Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 7, No. 1, February, 2023

Evaluation of Machine Learning Algorithms for Emotions Recognition using Electrocardiogram

Chy Mohammed Tawsif Khan [1*], Nor Azlina Ab Aziz [1*], J. Emerson Raja [1], Sophan Wahyudi Bin Nawawi [2], Pushpa Rani [3]

[1] *Faculty of Engineering & Technology, Multimedia University, 75450, Melaka, Malaysia.*

[2] *Fakulti Kejuruteraan Elektrik Universiti Teknologi Malaysia Skudai, Malaysia.*

[3] *Professor & Head, Department of Computer Science, Mother Teresa Women's University, India.*

Abstract

In recent studies, researchers have focused on using various modalities to recognize emotions for different applications. A major challenge is identifying emotions correctly with only electrocardiograms (ECG) as the modality. The main objective is to reduce costs by using single-modality ECG signals to predict human emotional states. This paper presents an emotion recognition approach utilizing the heart rate variability features obtained from ECG with feature selection techniques (exhaustive feature selection (EFS) and Pearson's correlation) to train the classification models. Seven machine learning (ML) models: multi-layer perceptrons (MLP), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting Decision Tree (GBDT), Logistic Regression, Adaboost and Extra Tree classifier are used to classify emotional state. Two public datasets, DREAMER and SWELL are used for evaluation. The results show that no particular ML works best for all data. For DREAMER with EFS, the best models to predict valence, arousal, and dominance are Extra Tree (74.6%), MLP and DT (74.6%), and GBDT and DT (69.8%), respectively. Extra tree with Pearson's correlation are the best method for the ECG SWELL dataset and provide 100% accuracy. The usage of Extra tree classifier and feature selection technique contributes to the improvement of the model accuracy. Moreover, the Friedman test proved that ET is as good as other classification models for predicting human emotional state and ranks highest.

Keywords:

Electrocardiogram (ECG);
Emotion Recognition System;
Exhaustive Feature Selection;
Gradient Boosting Decision Tree (GBDT);
Machine Learning.

Article History:

Received:	07	April	2022
Revised:	13	September	2022
Accepted:	25	September	2022
Available online:	07	November	2022

1- Introduction

Emotion affects and controls human perceptions and actions. For example, stress is a condition of mental or emotional pressure. Stress can be induced by some traumatic event or thought, which in turn makes the person feel angry, frustrated, or nervous. A stressed driver may have poor driving performance, which could be dangerous for the driver and other road users. Therefore, building machines that are able to understand human emotion can improve the quality of life by enhancing the interaction between humans and machines, e.g., a driver assistance system that can recognize signs of stress and remind the driver of dangerous driving. Emotion recognition in affective computing can make the machines understand human emotions using data collected from the person interacting with the machines [1]. Humans' emotions are reflected in body expressions and physiological measurements [2]. However, most research on recognizing emotions is based on physical features, such as audio-visual data [3] and facial expression data [4]. On the other hand, there is growing interest in recognizing emotion using data from physiological responses [5–11].

* **CONTACT**: 1191402716@student.mmu.edu.my; azlina.aziz@mmu.edu.my

DOI: http://dx.doi.org/10.28991/ESJ-2023-07-01-011

The main reason that attracts researchers to use physiological data for identifying emotion is because these physiological signals are unconscious responses that cannot be controlled or faked. Moreover, it does not require any continuous camera monitoring, which raises the issue of individual privacy. Some researchers combined several modalities as input data; this was done to obtain reliable emotional recognition system (ERS) data [3]. Emotion recognition systems have a lot of prospective applications, spanning healthcare, entertainment, e-learning, marketing, human monitoring, and security. According to Nikolova et al. [12], there were three major applications of emotion recognition systems, specifically using ECG signals.

Initially, human emotions and behavioral activities were monitored and assessed, respectively, in a critical circumstance. For instance, a driver's performance can be studied through the emotion identification technique [13].

Secondly, for giving the proper treatment or medication to the drug-addicted patient, the psychological response may be considered a clinical application. Now a day's emotion identification process is also utilized as a stress reduction system leads to promotes relaxation. Three particular emotional activities were considered in the designed architecture, such as excitement, amusement and relaxation [14].

Finally, emotion recognition can be utilized for marketing and website optimization [15], where the system has to collect information about visitors' attention to advertisements. Which will help to supply appropriate content according to audience demography.

This study aims to use a single modality, specifically using only electrocardiogram (ECG) data for ERS. This is to reduce the cost of building the ERS so that it is suitable to be embedded in consumer applications such as smart home controllers to ensure occupants' comfort. ECG was chosen because there was relatively little research considering only ECG physiological signals. For example, in a research study, Koldijk et al. [16] used different forms of data to detect stress. They had also tested using three physiological signals and obtained only 64.0997%.

This work used two public emotion datasets, namely DREAMER and SWELL, to build the emotion recognition system. Both of these datasets contain ECG signals. The DREAMER dataset grouped the data into three emotional dimensions: valence, arousal, and dominance. Each of the dimensions is then grouped into high and low. Meanwhile, SWELL data is labelled as positive and negative only. There is a greater amount of data in the SWELL dataset than in DREAMER. The findings show that the Extra Tree classifier and Pearson's correlation feature selection give the best accuracy for the SWELL dataset, while for the DREAMER dataset, the best classifier depends on the emotional dimension. No supreme classifier that works best for all emotional dimensions of valence, arousal, and dominance can be identified. The usage of EFS contributes to better performance compared to classification without feature selection and Pearson's correlation. Overall, the results show that ECG can be used as the single modality for building an emotion recognition system. However, like many classification problems, the data size for training a model is essential to building a better model. This is proven by the high accuracy found for the SWELL dataset, which is much better than the DREAMER.

