing is a field of study proposed by Rosalind Picard [2] that integrates computer science, psychology, cognition, and physiology to enable ERS [3].

A system that can recognize the emotional state of the user has huge potential in various fields. The industry that benefits from ERS includes healthcare, marketing, e-learning, entertainment, automotive, robotics, and security. In [4], ERS is used for driver's emotion detection to promote safe driving. Meanwhile, the application of facial-based ERS for smart

home automation can be found in [5]. For healthcare, various ERS applications had been proposed such as assisting in curing substance addictions [6], monitoring the emotional well-being of elderlies [7], and stress reduction therapy [8]. The ERS healthcare applications have been summarised in [9]. In [10], a multimodal ERS using facial and voice recognition is proposed to improve human–robot communication by recognizing the human's emotion and generating an appropriate affect response.

ERS can be categorized into two approaches which are multimodal and unimodal. Meanwhile, ERS that takes bio-signals as the modality for emotion recognition is known as a physiological-based ERS. The physiological-based multimodal approach combines different biosensors while the unimodal approach only utilises a single biosensor to detect emotions. The advantages of the unimodal over multimodal approach are that the data collection procedure is simpler, and the processing time and power required are significantly lower [11].

The heart signal collected using ECG is among the bio-signal used for ERS purposes. Nowadays, ECG device comes in mobile and wearable form with reliable signal quality.

✉ Nor Azlina Ab. Aziz  
azlina.aziz@mmu.edu.my

<sup>1</sup> Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia



However, the available ECG data collected for ERS are scarce and come in a small sample size [9]. This is due to the expensive procedures, and it is time-consuming. The COVID-19 pandemic makes it more challenging, as people are advised against physical interaction and contact. The lack of data hinders research and development of ECG-based ERS.

Hence, this work is focused on tackling the small affective ECG data challenge using a novel augmentation method of digital filters. A data augmentation increases a dataset size artificially. Although data augmentation is not common in ECG-based ERS, it is popular in cardiac pathological studies such as detecting arrhythmia [11, 12]. Here, the ECG signals are augmented by filtering the data using six filters; Neurokit, BioSPPy, Pan & Tompkins, Hamilton, Elgendi and EngZeeMod. The filters cleaned the data and removed noise using different mechanisms resulting in cleaned signals of dissimilar characteristics. The selected filters are commonly adopted for ECG filtering processes. Combining these filtered signals with un-augmented data increased the size of data by sixfold. The proposed method is then validated using our own collected data, A2ES, as well as datasets from other researchers namely AMIGOS [13] and DREAMER [14]. The data from these datasets are sourced from mobile devices namely Kardia and Shimmer. Both devices are reliable medical diagnostic tools available for consumers. All three datasets are considered small in size where the smallest is DREAMER with only 414 data. Five machine learning classifiers namely, k-nearest neighbours (KNN), support vector machine (SVM), decision tree (DT), random forest (RF) and multilayer perceptron (MLP) algorithms are used to classify the augmented and un-augmented data according to binary emotional model (BEM), affective dimensional model (ADM) of arousal and valence as well as discrete emotional model (DEM). The findings show that the proposed multi-filters augmentation offers an improvement of accuracy for all classifiers between 4 and 49%. KNN benefitted the most from the augmentation with the best accuracy obtained being 99%.

The next section is literature reviews relevant to the ECG-based ERS and data augmentation. Section 3 describes in detail the methodology of the study. The experimental settings are provided in Sect. 4. Section 5 contains the results and discussion, while Sect. 6 concludes the paper with suggestions for future works.

## 2 Literature Review

Due to the advantages of unimodal ERS and advantages of physiological based modalities which offer genuineness and hard to mask signals, several studies on unimodal ECG-based ERS had been reported. These works are carried either using open-sourced datasets or own datasets. For example,

Zong and Chetouani [15] utilized an open-source dataset, the AuBT dataset, in their study. The AuBT's ECG signal is filtered using an adaptive low-pass filter before the features are extracted using the fission and fusion of Hilbert Huang Transform (HHT). The data are then classified according to DEM classes using SVM, where 56% classification accuracy is achieved using fusion features and 69% using fission features.

Meanwhile, Bong, Murugappan, and Yaacob [16] designed an ECG-based ERS with their own collected dataset. The ECG signals are filtered using an elliptic band-pass filter and a discrete wavelet transform (DWT). Three-time domain features are being extracted and one of them is heart rate (HR). The results show that KNN performed better than SVM with around 10% differences between different training and testing splits. Xiefeng et al. [17] also use their own dataset. The pre-processing is done using Butterworth low-pass filter. The author extracted unspecified HRV features from the heart sound. A genetic algorithm is utilised to select the best feature combination, where 89.6% and 82.3% accuracy are achieved for valence and arousal, respectively. The accuracy for the combination of both scales using SVM is 72.9%. In other ERS work [18], ECG signals are collected for the study. The ECG signals are pre-processed using a bandpass filter with a removed 1–60 Hz baseline drift. The second filtration is done using a band-stop Butterworth filter at 49–51 Hz cut-off frequency. The features extracted are HR, HR stability and HR power and the classification is done using SVM. HR stability performed the best with 84.2% accuracy followed by HR power and HR.

Katsigiannis and Ramzan [14] proposed an ECG based ERS as part of their study. The raw ECG signals collected are directly extracted to get the PQRST statistical features and heart rate variability (HRV) features. The extracted features are classified using SVM and the accuracies achieved are around 62%. The dataset from this study is named as DREAMER and made available for other researchers. The ECG data from this dataset are adopted in this study.

Next, Correa et al. [13] proposed a dataset called AMIGOS, which is another dataset that is opened for other researchers. The dataset has ECG, electroencephalogram (EEG) and galvanic skin response (GSR) data. In AMIGOS' original work, a unimodal ERS is built from the ECG signals, where HR and HRV features are extracted. The classification is done using SVM. Since the data collected using short and long video scenarios, the accuracy presented is divided into three parts. Short video scenarios managed to get 53% and 55% accuracy for valence and arousal. The long video scenario gives out 55% and 54% accuracy while the combination of both scenarios results are 54% and 55%. AMIGOS is also adopted in this study.

Sarkar and Etemad [19] performed an ECG-specific study on AMIGOS and DREAMER datasets. The raw ECG signals are filtered using a high-pass infinite impulse response (IIR) filter with a bandpass of 0.8 Hz. The filtered signal is then normalised using Z-score normalisation. Rather unspecified spatiotemporal features are extracted and classified. The self-supervised convolutional neural network (CNN) results show a slightly better performance than the previous literature with around 85–89% accuracies on both datasets. Siddharth, Jung, and Sejnowski [20] also adopted AMIGOS and DREAMER in their study. Additionally, they also used the data from MANHOB-HCI [21]. Raw ECG signals from the datasets are pre-processed using a moving average filter with 0.25 s of window length. Then, HRV and spectrogram features are extracted. An extreme learning machine is used to classify the models. For AMIGOS, the individual accuracy for the ECG classification is approximately 82% for both valence and arousal, while for DREAMER, the results are around 80%. For MANHOB-HCI, the accuracies achieved are around 78%. The accuracy reported for both AMIGOS and DREAMER is better than the original work.

In [22], the authors adopted the AMIGOS dataset. The raw ECG signal is pre-processed using Pan-Tompkins QRS detection and filtered with 0.5–15 Hz cut-off frequency. Various features from time, frequency and nonlinear domain are extracted. The classifier used is a deep convolutional neural network (DCNN) and the accuracy reported for valence is 71% and for arousal, 81%.

Subramanian et al. [23] extracted ECG's HR and HRV features including other statistical features from their dataset ASCERTAIN. The classification results of the features using Naïve Bayes are better compared to SVM with 60% to 56% for valence and 59% to 57% for arousal. The ASCERTAIN dataset is opened to other researchers.

Chen et al. [24] use one of the largest affective dataset, the DECAF dataset [25]. The pre-processing steps are done by applying the Butterworth filter and extracting some features inclusive of HR. To the best of our knowledge, this study is the only ECG-based ERS study that has adopted data augmentation. A generative adversarial network (GAN) is applied to increase the number of ECG signals. GAN is the most popular data augmentation technique available [26] and some of the ordinary ways of doing it are through noise introduction, signal flipping, sine/cosine shifting, etc. [27]. The GAN generator creates fake but high-quality ECG signals while the discriminator validates them with the real signals. If the discriminator can no longer tell which signal is fake, then the generator has successfully created a string of ECG signals that is almost identical and close to the original one. The ERS performance shows an increase in accuracy when a higher ratio of augmented data is added alongside the original data in assisting the classification. Before augmentation, the study reported an accuracy of approximately 58%. The more

the synthetic ECG is introduced, the higher the accuracy is achieved. The average results for SVM and RF are above 63% respective to the combination of valence and arousal scale. The drawbacks of GAN are that reliability is questionable as well as the technique requires complex and high computing power due to its dependency on deep neural networks [28].

Table 1 summarises the reviewed works of ECG-based ERS. The sample size of the datasets reviewed is relatively small. This is one of the main challenges in the field. A learning algorithm triumphs on a large dataset allowing better pattern recognition during the training phase. Additionally, it is seen that by filtering the ECG signals, the ERS can achieve a better result. For example, [19] and [22] applied the filtering method and reported improvements in accuracy compared to the original work of AMIGOS and DREAMER. But existing works only used one type of filtering technique for data pre-processing or noise elimination, none of the work used multiple filters for data augmentation. Augmentation is a data manipulation technique that synthetically increases the data count by modifying the original data [29]. Even though data augmentation is popular for amplifying data size [16–18] it is not popularly adopted in ECG-based ERS. Among all the research reviewed here only one which is from [24] incorporates data augmentation, however, the accuracy reported is below 70%.