Choosing the proper classifier is essential to ensuring good emotion recognition performance. Linear classifiers [17] are popular because of their simplicity, interpretability, and speed. On the other hand, an experiment using EEG signals for the classification of five mental tasks shows that nonlinear classifiers are suitable for signal features and cognitive state classification [18]. Therefore, six machine learning classifiers are studied here: multi-layer perceptron (MLP), support vector machine (SVM), decision tree (DT), logistic regression (LR), and two Ensemble Learning models: the gradient boosting decision tree (GBDT) and the extra tree classifier. The MLP, DT, and GBDT are non-linear classifiers; the SVM contains both linear and non-linear classification; and the rest are linear classifiers. Ensemble Learning was used to improve the decision tree classifiers' performance. With the increase in the decision tree, there is usually a chance to become a vulnerable model to high-variance and might be overfitting. In this case. Ensemble learning was used with general rules to integrate regularization and overcome overfitting. This work also studies the usage of exhaustive feature selection (EFS) and Pearson's correlation-based feature selection in improving classification performance.

This section is followed by a discussion of related works in section 2. Next, section 3 contains the description of the methodology, including the data preprocessing of labeling, feature extraction, and selection, followed by the selected ML classifiers. In section 4, the results and discussion are presented. Finally, conclusions and future work are discussed in section 5.

2- Related Works

The demand for emotion recognition systems to enable more fluid human and machine interaction is growing along with technological advancement. One of the important aspects of building an emotion recognition system is the classification of emotions. According to Davidson et al. [19], the asymmetric behaviour of emotions is correlated to the dimension of arousal and valence. The positive or negative state is judged on valence, while the high and low excitation depends on arousal. Additionally, dominance is another dimension of emotion that represents whether a person is feeling in control or out of control. On the other hand, some researchers [20-22] adopt the discrete emotions classes of happy, sad, angry, surprised, etc. [20-22] while others [16, 23] adopt the simple two classes of emotions such as stress/not stress or positive/negative [16, 23].

Several successful research has been done on emotion recognition using audio/visual (voice, facial expression, etc.), physiological measurements (respiration, skin temperature, etc.), and relevant human activity (gesture, posture) [24]. However, physical modalities like voice, facial expression, gesture, and gait are vulnerable to fake emotions where a person can act in a certain way so that the machine perceives him/her to be in a particular state of feeling.

Katsigiannis and Ramzan [25] had provided a multimodality physiological signal dataset for emotion recognition; DREAMER that includes ECG signals. Their model performance accuracy is below 63%. The authors stated that combining ECG and EEG data provided better performance. Koldijk et al. (2014) [26] proposed a multimodal dataset known as SWELL-KW, which includes ECG data with the HRV extracted features. In another article Koldijk et al. (2018) [16] reported an emotion recognition model using multimodal with around 90% accuracy using the support vector machine (SVM) that detects a worker's mental stress level [16]. However, the multimodality prediction model is costly as it requires multiple sensors and devices.

Feature extraction, selections, and pre-processing techniques have been reported to contribute to the improvement of emotion recognition. Different researchers processed the data differently to extract the features. For example, Goshvarpour et al. [27], they tested three different feature extraction methods for feature extraction from ECG and galvanic skin responses (GSR). Various tools to generate the features have been proposed and available for the research community, such as the EEGLAB [28]. The EEGLAB extracts independent component analysis (ICA), time/frequency analysis, and event-related statistics from EEG, Magnetoencephalography (MEG), and other electrophysiological data.

Another well-known tool available is the Augsburg Biosignal Toolbox (AuBT) [29] which has the capability to extract emotionally related physiological features. Katsigiannis & Ramzan [25] used the AuBT tool to extract the HRV features from the ECG dataset. However, AuBT requires MATLAB to extract features. Therefore, it will be challenging to integrate this tool into real applications. The NeuroKit2 library developed by Makowski et al. [30] can be used to extract features from ECG signals. NeuroKit2 is a python-based library, which can easily be integrated into any python programming.

It is useful for pre-processing and extracting important features from various central and peripheral nervous system data, such as ECG, PPG, HRV, RSP, EDA, EMG, and EEG. The ECG features can be easily extracted by converting the signals into single heartbeats, i.e., heart rate variability (HRV), using NeuroKit2. These signals can easily be used in the selected model. The HRV features are strongly connected to the emotional state of the individual. For example, the heart rate (HR) value increases with anger or fear (arousal). Shu et al. [31] used HR data from a wearable smart bracelet to train the ML model for emotion detection. They used SelectKBest feature selection technique to select the best features according to k highest scores. They used Adaboost and GBDT classifiers with 70.7-84% performance for detecting three different emotions (happy, sad, and neutral). Bulagang et al. [32] trained three classifiers (KNN, SVM, and RF) using HR data from the Empatica E4 wristband, they reported SVM and KNN provided 80% accuracy while detecting emotions using HR.

Looking at the existing works, it could be seen that there is room further to enhance the performance of the emotion recognition model. A number of works mentioned feature engineering requirements to obtain the best features for training the model. The most common features of the ECG signal used are HRV features, which contain useful data about the physiological state of emotion of an individual. Therefore, in this study, a combination of feature extraction and feature selection was used together with the evaluation of several different learning algorithms for recognizing emotional states with better performance.

3- Methodology

In this section, details of the proposed system are presented. Figure 1 summarizes the steps for building the emotion recognition system model.