There are several non-ECG-based ERS research that adopted data augmentation. Luo [30] applied conditional Wasserstein GAN (CWGAN) framework to EEG data as an augmentation process to enhance the ERS. The technique generates realistic EEG data in differential entropy form raw EEG data. The generated data are classified into levels of data quality and only high-quality data are appended to the training models. After augmentation, the accuracy increases by 2.97%, 9.15% and 20.13%, respectively. Since CWGAN is just another variation of GAN, the drawbacks are as previously discussed. Chatziagapi et al. [31] implemented augmentation to rebalance the class labels of speech audio data. The GAN technique is used to generate synthetic spectrograms to increase the counts of the minority emotional classes. The magnitude of augmentation is recorded from 0.4, 0.6 and 0.8 to a fully balanced dataset. At fully balanced class labels, the accuracy improves by 5% and 10%. Data augmentation is popular with images. In a work on facial expression-based ERS [32], “Augmenter” an open-source library is used to augment the images by rotating, flipping, blurring, sharpening, embossing, and skewing them. The augmentation enlarges the data size to fit deep learning training. Meanwhile, image augmentation is used in [33] to increase the size of data 10 times and avoid overfitting.

Data augmentation is commonly used in medical applications. For example, a novel ECG augmentation technique is proposed in [34] where the study tries to solve the problems of an imbalanced dataset for atrial fibrillation (AF) detection.

**Table 1** Studies of ECG-based ERS

Papers	Dataset: Participants x Samples, Emotional Model	Pre-processing & Filtering	Features extracted	Best classifier: Result + Emotion
[15]	AuBT: 1 × 100, DEM	Adaptive low-pass filter	HHT (EMD and Hilbert transform) fission and fusion	SVM Fission: 69% Joy, Anger, Sadness, Pleasure SVM Fusion: 56% Joy, Anger, Sadness, Pleasure
[16]	Own: 5 × 15, BEM	Elliptic bandpass filter, DWT	Time domain: HR, MRamp, MRRI	KNN: 66.49% 60:40 train/test, 66.22% 70:30 train/test, 67.54% 80:20 train/test, Positive, Negative, Neutral 74.67% 60:40 train/test, 77.69% 70:30 train/test, 77.42% 80:20 train/test Positive, Negative SVM: 64.98% 60:40 train/test, 65.52% 70:30 train/test, 66.04% 80:20 train/test Positive, Negative
[14]	Own—DREAMER: 23 × 18, ADM	NA	- PQRST features: mean, med, std, min, max, range - HRV: RMMSD, PSD LF, PSD HF, LF/HF, total power	RBF SVM: 62.37% Valence (V), 62.37% Arousal (A), 61.57% Dominance
[13]	Own—AMIGOS: 40 × 16, ADM	NA	- Root mean square of inter-beat interval (IBI), mean IBI, 60 spectral powers, LF, MF, HF of HRV spectral power, HR - HRV statistics: mean, std, skewness, kurtosis, % of the time the future value above/below mean_std	Linear SVM: (Short video scenario): 53.5% V, 55.0% A (Long video scenario): 55.0% V, 54.3% A (Both): 54.5% V, 55.1% A
[23]	Own—ASCERTAIN: 58 × 36, ADM	NA	HR, HRV statistics: mean, std, skewness, kurtosis, % of the time the future value above/below means +std	SVM: 56% V, 57% A NB: 60% V, 59% A



**Table 1** (continued)

Papers	Dataset: Participants x Samples, Emotional Model	Pre-processing & Filtering	Features extracted	Best classifier: Result + Emotion
[18]	Own: 6 × 36, BEM	1–60 Hz bandpass filter, 49–51 Hz band stop Butterworth filter	HR, HR stability, HR power	Linear SVM: 69% HR Positive, Negative, Neutral, 84.2% HR stability Positive, Negative, Neutral, 70.4% HR power Positive, Negative, Neutral
[20]	AMIGOS, DREAMER, MANHOB-HCI: 27 × 20, ADM	Moving average filter with 0.25 s window length	HRV: pNN50 Spectrogram: 4096 features	Extreme Learning Machine (Individual): AMIGOS 81.89% V, 82.74% A, DREAMER 80.43% V, 80.68% A, MANHOB-HCI 78.76% V, 78.76% A
[24]	DECAF: 30 × 76, ADM	Butterworth filter	HR, DWT, QRS morphology: qrsWBR, qrsWRE, qrsABR, qrsARE, qrsMOR	SVM: 64.5% V, A RF: 63.4% V, A
[22]	AMIGOS	Pan-Tompkins QRS detection, 0.5–15 Hz cut-off frequency removal	- IBI time domain: meanNN, medNN, SDNN, rmSSD, pNN50, pNN20, coefVarSD, medADNN, coefVarNN, mCoefVarNN, Shannon Entropy, HRV triangular, numArtifacts - Frequency domain: peakHF, hTotalPowerRatio, normHF, peakLF, lfHfRatio, lfTotalPowerRatio, normLF, totalPower, ulfPeak, vhfPeak, vlfPeak - Nonlinear domain: correlation dimension, entropy, SVD, HF, LF, VLF; Shannon, fractal dimension Higushi and Petrosian, Fisher information	DCNN: 71% V, 81% A
[17]	Own: 16 × 96, ADM	Butterworth low pass filter	HRV	SVM: 72.9% V, A, 89.6% V, 82.3% A
[19]	AMIGOS, DREAMER	High-pass IIR filter with a bandpass of 0.8 Hz, Z-score normalization	High-level spatiotemporal features	Self-Supervised CNN: AMIGOS: 87.5% V, 88.9% A DREAMER: 85.0% V, 85.9% A

In a clinical setting, it is challenging to get ECG signals which contain AF traits from a diverse patient background. Thus, the majority of the ECG signals are from healthy patients, while the pathological signals suffer from data deficiency. First, the ECG signals are duplicated and concatenated. Then, the extrapolated ECG signals are resampled through a randomly selected augmented sequence. As a precautionary measure, the resampling permutations are ruled to not produce an exactly similar sample. The results claimed that after balancing the dataset with data augmentation, the training accuracy and the f1-scores increased significantly. This technique is similar to augmentation through a geometric transformation in image-based data. The repeatability issues in this technique may cause low variance and high bias to the augmented sample data. Therefore, careful attention is needed as this technique is not applicable everywhere [29].

Nonaka and Seita [35] tackle the issues of insufficient AF data with RandECG, a mixture of random ECG augmentation techniques. Various signal transformation methods are explored to introduce variations in the ECG signals. The transformation includes scaling, flipping, dropping, shifting, cut-out, and other noise additions such as square pulse, Gaussian noise, etc. The observation shows a relative improvement in detection accuracy by up to 3.51%. While the disadvantage of the transformation is the noise addition technique is considered a non-ideal in increasing the data count. This technique relies on a traditional signal-based augmentation process by adding more noise to an already noisy ECG signal. When too much noise is added to the raw signal, it pollutes the data and renders the affective information obsolete [36].

The works that used data augmentation for ERS and ECG are tabulated in Table 2.

### 3 Methodology

This work proposed a new application of filters which are for data augmentation. As discussed in the previous section, data augmentation improves system performance. Meanwhile, in the existing studies filtering ECG signals before classification contributed to better performance. However, the works implemented one filter only and the purpose is solely for data pre-processing. Given  $X$  number of signals, their filtering generated  $X$  number of filtered signals. In this study, multiple filters are applied to the signals and the output of all the filters is combined to increase the number of signals for classification training. Specifically, the ingested data are pre-processed with six different ECG filtration techniques and the filtered outputs are combined. To study the effectiveness of the proposed augmentation technique, ERS is built using several machine learning algorithms.

Figure 1 shows the flowchart of the processes. In the proposed ERS, there are four main phases namely pre-processing and data augmentation, feature extraction, data cleaning, as well as classification and performance assessment. During the pre-processing and data augmentation, the ECG's noises are cleaned using the six chosen filters. Besides cleaning the ECG signals, it also acts as the proposed data augmentation technique. The stream of raw ECG signals and the filtered ones are then combined, and the features are extracted. The extracted features include heart wave detection such as PQST detection and R-Peaks detection, HR statistical features, and HRV feature derivations.

Next, the data are cleaned and separated into two parts. The first part consists of HR and HRV features extracted from raw ECG signals only, while the second part is where the workflow combined the raw and filtered ECG features. The purpose of segmenting the pipeline is to compare the performance between before and after data augmentation. Cleaned data features are scaled and standardised before being split into training and testing in both pipelines.

The last process is the classification and performance assessment. An exhaustive classification using grid search upon five machine learning techniques is implemented on the training set. The assessment is done based on testing accuracy comparison and statistical analysis.

#### 3.1 Phase 1: Pre-processing and the Proposed Data Augmentation

Instead of noise addition or falsification, this study proposed a novel data augmentation technique through multiple ECG filtrations. Six types of ECG filters are chosen, and each of them cleans the raw and noisy signals in certain distinct ways.

The first filter is an inbuilt Neurokit's standard filter proposed by Makowski et al. [37]. The ECG signal and its sampling rate are the two parameters passed to the filtering algorithm. The Neurokit filtering method removes slow drift and DC offset using the 5th-order high-pass Butterworth filter. The Butterworth highpass filtering method is as shown in Eq. (1) where  $n = 5$ . The lowcut frequency,  $f_c$  is set to 0.5 Hz. The input frequency that is being filtered is represented by  $f$ . Neurokit method also filters out 50 Hz powerline noise by smoothing the signal with a moving average kernel with a width of one period of 50 Hz.

$$H(f) = \frac{1}{\sqrt{1 + \left(\frac{f_c}{f}\right)^{2n}}} \quad (1)$$

The second filtering method is BioSPPy, proposed by Carreiras et al. [38]. The filter removes the ECG signal frequencies which are below 3 Hz and above 45 Hz through a finite impulse response (FIR) bandpass filter. The technique

**Table 2** Data augmentation for ERS and ECG

Papers	Application	Data Source	Augmentation Technique	Advantage	Disadvantage
[24]	ERS	ECG	GAN	Average accuracy increases by 5%	Reliability is questionable when implemented in signal-based data as no ground truth verification can be made Complex Requires high computing power
[30]	ERS	EEG	CWGAN	2.97%, 9.15%, 20.13% increase in accuracy	Reliability is questionable when implemented in signal-based data as no ground truth verification can be made Complex Requires high computing power
[31]	ERS	Audio	GAN	Balancing imbalance dataset	Reliability is questionable when implemented in signal-based data as no ground truth verification can be made Complex Requires high computing power
[32]	ERS	Image	Augmenter	Increase data size for deep learning	Expensive
[33]	ERS	Image	Patches & reflection	Avoid overfitting	Applicable for image data
[34]	Medical – AF detection	ECG	Duplication, concatenation, & resampling	Balancing imbalance dataset	Repeatability issue which renders low variance & high bias augmentation set
[35]	Medical – AF detection	ECG	RandECG	Increase classification accuracy by 3.51%	Repeatability issue which renders low variance & high bias augmentation set Pollutes the raw data Might render the data to obsolescence

applies a linear digital filter twice, once forward and once backwards. The combined filter has zero phases and a filter order that of the original [39]. The order,  $N$  is decided based on  $0.3$  multiplied by the sampling rate with an addition of one, if the result is an even number. This is to enforce the order to be an odd number. Before getting the coefficient for the FIR filter, the frequency is normalised to Nyquist frequency. Then, the FIR filter is calculated based on Eq. (2). The  $x(n - i)$  is the input signals on each taps according to the order of the filter. The coefficient of the filter is represented by  $b_i$  where the range satisfy  $0 \leq i \leq N$ . The filtered output signal is represented by  $y(n)$ .

$$y(n) = \sum_{I=0}^N b_i x(n - i) \tag{2}$$

Pan and Tompkins [40] filtering method for ECG signal has been around for quite some time and it is famous for accurate pre-processing of QRS detection. In the first order ( $n = 1$ ), Butterworth bandpass filter is applied with a cut-off frequency of 5 Hz for low-pass,  $f_{cl}$  and 15 Hz for the high-pass,  $f_{ch}$  from Eq. (3). This method applies another derivative filter to highlight the frequency content and removes background noises. The “lfilter” from the SciPy library has the option of FIR or IIR filtration methods. The mathematical representation for the IIR filter is shown in Eq. (4). The feedforward and feedback filter order is represented by P and Q. The  $x(n - i)$  and  $x(n - j)$  are the input signals on each taps according to the order of the filter. The  $b_i$  and  $a_j$  are the feedforward and feedback filter coefficient while the filtered

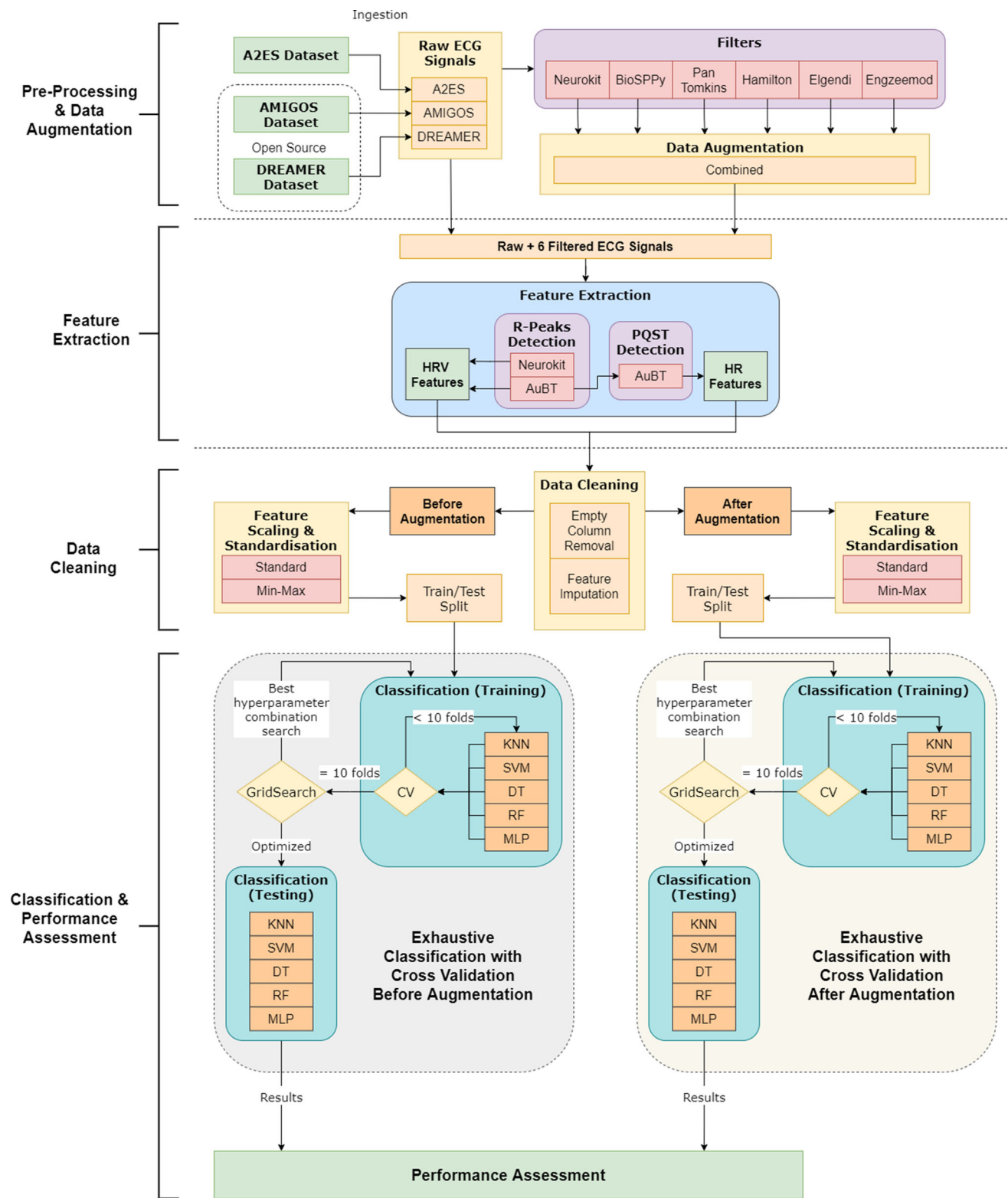


Fig. 1 Overview of the methodology



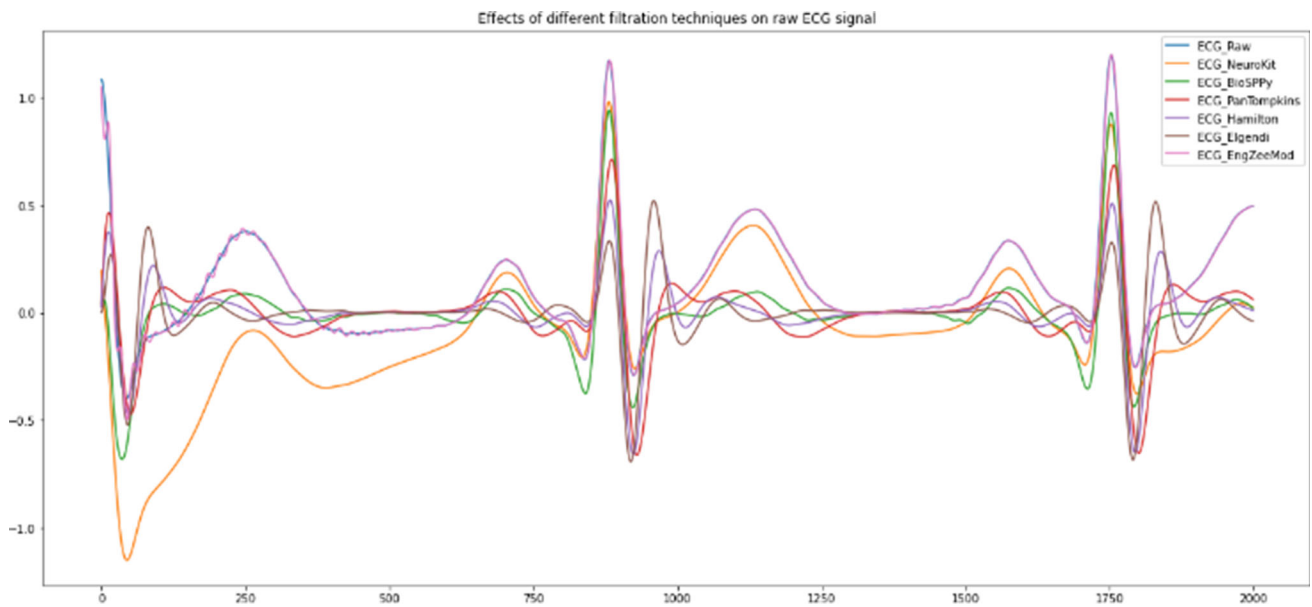


Fig. 2 Augmented ECG using various filters

Table 3 List of chosen ECG filters for ECG augmentation

No.	Filter name	Details	Justification
1	Neurokit	5th order highpass Butterworth filter with 0.5 Hz lowcut frequency, 50 Hz Powerline noise removal	The fastest and most accurate ECG filtering method
2	BioSPPy	FIR Bandpass filter ( $3 > x > 45$ Hz) with order = $0.3 \cdot \text{sampling rate}$	Open-sourced API for various bio-signals processing including ECG
3	Pan & Tompkins	1st order Butterworth Bandpass filter ( $5 > x > 15$ Hz), IIR/FIR filter	Highly cited and widely adopted ECG processing technique
4	Hamilton	1st order Butterworth Bandpass filter ( $8 > x > 16$ Hz), IIR/FIR filter	Improved version of Pan & Tompkins
5	Elgetti	2nd order Butterworth Bandpass filter ( $8 > x > 20$ Hz), IIR/FIR filter	Improved version of Pan & Tompkins
6	EngZeeMod	4th order Butterworth Bandstop filter ( $x < 48$ Hz, $52 < x < 52$ Hz), IIR/FIR filter	A modified version of the widely adopted ECG processing technique

output signal is represented by  $y(n)$ .

$$B(f) = \frac{1}{\sqrt{1 + \left(\frac{f}{f_{cl}}\right)^{2n}}} \cdot \frac{1}{\sqrt{1 + \left(\frac{f_{ch}}{f}\right)^{2n}}} \tag{3}$$

$$y(n) = \sum_{i=0}^P b_i x(n - i) - \sum_{j=1}^Q a_j y(n - j) \tag{4}$$

Hamilton [41] proposed a similar ECG filtering configuration with a slight variation in the cut-off frequency. In the first order ( $n = 1$ ), Butterworth bandpass filter is set at 8 Hz on the low-pass,  $f_{cl}$  and 16 Hz on the high-pass,  $f_{ch}$  threshold. The output signal becomes the coefficient for the IIR/FIR filter.