The system receives a physiological signal of ECG as its input. Overall, the system is made of two main phases; (1) pre-processing, which involves data relabeling, feature extraction, and feature selection, and (2) classification using machine learning algorithms. After feature selection, the data is split into training and testing data, with the majority of the data used for training the machine learning models, while the remaining are used for testing the model's performance. This system is able to classify arousal, valence, dominance, and stress based on the input data. The classes reflect the emotional state of an individual.

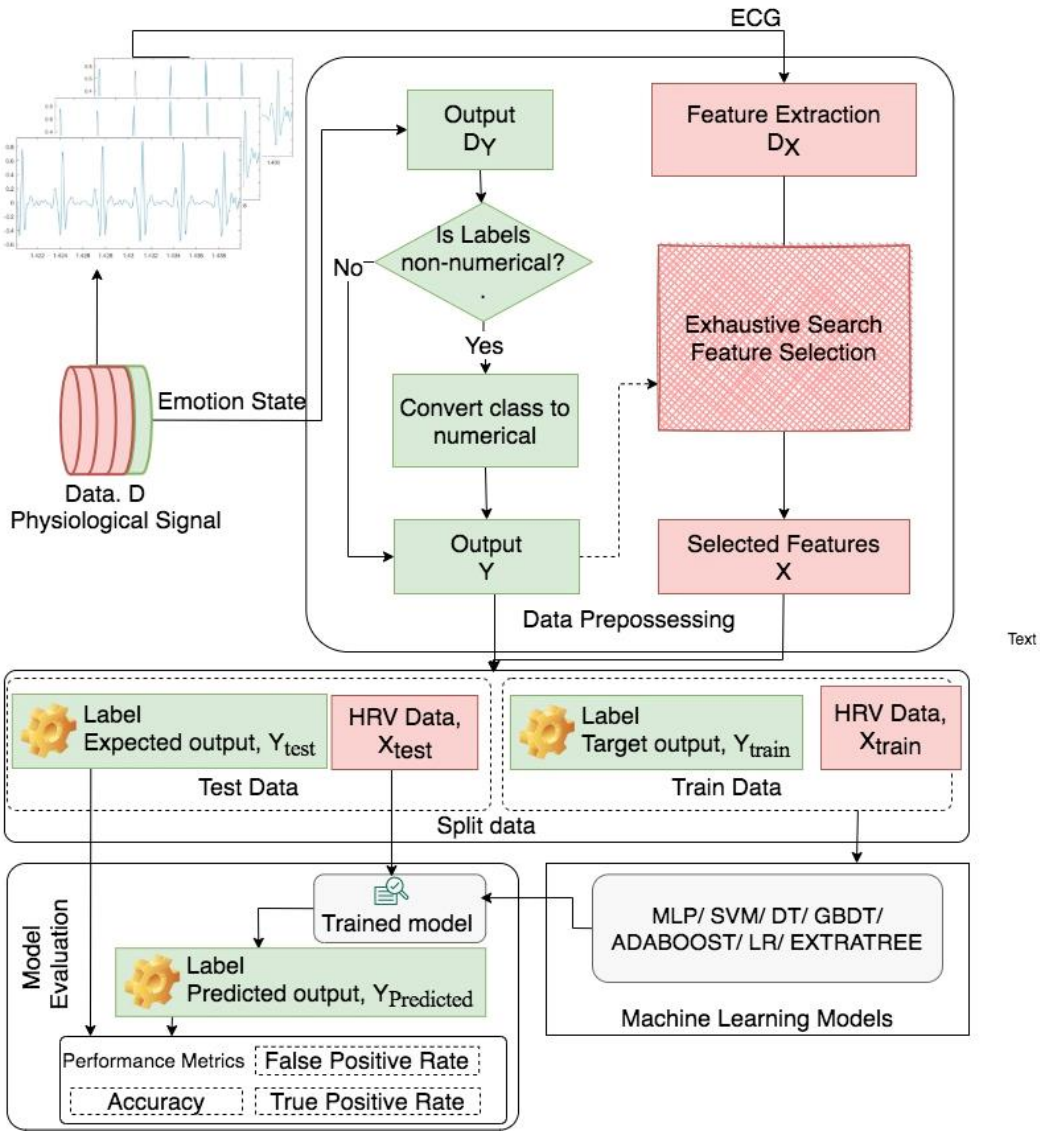


Figure 1. Overall System Flow to Develop Emotion Recognition Model

3-1- Data Pre-Processing

3-1-1- Data Relabelling

In Figure 1 the dataset is represented by 'D', where 'D' contains ECG signal 'DX' and their labels 'DY'. The proposed system framework is tested on benchmark datasets containing ECG data, DREAMER [25], and SWELL [26]. Table 1 briefly describes both datasets.

Table 1. Training and Testing Dataset Information

Dataset		DREAMER	SWELL
Participants		23	25
Data size		414	369000
Induction		video	working stressors
Data Available	Raw	✓	✓
	Pre-processed	×	✓
Physiological data, 'D'		ECG data at sampling rate 128/256 Hz	ECG
Features extracted, 'D_X'		27 HRV data	34 HRV data
Number of classes & categories, 'Y'		Valence (0,1)	0: no stress
		Arousal (0,1)	1: stress/ interrupted
		Dominance (0,1)	

Size of data is a significant challenge in this field; a large dataset can train the model better so that a better accuracy can be achieved. The SWELL dataset is significantly larger in comparison to DREAMER. SWELL dataset provided the pre-processed data of ECG's HRV features which are labelled into 3 classes of no stress, stress, and interrupted. These are non-numerical values (i.e., real numbers: $\mathbb{R}$); few of the selected classifiers are unable to train with non-numerical data. Therefore, in this work, the data is relabelled into two classes, 0 for no stress data while 1 for stress and interrupted. The new label is represented as Y in Figure 1.