Elgetti, Jonkman, and Deboer [42] configured a second order ( $n = 2$ ), Butterworth bandpass filter with the cut-off frequency of 8 Hz on the low-pass,  $f_{cl}$  and 20 Hz on the high-pass,  $f_{ch}$ . Upon returning the cleaned signal, another round of IIR/FIR filtration is done based on the output coefficients.

Engelse and C. Zeelenberg [43] as well as Lourenço et al. [44] proposed a fourth-order ( $n = 4$ ), Butterworth bandstop filter as shown in Eq. (5). The cut-off frequency is between 48 and 52 Hz for the  $f_{cl}$  and  $f_{ch}$ , respectively. Similarly, a digital IIR/FIR filter is configured afterwards to remove more background noises.

$$B(f) = \frac{1}{\sqrt{1 + \left(\frac{f}{f_{ch}}\right)^{2n}}} \cdot \frac{1}{\sqrt{1 + \left(\frac{f_{cl}}{f}\right)^{2n}}} \tag{5}$$

**Table 4** Statistical features derived from HR

Statistical features	References	Type
Mean	PP	Interval
Median	QQ	
Standard deviation	RR	
Min.	SS	
Max.	TT	
Range	PQ	Interval
	QS	
	ST	
	P amplitude	
R amplitude		
S amplitude		

Total: 66 features

Based on the technique proposed, it is observed that the multi-filter generates smooth signals with slight variations in ECG patterns. Table 3 shows the summary of the listed filters used to augment the raw ECG signals for this study. Figure 2 visualized the effects of raw ECG signal when cleaned with the listed filtering techniques. The filters removed noise, smoothed, and amplified the signal differently.

### 3.2 Phase 2: Feature Extraction

Feature extraction is done using the Neurokit and AuBT toolboxes. Before extracting the HR and HRV features, the PQRST wave detection is done. Allocating these heart wave points is the foundation of feature extraction in ECG analysis. The heart wave detection is performed using Neurokit and AuBT toolboxes for R peak detection and only AuBT is used for PQST wave detection.

The HR is measured in beats per minute. Normally, a lower HR implies a healthier heart and vice versa. The advantage of HR is that it is easy to measure and does not need extreme signal accuracy to acquire it. One cycle of a heartbeat can be measured between any two peaks. Using RR peaks is the most common way of detecting HR. The use of HR features for affective recognition is seen in various studies [45]. In this study, a total of 66 HR features are extracted using AuBT. Table 4 shows the summary of the statistical features derived from HR with a different type of reference.

HRV measures the variability or specific changes in time between successive heartbeats or known as the RR peaks (interval). Low HRV indicates the physiological states of stress while high HRV indicates a recovery state of a person from the condition [46]. With a proper analysis method, HRV is considered the most precise non-invasive/intrusive

method to detect ANS activity [47] but it is difficult to measure while moving or during exercise [48]. HRV also contains evidence of ANS activity traits including emotional changes within an individual [49]. HRV features are the most used indicators for identifying emotions in a physiological-based system [50].

There are three domains from which HRV features are derived, namely time/temporal and geometric, frequency/spectral, and nonlinear domains. Neurokit features include all three domains while AuBT features are only available in the time domain. There are 52 and 14 HRV features extracted using Neurokit and AuBT respectively. The summary of the features is recorded in Table 5.

### 3.3 Phase 3: Data Cleaning

#### 3.3.1 Data Cleaning

Missing and incomplete data are common in real-world studies. However, these may handicap the statistical prediction as well as introduce bias to the results if not handled properly [51]. So, after the features have been extracted, the data cleaning processes that include empty column removal and feature imputation are conducted.

The empty column removal is a straightforward cleaning process where any feature that does not return any value is discarded. ULF and VLF features from HRV return an empty column for all ECG signals. This is because both features need a longer period of ECG recordings to literally detect the frequency bands. Thus, these feature columns are discarded.

The second cleaning method is through feature imputation. The returned features being extracted are not always clean. There are three possible returned feature states of the extracted data. For the empty cells and the '#NAME', the imputation is done based on averaging the columns, and then the cell is filled or replaced with the value. Although this technique is vulnerable to bias, it is the most common imputation technique practised in data science [52]. For the 'inf' cells, the replacement is done using the largest value in the column.

#### 3.3.2 Feature Scaling and Standardization

Applying feature scaling or standardization is important to reduce inaccuracy in machine learning models. Different scaling and standardization methods have been proven to affect the model's performance [53]. Scaling data does not change the shape of the distribution, but it changes the range of the values. Meanwhile, standardizing data changes the values so that the distribution's standard deviation equals one. Machine learning algorithms such as KNN, SVM, and multi-layer perceptron (MLP) are known to converge faster with scaled or standardized data. In this study, two methods

**Table 5** List of HRV features extracted

Domain	Neurokit features	AuBT features	
Time and geometric domain	RMSSD	Mean	
	MeanNN	Median	
	SDNN	Standard deviation	
	SDSD	Min.	
	CVNN	Max.	
	CVSD	Ranges	
	MedianNN	pNN50	
	MadNN	Distribution mean	
	MCVNN	Distribution median	
	IQRNN	Distribution standard deviation	
	pNN50	Distribution Min.	
	pNN20	Distribution Max.	
	TINN	Distribution range	
	HTI	Distribution triangular index	
	Frequency domain	ULF ( $\leq 0.003$ Hz)	LFHF
VLF (0.0033–0.04 Hz)		LFn	
LF (0.04–0.15 Hz)		HF <sub>n</sub>	
HF (0.15–0.4 Hz)		LnHF	
VHF ( $\geq 0.4$ Hz)			
Non-linear domain	SD1	C1d	
	SD2	C1a	
	SD1SD2	SD1d	
	S	SD1a	
	CSI	C2d	
	CVI	C2a	
	CSI MODIFIED	SD2d	
	PIP	SD2a	
	IALS	Cd	
	PSS	Ca	
	PAS	SDNNd	
	GI	SDNNa	
	SI	ApEn	
	AI	SampEn	
	PI		
	Total	52 features	14 features

are implemented which are Standard Scaler and Min–Max Scaler, which are adapted from [54].

### 3.3.3 Train/Test Split

Before feeding the machine learning models with the scaled and standardized features, the data is split into training and testing sets. The splitting method is done using Scikit Learn

[54] where each feature set is divided into an 80:20 ratio of training and testing. The proportion of class labels in the training set is identical to the samples for every dataset. This is achieved through stratifying the train/test split. The random state is set to an integer for reproducible output across multiple function calls. The deterministic nature of the random state also acts to control the shuffling applied to the data before proceeding with the split.

### 3.4 Phase 4: Classification and Performance Assessment

The classification and performance assessment are done using five supervised machine learning classifiers. These classifiers are chosen to evaluate the effectiveness of the multi-filter data augmentation proposed. The classifiers are KNN, SVM, DT, RF and MLP.

#### 3.4.1 K-Nearest Neighbour

KNN is a non-parametric classification algorithm that is known as a lazy learner. KNN keeps all the training data to make future predictions by computing the similarity between an input sample and its training instance. The tuned hyperparameter values in this study are the number of neighbours, weights and distance metrics. There are various distance metrics available for the KNN algorithm, but Manhattan and Hamming are selected because of their ability to learn the data well.

#### 3.4.2 Support Vector Machine

SVM is a supervised classification algorithm that separates data into classes using hyperplanes. It also uses kernel tricks to transform the data and optimize the decision boundaries. The hyperparameters tuned for SVM are the kernel function, gamma, and C. Since the assessment involves multidimensional classification, only the radial basis function (RBF) kernel is considered. Gamma is the degree of curvature of the hyperplanes while C is the degree of the error margin.

#### 3.4.3 Decision Tree

DT has a flowchart-like tree structure, and it is non-parametric. The highest node is considered the root with the branches that represent the decision rule with an outcome leaf node. The hyperparameters tuned for DT are the splitting criteria, minimum sample leaf, minimum sample split and maximum depth. The splitting criteria considered are Gini and Entropy.

#### 3.4.4 Random Forest

RF creates decision trees for different samples and randomly selects the best solution by the means of voting. The logic behind RF is that the more trees are sampled, the more it reduced the bias, and the better it generalized the data. Thus, many sample decision trees make up a forest. The hyperparameter values tuned for this algorithm are the number of estimators, the maximum number of features, the maximum depth, and the criterion. Again, the criterion is actually the

splitting criteria, and the considered techniques are Gini and Entropy.

#### 3.4.5 Multi-layer Perceptron

MLP is a neural network based supervised learning algorithm that trains using backpropagation. MLP algorithm from Scikit-Learn is considered a basic deep learning model that propagates the error in a backward direction to update the weights of the hidden layers. The tuned hyperparameters are the activation function, the hidden layer size, the solver, the alpha and the learning rate.

## 4 Experimental Setting

### 4.1 Dataset

This research uses raw ECG signals from two open-source datasets as well as our own primary dataset. Our dataset is named Asian Affective and Emotional State (A2ES) dataset which comes with ECG and PPG signals along with DEM-labelled emotions. Only ECG data are considered in this research. The raw signals are recorded from 47 participants of various Asian backgrounds with 25 samples each. The ECG is collected using KardiaMobile by AliveCor. The stimuli used to elicit the emotions are a collection of videos targeting different emotions.

AMIGOS and DREAMER are open-sourced datasets available for research purposes. AMIGOS dataset [13] consists of physiological signals inclusive of ECG, EEG and GSR recorded from 40 participants with 16 samples each. The ECG signals are recorded using a mobile ECG device called Shimmer. The participants labelled their emotions according to DEM and ADM. Around 51 to 150 s of videos are used as stimuli for emotional elicitation.

DREAMER [14] is a popular physiological-based affective dataset. The dataset holds ECG signals, EEG signals as well as emotion class labels in the format of ADM scales. The size of the dataset is 23 by 18 samples, and the ECGs are recorded using Shimmer as well. The stimuli used for emotion elicitation are 65 to 393 s film clips. Table 6 summarizes the details of the dataset used in this study.

Due to technical errors during data collection not all the data from these datasets can be used as the ECGs recorded suffer a loss of signal or have poor signal quality. These contribute to ineffective feature extraction on some ECG signals in the A2ES and AMIGOS dataset. The actual data used from A2ES, AMIGOS, and DREAMER are 1163 out of 1175, 1258 out of 1280, and 828. The same goes for the augmented samples where the feature extraction had trouble in processing some of the badly augmented ECGs. After augmentation, the size of the sample data is expanded to 8068 out of 8225

**Table 6** Dataset details

Dataset	Stimuli	Participants x Samples	Physiological modalities	ECG Device	Emotional annotation	Emotional model
A2ES	60–300 s videos	47 × 25	ECG, PPG	AliveCor, Kardia	Happy, Sad, Anger, Fear, Disgust, Surprise & Neutral	DEM
AMIGOS [13]	51–150 s videos	40 × 16	ECG, EEG, GSR	Shimmer	Happy, Sad, Anger, Fear, Disgust, Surprise & Neutral, Valence, Arousal & Dominance	DEM & ADM
DREAMER [14]	65–393 s film clips	23 × 18	ECG, EEG	Shimmer	Valence, Arousal & Dominance	ADM

for A2ES, 8806 out of 8960 for AMIGOS, and 5796 for DREAMER.

The distribution of class labels after augmentation in A2ES, AMIGOS, and DREAMER datasets is shown in Fig. 3. For BEM, the class labels are either positive or negative whereas the negative emotions are emotions which contribute towards stress. For ADM, scales of high, neutral, and low are considered for both valence and arousal. The class labels for DEM in this study are happy, sad, anger, fear, disgust, surprise and neutral.

Out of 8086 samples, 58% are labelled positive and 42% are labelled negative in A2ES BEM. For ADM-valence, 42% are labelled as low, 32% as high labels the remaining as neutral. For ADM-arousal, slightly above half of the samples are labelled high while the other half are divided almost equally between neutral (26%) and low (23%). Lastly, for DEM, the most sampled data are neutral at 26% and followed by happy at 20%. The least sampled data is anger with 657 signals out of 8086 (8%).

From the AMIGOS dataset, the distribution ratio of the class sample for BEM is 31:19 for positive and negative respectively. For ADM-valence, 41% of the data is labelled as high while low and neutral are divided approximately equal. The same goes for ADM-arousal, 43% of the data are labelled high while neutral is only 19% the remaining are low. The class distribution for DEM is the worst where the data are not evenly distributed across the 7 classes. The majority of 2648 out of 8068 signals are labelled as neutral. The smallest portion is sadness with only 212 sample data. The imbalance between emotion class labels in this dataset is huge.

Finally, from the DREAMER dataset, the distribution class sample for BEM is 61:39 for positive and negative respectively. For ADM-valence, the distribution between high and low are equal with 39% each while the rest of

the 22% are neutral. For ADM-arousal, 44% of the samples are labelled as high, 29% neutral, and 27% low. Since the DREAMER dataset does not come with DEM labels, no pie chart is presented for the DEM.

## 4.2 Hyperparameters

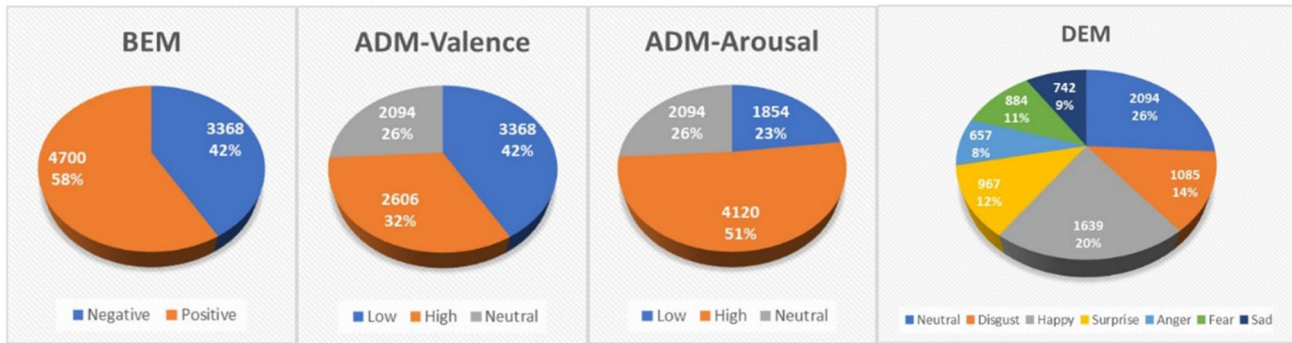
Since the performance of machine learning models is dependent on their hyperparameter settings, tuning them is necessary for the best results. Table 7 summarizes the hyperparameters settings with the number of variations being explored for every classifier.

## 4.3 Evaluation Metrics

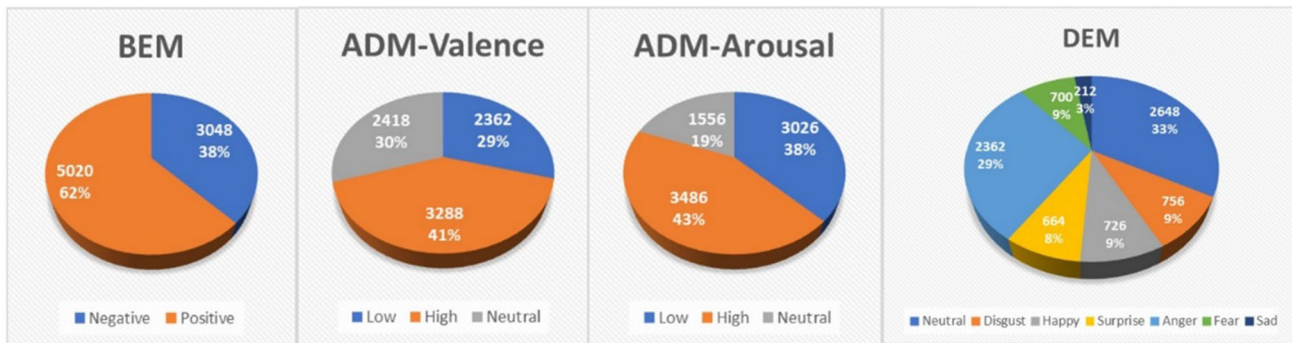
The most common evaluation metric to assess an ERS performance is using accuracy. It can be defined as the ratio of accurately classified data items to the total number of observations. Equation (6) shows the accuracy is calculated by dividing the summation of predicted true for positive (TP) and negative (TN) labelled data with the total data. The total data is calculated by summing up the TP, TN, false positive (FP), and false negative (FN).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

## A2ES Dataset



## AMIGOS Dataset



## DREAMER Dataset

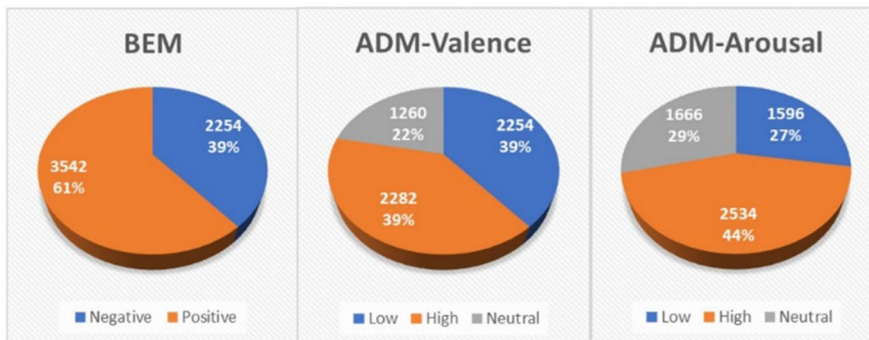


Fig. 3 Distribution of the datasets' emotional class labels

## 5 Results and Discussion

### 5.1 A2ES Dataset

The results obtained for BEM, ADM and DEM are presented in Table 8. For all emotional models, the testing accuracy of which augmentation takes place shows a significant improvement especially the one with the KNN classifier.

Based on the observation, the best classifier before data augmentation for BEM are RF with standard scaler and MLP with minmax scaler at 61%. The testing accuracy of the rest of the classifiers is not as good with the range of 48–56%. After

data augmentation, a huge leap in percentage accuracy is seen across all classifiers. The best one is KNN using standard scaler with 94% testing accuracy. Using the minmax scaler, KNN achieved 1% lesser in comparison to the previous one. More than a 40% increment in testing accuracy is observed for KNN with the introduction of augmented features. The second-best performing classifier is SVM with 81% testing accuracy, followed by MLP, RF, and DT. The later classifiers only manage to get around 64–79%.

Next, the testing accuracy recorded based on ADM-Valence before augmentation is in the range of 38–45%. KNN and SVM using standard scaler performed the best with 45%

**Table 7** Classifiers with the hyperparameter’s value for grid search

Classifier	Hyperparameter	Values	Number of variations
KNN	n_neighbors	3,5,7, sqrt(y), sqrt(y)/2	5
	Weights	Distance	1
	Metric	Hamming, Manhattan	2
Number of possible combinations			10
SVM	Kernel	RBF	1
	C	1, 10, 100	3
	Gamma	0.01, 0.1, 1	3
Number of possible combinations			9
DT	Criterion	gini, entropy	2
	max_depth	range (8,10)	2
	min_samples_split	range (5,7)	2
	min_samples_leaf	range (1,3)	2
Number of possible combinations			16
RF	n_estimator	200, 500	2
	max_features	auto	1
	max_depth	8	1
	criterion	gini, entropy	2
Number of possible combinations			4
MLP	hidden_layer_size	(200,200,200,)	1
	Activation	relu, tanh	2
	Solver	adam, lbfgs	2
	Alpha	0.05	1
	Learning_rate	constant	1
Number of possible combinations			4

testing accuracy. The worst performing algorithms are DT and MLP with standard scaler which recorded merely 38% testing accuracy. With the introduction of data augmentation, the accuracy of KNN increases by twofold. In classifying the high, neutral, and low classes of valence, KNN manages to achieve 91% and 90% testing accuracy using minmax and standard scaler respectively. For SVM and MLP, the algorithms did not perform as well as KNN but still manage to obtain 73% and 72% followed by RF and DT. However, MLP using the minmax scaler performed the worst of all as only a 9% increment is observed in the testing accuracy.