Moreover, the DEAMER dataset contains raw ECG data. Therefore, to extract the HRV features NeuroKit2 library was used. The trained model outputs using the DREAMER dataset depend on emotional dimensions of valence, arousal, and dominance, where the values are ranged from 1 to 5. This research transforms the data into binary classes of high (1) and low (0) valence/arousal/dominance.

3-1-2- Feature Extraction

ECG data is a bio-signal that requires feature extraction prior to training the learning models. In Figure 1 the extracted features are represented as 'DX'.

The ECG measurements can be used to calculate HRV by different intervals (such as RR intervals), the time difference between two close R-wave peaks. ECG's HRV is known to be used to detect the emotional state of any individual [26, 27]. HRV value might reduce when a person is feeling happy, sad, or fearful, while the heart rate peak value might gradually increase in the state of anxiety 25. Additionally, the application of HRV to train a model for the classification of different problems is proven to give a good performance [33]. Therefore, HRV data is used here.

The DREAMER's ECG signals are converted into HRV features. NeuroKit2 tool developed by Makowski et al. [30] was used to get the following features from the ECG signal:

- Time-domain features; the root mean square of successive heartbeat interval difference (RMSSD), the mean interval between two heartbeats (MeanNN), the standard deviation of two heartbeats (SDNN);
- Frequency domain features; spectral power density in various frequency bands;
- Nonlinear domain features; spread of RR intervals, cardiac sympathetic index (CSI), cardiac vagal index (CVI), modified CSI, and sample entropy (SampEn).

On the other hand, the SWELL dataset in the Kaggle database comes with the pre-processed HRV features; hence, feature extraction is not required. A total of 27 features are generated from the DREAMER ECG signals, while the SWELL dataset provided 34 HRV features.

3-1-3- Feature Selection

Feature selection selects a subset of features from the whole set of features to obtain faster processing and improve classification performance. The selected features are represented as 'X' in Figure 1.

Here, the exhaustive feature selection (EFS) process is used to obtain the best features. The concept of EFS is visualized in Figure 2. EFS is one of the wrapper methods to choose the features. Due to the exhaustive nature of selecting a feature, it is known as EFS [34, 35]. Basically, EFS divides the features into subsets and obtains the best model by calculating performance metrics, such as Accuracy, precision, etc. EFS is similar to the forward selection method. However, forward selection suffers from greediness.

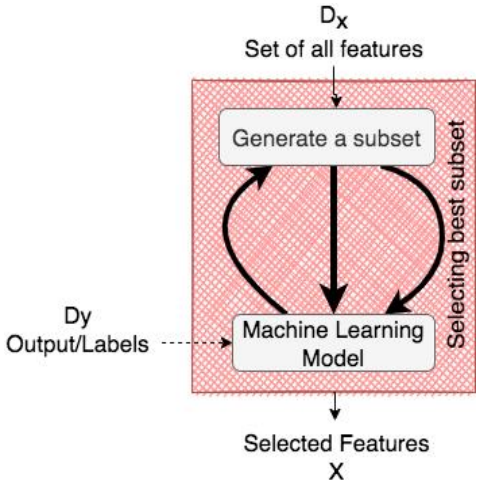
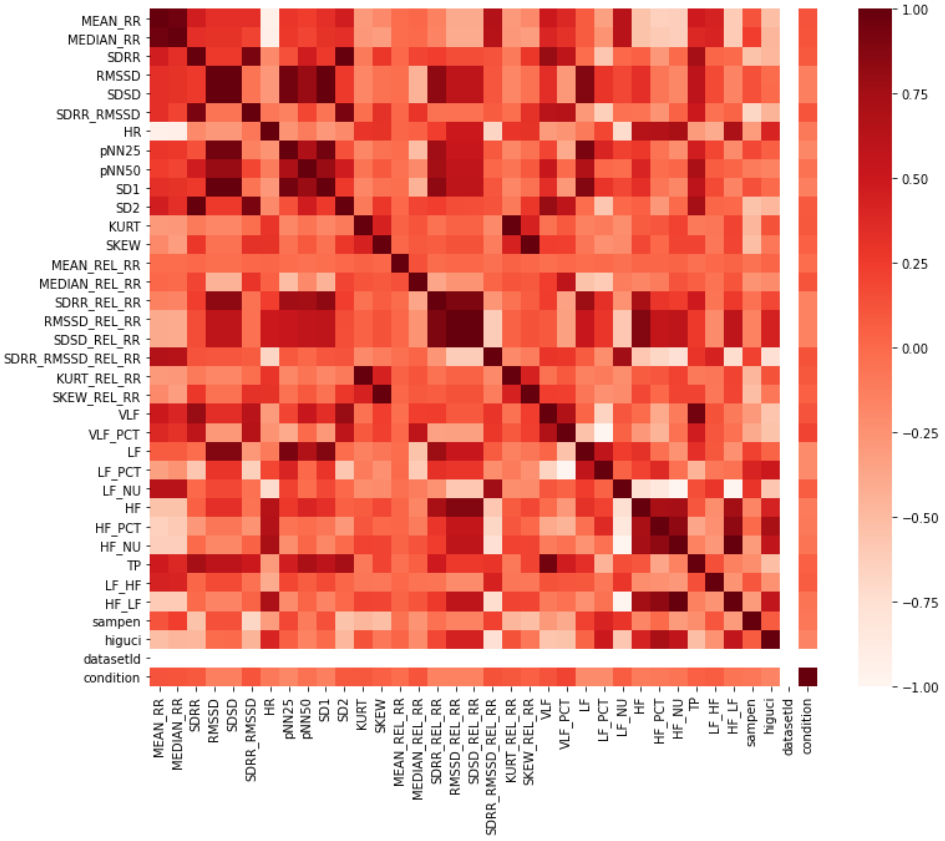


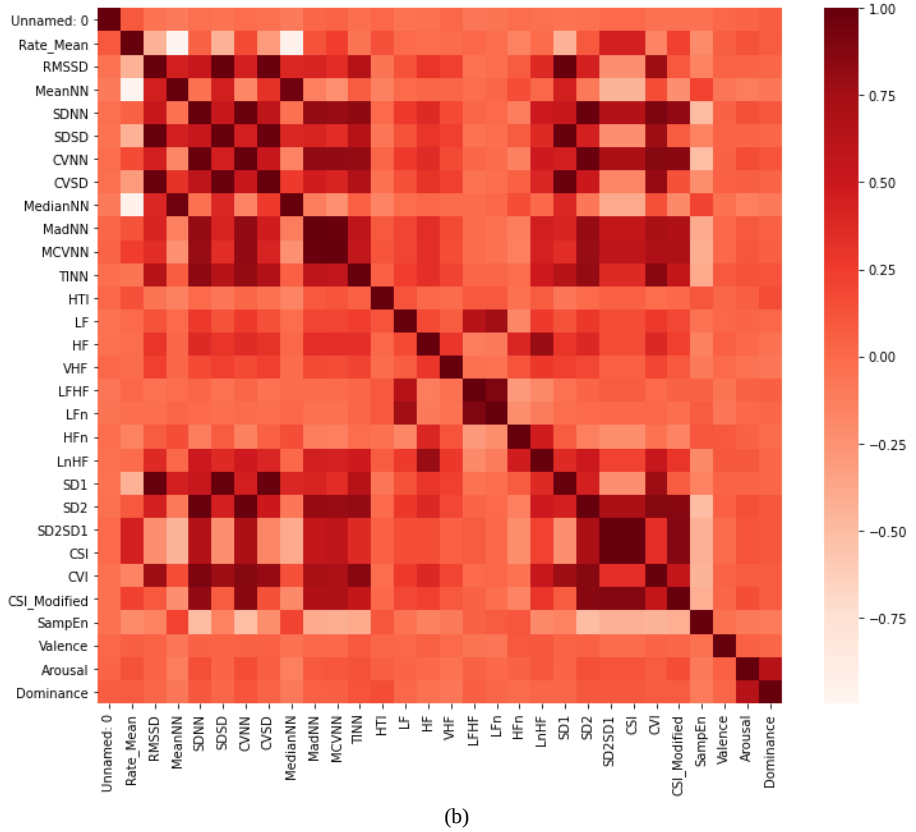
Figure 2. Exhaustive Search Feature Selection approach integrated into Overall Process Flow

Feature selection using Pearson's Correlation is also studied here. The correlation matrix visualization is presented in Figure 3; the values were obtained using Equation 1.

$$\boldsymbol{R} = \frac{\boldsymbol{n}\sum(\boldsymbol{x}\times\boldsymbol{y})-(\sum\boldsymbol{x})(\sum\boldsymbol{y})}{[\boldsymbol{n}\sum(\boldsymbol{x}^2)-\sum(\boldsymbol{x}^2)]\times[\boldsymbol{n}\sum(\boldsymbol{y}^2)-\sum(\boldsymbol{y}^2)]} \quad (1)$$



(a)



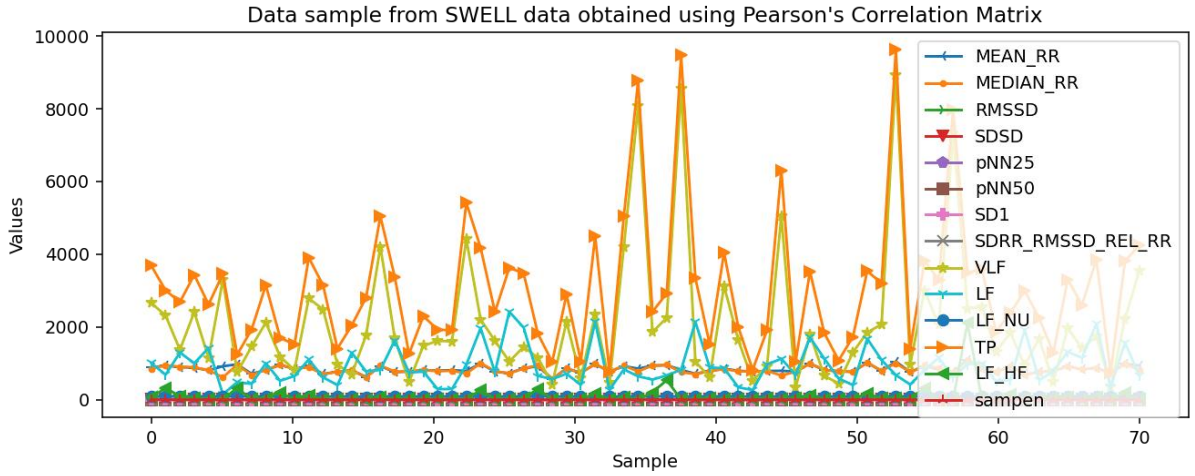
(b)

Figure 3. Pearson's Correlation Matrix indicating dependency of features in (a) SWELL dataset (b) DREAMER dataset