The next part shows the performance comparison recorded for classifying the ADM-arousal scale. Before augmentation, the testing accuracy achieved for all classifiers is within the range of 42–54%. The best classifier is RF for both the standard scaler and minmax scaler while the underperformed algorithm is DT with the standard scaler. After augmentation, KNN manage to get 91% of testing accuracy in classifying the features that are scaled with the standard and minmax scaler. The rest of the classifiers did manage to show some improvements with the introduction of augmentation which SVM, MLP, RF, and DT in orderly decreasing performance.

Finally, the performance of classifying DEM before data augmentation is very poor. Since the class labels are imbalanced, the machine learning algorithms suffered to recognize and generalize each distinctive emotion. Additionally, DEM has 7 classes which increases the complexity of the classification process. The testing accuracies are ranging within ~ 20% only which is considered very poor. At 27%, SVM and MLP with standard scaler reported the highest achievable accuracy. After data augmentation, the DT algorithm shows the slightest increase in performance. However, KNN, SVM, RF, and MLP manage to achieve more than twofold performance improvement. The best is KNN with 88% testing accuracy when paired with standard and minmax scalars. In this case, KNN has the largest gap in terms of performance gain compared to other classifiers. Around 65% increment is observed. Here, it shows that data augmentation is necessary to increase the count of small class samples and help the machine learning algorithm to improve the overall performance. SVM using minmax scaler manage to get 64% testing accuracy which is the second-best performance reported. The

**Table 8** Results from A2ES Dataset for BEM, ADM, and DEM

		Testing accuracy (%)			
		Before augmentation		After augmentation	
		Standard scaler	MinMax scaler	Standard scaler	MinMax scaler
BEM	KNN	51	48	94	93
	SVM	55	53	81	81
	DT	55	55	66	67
	RF	61	55	74	74
	MLP	56	61	79	64
ADM-V	KNN	45	42	90	91
	SVM	45	43	73	73
	DT	38	40	58	58
	RF	43	43	65	65
	MLP	38	40	72	49
ADM-A	KNN	51	51	91	91
	SVM	51	49	74	73
	DT	42	44	60	61
	RF	54	54	66	66
	MLP	52	51	71	58
DEM	KNN	24	22	88	88
	SVM	27	24	52	64
	DT	24	22	38	38
	RF	25	24	52	52
	MLP	27	20	62	51

worst classifier reported is DT where for both scalers the outcomes are around 38%. For MLP and RF, the observation reported ranged between KNN and DT.

Based on the results from BEM, ADM and DEM, the implementation of data augmentation that increases data size improves classifiers' performance and changes the rank of the classifier with KNN reported as the best for all emotional models.

## 5.2 AMIGOS Dataset

In the AMIGOS dataset, the results obtained for BEM, ADM and DEM are presented in Table 9. Same as the previous dataset, testing accuracy is presented. Overall, the effects of augmentation are significant, and the best-performing classifier after augmentation is also KNN. As a comparison, result from AMIGOS' original work [13] is also presented. However, only results of ADM are available.

Firstly, the results for positive and negative emotional classification are observed. Before data augmentation, the RF classifier performed the best with over 70% testing accuracy for both standard scalar and minmax scalar. Besides, the classification using SVM with standard scalar also manage to gain more than 70%. But the SVM with minmax scaler

only manage to achieve 69%. For the other three classifiers, the classification performance is within the 60% range. After data augmentation, the performance accuracy for all classifiers increased. Same as the A2ES dataset, the KNN algorithm performed best by attaining around 95% testing accuracy for both scalers. MLP and SVM with minmax scaler also show quite a significant leap with more than 80% testing accuracy after augmentation. The rest of the classifiers manage to recognize BEM with lesser accuracy but no less than 70%.

The part is the results achieved when the model is trained and tested using the ADM valence scale. For all five classifiers, the testing accuracy does not exceed 60% before data augmentation. RF on both scalers reported 55% testing accuracy. SVM also manage to get 55% testing accuracy using the standard scaler, but not for the minmax scaler. This accuracy is similar to the original study in [13]. The maximum accuracy reported is through KNN with a standard scaler which is 56%. MLP scores the lowest with only 52% on both scalers. After augmentation, the accuracy of all classifiers improved. The results for DT improved the least where the accuracy for both the standard scalar and the minmax scalar is still below 60%. The accuracy for SVM, RF and MLP is in the range of 60–80%. Meanwhile, the best classifier is KNN with



**Table 9** Results from AMIGOS Dataset for BEM, ADM, and DEM

		Testing Accuracy (%)			
		Before augmentation		After augmentation	
		Standard scaler	MinMax scaler	standard scaler	MinMax scaler
BEM	KNN	67	66	95	95
	SVM	71	69	78	84
	DT	67	61	74	73
	RF	72	71	77	76
	MLP	69	65	86	83
ADM-V	KNN	56	53	93	94
	SVM	55	53	65	73
	DT	54	52	58	58
	RF	55	55	73	74
	MLP	52	52	80	76
[13]	SVM	55			
ADM-A	KNN	56	57	93	93
	SVM	54	53	68	74
	DT	52	52	73	74
	RF	60	60	82	82
	MLP	52	53	79	77
[13]	SVM	55			
DEM	KNN	52	53	92	92
	SVM	44	44	55	68
	DT	33	33	66	67
	RF	46	47	77	77
	MLP	39	40	75	70

93% and 94% accuracy with respect to standard scalar and minmax scalar.

The performance of ADM arousal classification is discussed next. Before augmentation, the testing accuracy for KNN, SVM, DT, and MLP is within the range of 50% while for RF, the testing accuracy is 60% using both standard scalar and minmax scalar. The accuracy obtained in [13] is also within this range. A significant accuracy improvement is observed when the multi-filter augmentation is utilised to increase the data size. For SVM, DT, and MLP, the results are improved to between 68 and 79% for testing while for RF, the accuracy is over 80% on both scalars. The leading algorithm is KNN with 93% accuracy on standard and minmax scalars.

The last part is the results for DEM where seven basic emotions are classified together. Here, the overall recognition performance before augmentation is comparably better than A2ES. The worst testing accuracy achieved is 33% by DT. The highest reported testing accuracy is by KNN with above 50% accuracy using both scalars. Augmentation helps to improve classifiers’ performance in classifying unbalanced and multiclass datasets. This is also observed before in the

A2ES dataset. The least improved classifier is SVM which reported only 55% and 68% accuracy. For DT, the testing accuracy reported is 66% and 67%. For RF and MLP, around 70–77% accuracy is reported. Alas, KNN is once again the best-performing classifier with 92% accuracy for standard scalar and minmax scalar.

### 5.3 DREAMER Dataset

In the DREAMER dataset, only BEM and ADM results are available. Therefore, no classification performance for DEM is reported for this dataset. The testing accuracies for each classifier for BEM and ADM are shown in Table 10. As a comparison, ADM result from [14] is also presented.

For BEM before augmentation, DT is the worst classifier with 65% and 69% testing accuracy for the standard scaler and minmax scaler. MLP with a standard scaler also gives out low accuracy at 67% while with a minmax scaler the result is better at 70% which is similar to KNN with a standard scaler and SVM with a minmax scaler. The highest classification accuracy is achieved through RF where the recorded results show an accuracy of 75% for the standard scaler and minmax

**Table 10** Results from DREAMER Dataset for BEM and ADM

		Testing accuracy (%)			
		Before augmentation		After augmentation	
		Standard scaler	MinMax scaler	Standard scaler	MinMax scaler
BEM	KNN	70	72	99	99
	SVM	72	70	84	93
	DT	65	69	86	86
	RF	75	75	92	92
	MLP	67	70	94	92
ADM-V	KNN	67	64	98	98
	SVM	70	64	82	90
	DT	61	61	71	71
	RF	69	69	90	90
	MLP	67	67	91	92
[14] ADM-A	SVM	62			
	KNN	76	77	99	99
	SVM	67	68	76	90
	DT	52	53	72	72
	RF	71	74	94	94
[14]	MLP	67	60	87	89
	SVM	62			

scaler. After applying the proposed augmentation method, a staggering 99% testing accuracy is recorded by KNN. MLP and RF also passed the 90% accuracy for both scalers. For SVM with a minmax scaler, the accuracy is reported to be 93% while the one with a standard scaler is 84% which is the lowest among all. Both scalers for DT reported 86% for testing accuracy.

The results for the classification of valence show that before augmentation the highest testing accuracy achieved is 70%. This is achieved using SVM with a standard scaler. The lowest accuracy reported is 61% by DT with standard scaler and minmax scaler. Using the minmax scaler, KNN and SVM achieved 64%, while the rest are reported above that, but no more than 70% including the original study from [14]. With a 10% increase, DT is the least-performing classifier after data augmentation. Next, is SVM with a standard scaler that managed to gain a 12% increase in accuracy while the rest of the classifiers manage to achieve 90% and above for testing accuracy. SVM with minmax scaler and RF with both scalers achieved 90%. MLP on the other hand, manage to achieve 91% and 92%. Ultimately, KNN reported 98% for both standard and minmax scaler.

Lastly, the observation results for the three-scale arousal are discussed. Before augmentation, the best testing accuracies are 77% and 76% from KNN. The second-best algorithm is RF with 74% and 71%. For SVM and MLP, the testing accuracy observed is within 60–70%, including the results

from [14]. The lowest accuracy achieved is 52% and 53% by DT. Post augmentation, the best classifier for classifying arousal is KNN. Both scalers give 99% testing accuracy, which is an almost perfect performance. With slightly lower accuracy, RF using both scaler and SVM for minmax scaler reported 94% and 90% testing accuracy respectively. However, SVM with a standard scaler did not perform as well, where the reported accuracy is 76%. MLP with a minmax scaler shows a 2% higher accuracy compared to the one with a standard scaler. Although DT only achieved 72% testing accuracy, the improvement gained from augmentation is more than 20%.