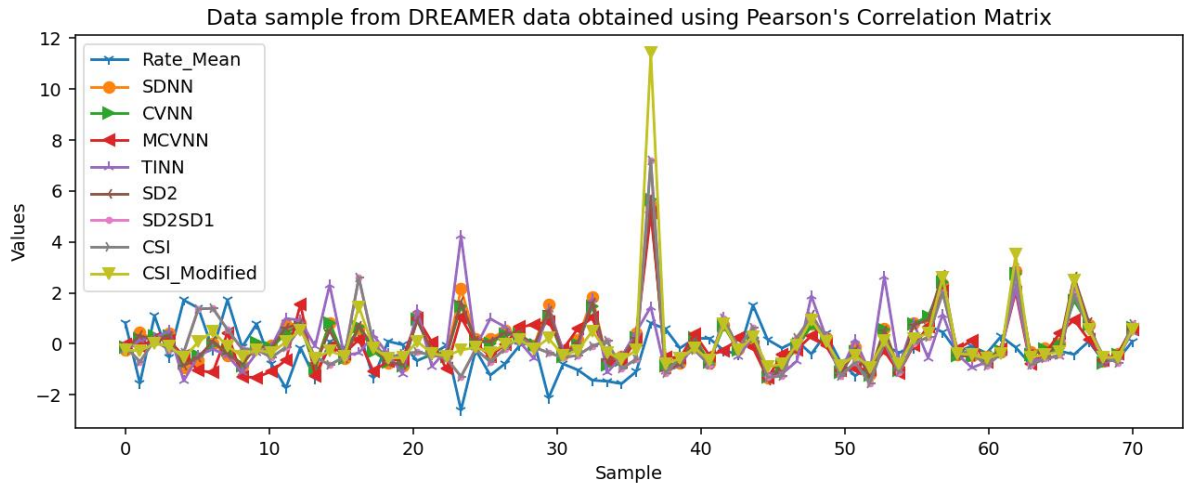
If the number of correlations between two features is closed to 1, it is considered an excellent correlated value, like the correlation value for MEAN_RR and MEADIAN_RR shown in Figure 3-a. The Pearson's Correlation Matrix indicates the dependency of the features. Next, using the function feature_selection() presented in Algorithm 1, the best features are selected to train the ML model (sample of selected features from SWELL and DREAMER datasets are shown in Figure 4).

Algorithm 1. Pearson's Correlation Matrix Feature Selection

```
DX=list of all features
X= Selected Features
correlation = pearson_correlation(DX)
threshold = 1
function feature_selection(correlation,threshold)
        X = []
        for i in range(correlation.shape[0]):
                if correlation in ith postion with selected No_of_Features is greater than
                threshold:
                        Insert ith feature into
        end for
        return X
end function
function pearson_correlation(dataset)
x=first set of data from input values
y=second set of data from input values
Calculate value R using equation
return R
end function
```



(a)



(b)

Figure 4. Training sample data obtained from (a) SWELL (b) DREAMER data with Pearson's Correlation Matrix

3-2- Model Classification and Evaluation

DREAMER dataset contains three dimensions for classification; valence, arousal and dominance. An emotion recognition system is built for each dimension. Meanwhile, for the SWELL dataset, only one model is required to predict whether stressful or not.

First, the selected features and the labels {X, Y} are split into training {X_{train}, Y_{train}} and testing {X_{test}, Y_{test}} subsets. In this work, the ratio is 80%:20%. Afterwards, X_{train} is used to train different models.

In this study, seven machine learning (ML) algorithms are used: multi-layer perceptron (MLP), support vector machine (SVM), decision tree classifier (DT), gradient boosting decision tree (GBDT), AdaBoost, logistic regression (LR) and Extra Tree. Each model parameter was tuned to obtain better models performances. The trained models are then tested using {X_{test}, Y_{test}} subsets. The test accuracy of each ML to classify the emotions is then compared in the next section.

3-2-1- Multi-Layer Perceptron (MLP)

MLP is an implementation of a feed-forward artificial neural network. MLP is developed using three layers (i.e., input, hidden and output layer). Each layer contains nodes, where each node in the hidden and output layer is known as a neuron with a nonlinear activation function. Equation 2 represents the ReLU (Rectified Linear Unit) function is used to calculate the activation function.

$$f(n) = \max(0, n) \tag{2}$$

where n is input data into a neuron, which is calculated using Equation 3:

$$n = \sum_{i=1}^{p} w_i x_i \tag{3}$$

where w_i is the weight assigned for ith input features and x_i is ith input feature.

3-2-2- Support Vector Machine (SVM)

Here, SVM is used to obtain the hyperplane in the N-dimensional space, where N is dependent on number of features. The hyperplane helps the process of classifying the data points. The plane should have maximum distance between the different classes. In this research, non-linear (RBF) kernel is used for Support vector classification. The kernel helps obtain the hyperplane for classifying different classes using Equation 4:

$$k(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \tag{4}$$

where $\|x_i - x_j\|^2$ is the squared Euclidean distance between two data input x_i and x_j vectors. For the study presented in this research, the misclassification rate, C is set to 2. The misclassification rate is the percentage of incorrectly classified by the trained model. The minimum value was selected to indicate a lower error bound and obtain the more significant margin gap between classes.

3-2-3- Decision Tree Classifier (DT)

The DT classifier consists of a tree with nodes, and those nodes are selected by generating an optimum split of the input features. The tree root and split of data provide the largest information gain (IG). To avoid model overfitting, the tree prune was set with a maximum depth of the tree is 8. To obtain IG for the parent node Equation 5 was used.

$$IG(D_p) = ID_p - \frac{N_{left}}{N_p} ID_{left} - \frac{N_{right}}{N_p} ID_{right} \tag{5}$$

where D_p, D_{left}, and D_{right} is a set of a parent, left and right dataset. In this study, to obtain the quality of the split entropy criterion was used, I is the entropy. The entropy was calculated by using Equation 6.

$$I = -\sum_i p_i \cdot \log_2 p_i \tag{6}$$

where p_i is the probability of target value i. Further to calculate classification error, Equation 7 was used

$$\text{DT}model_{error} = 1 - max(p_i) \tag{7}$$

3-2-4- Logistic Regression (LR)

Similar to SVM, LR creates a boundary between the different classes. Which later can be used to predict the output of the input data. Basically, for classification sigmoid function is calculated using equation 8, where the return value is between [0,1].

$$\emptyset(z) = \frac{1}{1+e^{-z}} \quad (8)$$

where z is "$W^t.x_i * y_i$", The parameter weight, W of the LR is chosen by maximizing the conditional data likelihood [36]. y_i is the output of the position i^{th} and x_i is input vector of the i^{th} position. Depending on the sigmoid output, the final decision is obtained, if $\emptyset(z) \geq 0.5$ then $\hat{y} = 1$ else $\hat{y} = -1$, where $\hat{y}$ is predicted output.

In this trained model, the inverse of the regularization strength parameter, C is set to 0.55, and the maximum number of iterations is set to 1000.

3-2-5- Ensemble Learning Model

In this study, an ensemble learning model based on GBDT [37] and Extra Trees Classifier was implemented. GBDT uses multiple DTs as base learning. The normal gradient boosting approach may lead to more significant misclassification [38], due to dependency between the increasing numbers of trees. However, for GDBT classifier, the input of the next tree is residual from the previous tree result, this helps to reduce the loss of the model. While in the Extra-Tree classifier (extremely randomized trees) [39], the DTs are randomly built using numerical input feature, the choice of the optimal cut-point of the process generates a huge number of variance of the induced tree. Extra Trees Classifier is like the "Random" in Random Forest as it uses a random subset of the dataset. Here, the threshold split values are chosen randomly. Determining feature and threshold split at each node helps obtain the model's "largest information gain" for the model.

Initially, to implement the ensemble learning model, the total data is randomly divided into m number of subsets, represented by $X_{i,j} = \{x_1, x_2, x_3, .., x_m\}$. For each set of $X_{i,j}$ a decision tree model is generated for the prediction of emotions. Then a set of trained models is obtained and represented as $P_{i,j} = \{p_1, p_2, p_3, .., p_m\}$. Finally, the interaction score of the pair is calculated using Equation 9, by summarizing all decision trees' scores.

$$\text{GBDT}model_{score}(i,j) = \frac{1}{m}\sum_{n=1}^{m} \lambda_n P_n(x_n) \quad (9)$$

where $P_n(x_n)$ is the score obtained for the decision tree P_n. For adjusting the contribution of the tree P_n, a constant value for that tree was used, i.e., λ_n. The loss of the GBDT is calculated using Equation 10

$$\text{GBDT}model_{loss} = \sum_{i,j} log\left(1 + e^{-2Y_{ij}\hat{Y}_{ij}}\right) \quad (10)$$

where, Y_{ij} is the actual interaction between every input data (s) and its target value. And $\hat{Y}_{ij}$ is calculated score using $\text{GBDT}model_{score}(i,j)$.

3-2-6- Adaboost

Adaboost algorithm is used here to boost up the performance of Logistic Regression ($F(s)$), by adding its parameters with additional parameters, such as learning rate and a number of estimators. The weak classifier (LR) generates an output hypothesis, $h(s)$ for each input sample, s. The classifier is assigned a coefficient α, and obtained the summation of training error by using Equation 11.

$$\text{ADA}model_{error} = \sum_i error[\alpha_i h_i(s)] \quad (11)$$

where $\alpha h(s)$ is the weak classifier obtained by input s for addition to the final classifier. For the training process of LR, a weight w_i is added to each sample in the training set.

4- Results and Discussion

All the seven machine learning algorithms are trained using DREAMER – valence, DREAMER – arousal, DREAMER – dominance, and SWELL data. The training is done using data that have to go through feature selection and without feature selection. Each machine learning algorithm's performance is compared, and the effect of feature selection is observed. The performance is measured using model's test accuracy and ROC (receiver operating characteristic) curve and benchmarked against the original work of each dataset. The accuracy is calculated using Equation 12.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False negative. Further ROC curve contains plots of two parameters, i.e., True Positive Rate (TPR) using Equation 13 and False Positive Rate (FPR) using Equation 14.

$$TPR = \frac{TP}{TP+FN} \quad (13)$$

$$FPR = \frac{FP}{FP+TN} \quad (14)$$

4-1- Model's Accuracy

For the DREAMER dataset, the classification accuracy is presented in Table 2. The accuracy of classification without feature selection shows that for valence, MLP and Adaboost are the best classifiers among the seven tested algorithms with 59% accuracy. However, comparing to original research [25] the performance is not as good as the one reported in the work of 62.37%. Extra tree classifier gave the best accuracy 65.6%, for arousal classification, even better than the original work, 62.37%. For the dominance classification, DT, GBDT and LR reported 61.4% accuracy, the best among the seven algorithms. These accuracies are slightly lower than what are reported in the original work, 61.57%.