## 5.4 Statistical Analysis

Non-parametric statistical analysis is applied to examine the significant difference in the overall results recorded. The Wilcoxon signed ranks test [55] and Friedman test [56] with Holm's post hoc test [57] were conducted as suggested by [58, 59]. The Wilcoxon signed ranks test is done to find the significant difference in the results before and after augmentation as well as between the Standard scaler and MinMax scaler. Next, the Friedman statistical test is applied to find the best classifier before augmentation and the best classifier after augmentation. Lastly, the Holm post hoc test is done if the Friedman test found a significant difference between the

**Table 11** Average ranking of Friedman test for classifiers comparison before and after augmentation

Algorithm	Ranking of classifiers before augmentation	Ranking of classifiers after augmentation
KNN	2.7045	1
SVM	2.6818	3.2045
DT	4.5909	4.7045
RF	1.5455	3.0227
MLP	3.4773	3.0682

classifiers. The analysis is done using Knowledge Extraction based on Evolutionary Learning (KEEL) software [60].

#### 5.4.1 Comparison of Before and After Augmentation

A Wilcoxon Signed-Rank test is applied to find the correlation between the results collected before and after augmentation regardless of the classifier or the scaler used.

The value of  $W$  obtained is 0 and the distribution is approximately normal. The  $z$ -value is -9.1035 and the  $p$ -value is  $p < 0.00001$ . The null hypothesis is rejected as the result is significantly different with a significant level of 5%. Thus, the augmentation technique proposed has significantly improved the performance of the ERS.

#### 5.4.2 Comparison of Classifiers Before Augmentation

A multiple classifiers comparison using the Friedman test is conducted to find the significant difference and the ranking of the best classifier before data augmentation is implemented. Table 11 shows the average ranking of the algorithms where the best-ranked classifier is RF. SVM, KNN, and MLP are ranked second, third, and fourth while DT is the last one. Friedman's statistic considering reduction performance (distributed according to chi-square with 4 degrees of freedom) is 44.5545. The  $p$ -value computed by the Friedman test is  $p < 0.00001$ . Therefore, the null hypothesis that stated that all the classifiers are on par with each other is rejected. The result is significantly different at  $p < 0.05$ .

Table 12 shows the algorithms compared using Holm's post hoc test. The  $z$ -value,  $p$ -value, and Holm value are tabulated. Holm's procedure rejects those hypotheses that have an unadjusted  $p$ -value  $\leq 0.008333$ . This means that for i1-i6 the pairs are statistically on par with each other while for i7-i10 there is a significant difference between the classifiers' performance.

#### 5.4.3 Comparison of Classifiers After Augmentation

Multiple comparisons using the Friedman test are conducted to find the significant difference and the ranking of the best classifier after data augmentation implementation. Table 11 shows the average ranking of the algorithms where the best-reported classifier is KNN. Followed by RF, MLP and SVM while DT is the last one. Friedman's statistic considering reduction performance (distributed according to chi-square with 4 degrees of freedom) is 61.1818. The  $p$ -value computed by the Friedman test is  $< 0.00001$ . Thus, the null hypothesis is rejected as the result is significantly different at  $p < 0.05$ .

Table 12 shows the algorithms compared using Holm's post hoc test. The  $z$ -value,  $p$ -value, and Holm value are tabulated. Holm's procedure rejects those hypotheses that have an unadjusted  $p$ -value  $\leq 0.016667$ . This means that from i1-i3 there is no significant difference between the classifier being compared while from i4-i10 there is a significant difference. This shows that SVM, RF, and MLP are statistically on par with each other. Most importantly classification of augmented data using KNN is significantly better than other classifiers.

#### 5.4.4 Comparison of Scalers

A Wilcoxon Signed-Rank test is applied to find the correlation between the results collected using a standard scaler and minmax scaler regardless of the classifier and augmentation.

The value of  $W$  is 1597 and the distribution is approximately normal. The  $z$ -value is  $-0.8384$  and the  $p$ -value is 0.4009. Thus, the null hypothesis is retained as the result is not significantly different at  $p < 0.05$ . Thus, the performance of ERS is not determined by the type of scaler used in this study.

## 6 Conclusion and Future Works

This study is dedicated to tackling the problem of insufficient sample data in designing an ECG-based ERS that causes low accuracy. The application of multiple filters for data augmentation is proposed here. The augmentation technique proposed can increase the number of ECG data and increases the training samples available for the classifiers to learn on. This method is simple, and the selected filters had been proven to be good for ECG signal filtering. ERS models are built to evaluate the effectiveness of the proposed augmentation method. The results from three selected datasets, A2ES, AMIGOS, and DREAMER show that the classification accuracy increase after data augmentation is introduced. This is validated by the Wilcoxon-Signed statistical test. The KNN classifier benefits the most from the introduced augmentation as observed from the statistical test conducted.

**Table 12** Holm post hoc test *p*-values table for classifiers comparison before and after augmentation

<i>i</i>	Algorithms	$z = (R_0 - R_i)/SE$	<i>p</i>	Holm
Before augmentation				
10	DT vs. RF	6.388199	0	0.005
9	RF vs. MLP	4.052216	0.000051	0.005556
8	SVM vs. DT	4.004543	0.000062	0.00625
7	KNN vs. DT	3.95687	0.000076	0.007143
6	KNN vs. RF	2.43133	0.015044	0.008333
5	SVM vs. RF	2.383656	0.017142	0.01
4	DT vs. MLP	2.335983	0.019492	0.0125
3	SVM vs. MLP	1.66856	0.095205	0.016667
2	KNN vs. MLP	1.620886	0.105042	0.025
1	KNN vs. SVM	0.047673	0.961977	0.05
After augmentation				
10	KNN vs. DT	7.77072	0	0.005
9	KNN vs. SVM	4.624294	0.000004	0.005556
8	KNN vs. MLP	4.338255	0.000014	0.00625
7	KNN vs. RF	4.242909	0.000022	0.007143
6	DT vs. RF	3.527812	0.000419	0.008333
5	DT vs. MLP	3.432465	0.000598	0.01
4	SVM vs. DT	3.146427	0.001653	0.0125
3	SVM vs. RF	0.381385	0.702918	0.016667
2	SVM vs. MLP	0.286039	0.774848	0.025
1	RF vs. MLP	0.095346	0.92404	0.05

For the scaler used, the study shows that either standard or minmax can be used without any significant effect on the accuracy performance of the ECG-based ERS.

The suggestions for future works include the use of the same augmentation method to balance the class labels in the dataset. This is to ensure that the classification bias is reduced minimally due to the imbalance class. Furthermore, extending the data augmentation via a multi-filtering method to other modalities such as EEG, PPG, etc. This is to observe the effectiveness of multi-filtering augmentation. Finally, investigate the number of filters concerning the original data size and the best filter combination for augmentation.

**Author Contributions** MAH, NAAA, and AAA contributed to the experimental design and analysis of the results. MAH wrote the first draft of the manuscript. NAAA and AAA contributed to the manuscript's revision and proofreading.

**Funding** This research was funded by the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme, Grant No. FRGS/1/2019/ICT02/MMU/02/ 15, which is awarded to Multimedia University.

## Declarations

**Conflict of interest** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Ethics Approval** The Multimedia University Research Ethics Committee (REC) has approved the data collection of A2ES (Approval No.: EA0282021).

## References

- Shiomi, M.; Zheng, X.; Minato, T.; Ishiguro, H.: Implementation and evaluation of a grip behavior model to express emotions for an android robot. *Front. Robot. AI* 8(October), 1–10 (2021). <https://doi.org/10.3389/frobt.2021.755150>
- Picard, R. W.: *Affective Computing*, (1995)
- Strauss, M., et al.: "Affective computing: a review, Vol. 3784, p. 699–706. Springer, Berlin (2005) <https://doi.org/10.1007/11573548>
- Braun, M.; Chadowitz, R.; Alt, F.: "User experience of driver state visualizations: a look at demographics and personalities," In *Human-Computer Interaction – INTERACT 2019. INTERACT 2019. Lecture Notes in Computer Science*, 2019, vol. 11749, pp. 158–176, doi: <https://doi.org/10.1007/978-3-030-29390-1>.

5. Jaihar, J.; Lingayat N.; Vijaybhai, P.S.; Venkatesh, G.; Upla, K.P.: “Smart home automation using machine learning algorithms”, In 2020 International Conference for Emerging Technology INCET 2020, 20–23, (2020) <https://doi.org/10.1109/INCET49848.2020.9154007>
6. Hovsepian, K.; Al’absi, M.; Ertin, E.; Kamarck, T.; Nakajima, M.; Kumar, S.: “CStress: towards a gold standard for continuous stress assessment in the mobile environment,” (2015), doi: <https://doi.org/10.1145/2750858.2807526>
7. Jiang, Z.; Lu, L.; Huang, X.; Tan, C.: “Design of wearable home health care system with emotion recognition function,” (2011), doi: <https://doi.org/10.1109/ICECENG.2011.6057832>
8. Tivatansakul, S.; Ohkura, M.: “Healthcare system focusing on emotional aspects using augmented reality: Implementation of breathing control application in relaxation service,” Proc. - 2013 Int. Conf. Biometrics Kansei Eng. ICBAKE 2013, no. July 2013, pp. 218–222, (2013), doi: <https://doi.org/10.1109/ICBAKE.2013.43>
9. Hasnul, M.A.; Aziz, N.A.A.; Alelyani, S.; Mohana, M.; Aziz, A.A.: Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review. *Sensors* **21**(15), 5015 (2021). <https://doi.org/10.3390/s21155015>
10. Churamani, N.; Barros, P.; Gunes, H.; Wermter, S.: Affect-driven learning of robot behaviour for collaborative human-robot interactions. *Front. Robot. AI* **9**(February), 1–19 (2022). <https://doi.org/10.3389/frobt.2022.717193>
11. Pantic, M.; Caridakis, G.; André, E.; Kim, J.; Karpouzis, K.; Kollias, S.: Multimodal emotion recognition from low-level cues. *Cogn. Technol.* (2011). [https://doi.org/10.1007/978-3-642-15184-2\\_8](https://doi.org/10.1007/978-3-642-15184-2_8)
12. Hatamian, F.N.; Ravikumar, N.; Vesal, S.; Kemeth, F.P.; Struck, M.; Maier, A.: “The effect of data augmentation on classification of atrial fibrillation in short single-lead ecg signals using deep neural networks,” In ICASSP, IEEE International conference on acoustics, speech and signal processing - proceedings, (2020), vol. 2020-May, doi: <https://doi.org/10.1109/ICASSP40776.2020.9053800>
13. Miranda Correa, J.A.; Abadi, M.K.; Sebe, N.; Patras, I.: AMIGOS: a dataset for affect, personality and mood research on individuals and groups. *IEEE Trans. Affect. Comput.* (2018). <https://doi.org/10.1109/TAFFC.2018.2884461>
14. Katsigiannis, S.; Ramzan, N.: DREAMER: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE J. Biomed. Heal. Inf.* **22**(1), 98–107 (2018). <https://doi.org/10.1109/JBHI.2017.2688239>
15. Zong, C.; Chetouani, M.: “Hilbert-Huang transform based physiological signals analysis for emotion recognition,” In 2009 IEEE International symposium on signal processing and information technology (ISSPIT), (2009), pp. 334–339, doi: <https://doi.org/10.1109/ISSPIT.2009.5407547>
16. Bong, S.Z.; Murugappan, M.; Yaacob, S.: Analysis of electrocardiogram (ECG) signals for human emotional stress classification, p. 198–205. Springer, Berlin (2012) [https://doi.org/10.1007/978-3-642-35197-6\\_22](https://doi.org/10.1007/978-3-642-35197-6_22)
17. Xiefeng, C.; Wang, Y.; Dai, S.; Zhao, P.; Liu, Q.: Heart sound signals can be used for emotion recognition. *Sci. Rep.* **9**(1), 1–11 (2019). <https://doi.org/10.1038/s41598-019-42826-2>
18. Liu, X., et al.: Human emotion classification based on multiple physiological signals by wearable system. *Technol Health Care* (2018). <https://doi.org/10.3233/THC-174747>
19. Sarkar, P.; Etemad, A.: Self-supervised ECG representation learning for emotion recognition. *IEEE Trans. Affect. Comput.* (2020). <https://doi.org/10.1109/TAFFC.2020.3014842>
20. Siddharth, S.; Jung, T.-P.; Sejnowski, T.: Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing. *IEEE Trans. Affect. Comput.* **1**, 99 (2019)
21. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M.: A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput.* **3**(1), 42–55 (2012). <https://doi.org/10.1109/TAFFC.2011.25>
22. Santamaria-Granados, L.; Munoz-Organero, M.; Ramirez-Gonzalez, G.; Abdulhay, E.; Arunkumar, N.: Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS). *IEEE Access* **7**, 57–67 (2019). <https://doi.org/10.1109/ACCESS.2018.2883213>
23. Subramanian, R.; Wache, J.; Abadi, M.K.; Vieriu, R.L.; Winkler, S.; Sebe, N.: Ascertain: emotion and personality recognition using commercial sensors. *IEEE Trans. Affect. Comput.* (2018). <https://doi.org/10.1109/TAFFC.2016.2625250>
24. Chen, G.; Zhu, Y.; Yang, Z.; Hong, Z.: Emotionalgan: generating ECG to enhance emotion state classification, (2019), doi: <https://doi.org/10.1145/3349341.3349422>
25. Abadi, M.K.; Subramanian, R.; Kia, S.M.; Avesani, P.; Patras, I.; Sebe, N.: DECAF: MEG-based multimodal database for decoding affective physiological responses. *IEEE Trans Affect Comput* **6**(3), 209–222 (2015). <https://doi.org/10.1109/TAFFC.2015.2392932>
26. Goodfellow, I., et al.: Generative adversarial networks. *Commun. ACM* **63**(11), 139–144 (2020). <https://doi.org/10.1145/3422622>
27. Pei, Y., et al.: Data augmentation: using channel-level recombination to improve classification performance for motor imagery EEG. *Front. Hum. Neurosci.* (2021). <https://doi.org/10.3389/fnhum.2021.645952>
28. Bowles, C., et al.: GAN augmentation: augmenting training data using generative adversarial networks, (2018)
29. Shorten, C.; Khoshgoftaar, T.M.: “A survey on image data augmentation for deep learning. *J Big Data* (2019). <https://doi.org/10.1186/s40537-019-0197-0>
30. Luo, Y.: EEG data augmentation for emotion recognition using a conditional Wasserstein GAN. In Proceedings of the annual international conference of the IEEE engineering in medicine and biology society, EMBS, (2018), vol. 2018-July, doi: <https://doi.org/10.1109/EMBC.2018.8512865>
31. Chatziagapi, A., et al.: “Data augmentation using GANs for speech emotion recognition,” In proceedings of the annual conference of the international speech communication association, INTER-SPEECH, (2019), vol. 2019-Sept, doi: <https://doi.org/10.21437/Interspeech.2019-2561>
32. Sajjad, M.; Zahir, S.; Ullah, A.; Akhtar, Z.; Muhammad, K.: Human behavior understanding in big multimedia data using CNN based facial expression recognition. *Mob. Networks Appl.* **25**(4), 1611–1621 (2020). <https://doi.org/10.1007/s11036-019-01366-9>
33. Kartali, A.; Roglic, M.; Barjaktarovic, M.; Duric-Jovicic, M.; Jankovic, M.M.: “Real-time algorithms for facial emotion recognition: a comparison of different approaches”, In 2018 14th Symp Neural Networks Appl. NEUREL, (2018) 2018–2021, <https://doi.org/10.1109/NEUREL.2018.8587011>
34. Cao, P., et al.: A novel data augmentation method to enhance deep neural networks for detection of atrial fibrillation. *Biomed. Signal Process. Control* **56**, 101675 (2020). <https://doi.org/10.1016/j.bspc.2019.101675>
35. Nonaka, N.; Seita, J.: RandECG: data augmentation for deep neural network based ECG classification, (2021)
36. Iwanaid, B.K.; Uchida, S.: An empirical survey of data augmentation for time series classification with neural networks, (2021), doi: <https://doi.org/10.1371/journal.pone.0254841>
37. Makowski, D. et al.: NeuroKit2: a python toolbox for neurophysiological signal processing. *Behav. Res. Methods* **53**(4), 1689–1696 (2021). <https://doi.org/10.3758/s13428-020-01516-y>
38. Carreiras, C.; Alves, A.P.; Lourenço, A.; Canento, F.; Silva, H.; Fred, A.: BioSPPy: biosignal processing in python. Accessed on **3**(28), 2018 (2015)



39. Gustafsson, F.: Determining the initial states in forward-backward filtering. *IEEE Trans. Signal Process* **44**(4), 1996 (1996). <https://doi.org/10.1109/78.492552>
40. Pan, J.; Tompkins, W.J.: A real-time QRS detection algorithm. *IEEE Trans. Biomed. Eng.* **BME-32**(3), 230–236 (1985). <https://doi.org/10.1109/TBME.1985.325532>
41. Hamilton, P.: Open source ECG analysis. *Comp Cardiol* (2002). <https://doi.org/10.1109/cic.2002.1166717>
42. Elgendi, M.; Jonkman, M.; Deboer, F.: Frequency bands effects on QRS detection, (2010), doi: <https://doi.org/10.5220/0002742704280431>
43. Engelse, W.A.H.; Zeelenberg, C.: Single scan algorithm for QRS-detection and feature extraction, (1979)
44. Lourenço, A.; Silva, H.; Leite, P.; Lourenço, R.; Fred, A.: Real time electrocardiogram segmentation for finger based ECG biometrics. *Signals* (2012). <https://doi.org/10.5220/0003777300490054>
45. Xia, L.; Malik, A.S.; Subhani, A.R.: A physiological signal-based method for early mental-stress detection. *Biomed. Signal Process. Control* **46**, 18–32 (2018). <https://doi.org/10.1016/j.bspc.2018.06.004>
46. Kim, H.G.; Cheon, E.J.; Bai, D.S.; Lee, Y.H.; Koo, B.H.: Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry Investig.* **15**(3), 235–245 (2018). <https://doi.org/10.30773/pi.2017.08.17>
47. Sztajzel, J.: Heart rate variability: a noninvasive electrocardiographic method to measure the autonomic nervous system. *Swiss Medical Weekly* **134**(35–36), 514–522 (2004)
48. Michael, S.; Graham, K.S.; Oam, G.M.D.: Cardiac autonomic responses during exercise and post-exercise recovery using heart rate variability and systolic time intervals—a review. *Front Physiol* (2017). <https://doi.org/10.3389/fphys.2017.00301>
49. Rainville, P.; Bechara, A.; Naqvi, N.; Damasio, A.R.: Basic emotions are associated with distinct patterns of cardiorespiratory activity. *Int. J. Psychophysiol.* **61**(1), 5–18 (2006). <https://doi.org/10.1016/j.ijpsycho.2005.10.024>
50. Ferdinando, H.; Seppanen, T.; Alasaarela, E.: Comparing features from ECG pattern and HRV analysis for emotion recognition system, (2016) <https://doi.org/10.1109/CIBCB.2016.7758108>.
51. Hayati Rezvan, P.; Lee, K.J.; Simpson, J.A.: “The rise of multiple imputation: a review of the reporting and implementation of the method in medical research Data collection, quality, and reporting. *BMC Med. Res. Methodol.* (2015). <https://doi.org/10.1186/s12874-015-0022-1>
52. Donders, A.R.T.; van der Heijden, G.J.M.G.; Stijnen, T.; Moons, K.G.M.: Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* **59**(10), 1087 (2006). <https://doi.org/10.1016/j.jclinepi.2006.01.014>
53. Ahsan, M.M.; Mahmud, M.A.P.; Saha, P.K.; Gupta, K.D.; Siddique, Z.: Effect of data scaling methods on machine learning algorithms and model performance. *Technologies* **9**(3), 52 (2021). <https://doi.org/10.3390/technologies9030052>
54. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
55. Wilcoxon, F.: Individual comparisons of grouped data by ranking methods. *J Econ Entomol* **39**, 1946 (1946). <https://doi.org/10.1093/jee/39.2.269>
56. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **32**(200), 675–701 (1937). <https://doi.org/10.1080/01621459.1937.10503522>
57. Holm, S.: A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**(2), 65–70 (1979)
58. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006)
59. Singh, P.K.; Sarkar, R.; Nasipuri, M.: Significance of non-parametric statistical tests for comparison of classifiers over multiple datasets. *Int. J. Comput. Sci. Math.* **7**(5), 410 (2016). <https://doi.org/10.1504/IJCSM.2016.080073>
60. Alcalá-Fdez, J., et al.: KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Comput.* **13**(3), 307–318 (2009). <https://doi.org/10.1007/s00500-008-0323-y>

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.