Table 2. Performance Evaluation Using Dreamer Dataset

Model	Accuracy (100%) All features			Accuracy (100%) Selected Features using EFS			Accuracy (100%) Selected Features Using Pearson's correlation (No of features = 9)		
	Valence	Arousal	Dominance	Valence (No of features)	Arousal (No of features)	Dominance (No of features)	Valence	Arousal	Dominance
Multi-layer perceptron	59.0	62.6	56.6	66.2 (22)	74.6 (14)	66.2 (21)	55.42	69.03	60.24
Support Vector Machine	57.8	62.6	60.2	56.6 (21)	67.4 (14)	66.2 (22)	60.24	60.24	54.21
Decision Tree Classifier	48.2	55.42	61.4	68.6 (22)	74.6 (14)	69.8 (21)	53.01	62.65	42.16
Logistic Regression	57.8	60.24	61.4	62.4 (23)	67.4 (14)	66.2 (21)	60.24	59.03	61.44
Gradient Boosting Decision Tree	56.6	57.83	61.4	67.4 (22)	69.8 (21)	69.8 (22)	57.83	57.83	53.01
Extra tree	56.7	65.6	57.8	74.6 (21)	68.2 (14)	62.2 (21)	59.03	60.24	55.42
Adaboost	59.0	59.03	60.2	61.4 (25)	67.4 (14)	65.0 (21)	61.44	59.03	57.83
Previous Accuracy [25] using ECG only	62.37	62.37	61.57	-	-	-	-	-	-

The introduction of EFS improves the performance of all classifiers with the exception of SVM for valence, it also changes the best classifiers for each dimension. Extra tree is the best classifier for valence data, MLP and DT are for arousal data and DT and GBDT are for dominance data. All of these models reported accuracy better than the original work. The extra tree classifier with the EFS technique gave the best accuracy of 74.6% for valence. MLP and DT with EFS give 74.6% accuracy for arousal. Finally, for dominance, DT and GBDT model provides models with 69.8% accuracy. The best accuracy for valence is achieved with 21 features, while arousal with 14 features only and dominance with 21 (DT) and 22 (GBDT).

Further, the model was trained with Pearson's correlation features selection that selected 9 features using Algorithm 1. However, unlike EFS, the Pearson's correlation features selection improves performance of only some of the classifiers, MLP for arousal and dominance, SVM for valence, DT for valence and arousal, LR for valence, GBDT for valence, extra tree for valence and Adaboost for valence, while either not improving or lowering performance of the others. However, the classification using features selected using ERS is better than using features from Pearson's correlation for DREAMER dataset.

The test accuracy of the models trained using SWELL is tabulated in Table 3. The SWELL dataset has the advantage of larger amount of data. The accuracy found by the machine learning are significantly higher in comparison to their accuracy for DREAMER, except for LR. The extra tree classifier is the best classifier with and without feature selection, achieving 99.8%-100% and 97.4% accuracy respectively. Interestingly, unlike the DREAMER data, the EFS contributed to better performance of extra tree, GBDT, Adaboost and LR while the introduction of EFS prior to classification lowered the accuracy of MLP, SVM and DT. Moreover, Pearson's correlation feature selection had only improved ensemble learning models (i.e., GBDT and Extra tree), LR and Adaboost.

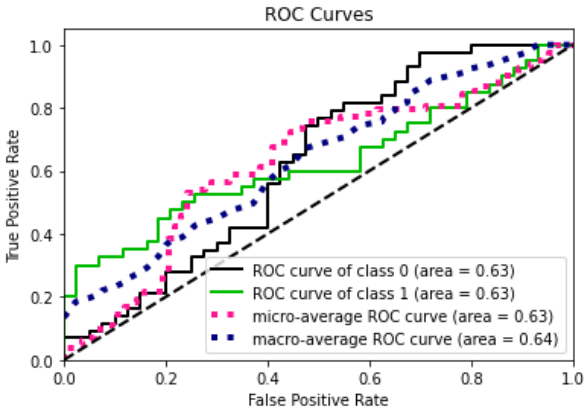
Table 3. Performance Evaluation Using SWELL Dataset

	Accuracy (100%) All features	Accuracy (100%) Selected Features using EFS (No of features = 8)	Using positive Pearson's correlation (No of features = 14)
Multi-layer perceptron (MLP)	82.2	73	65.62279
Support Vector Machine (SVM)	83.4	67	69.42704
Decision Tree Classifier (DT)	95.1	80	78.93
Logistic Regression (LR)	59	64	65.62279
Boosting Decision Tree (GBDT),	95.7	97	99.19333
Extra tree	97.4	99.8	100
Adaboost	63.5	66.7	65.2986
Previous Accuracy [16] using ECG only	-	64.0997	-

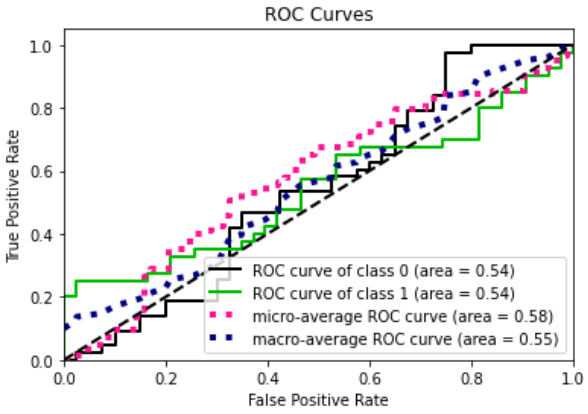
The trained Extra Trees classifier works perfectly for the SWELL dataset. It is due to its training approach; that the training input splits randomly. However, Extra Trees Classifier can be a higher bias and lower variance. The training process is held randomly and fits a number of randomized decision trees on different subsamples of the SWELL dataset. Then the value is averaged to enhance the predicted output and able to avoid over-fitting model.

4-2- ROC Curve

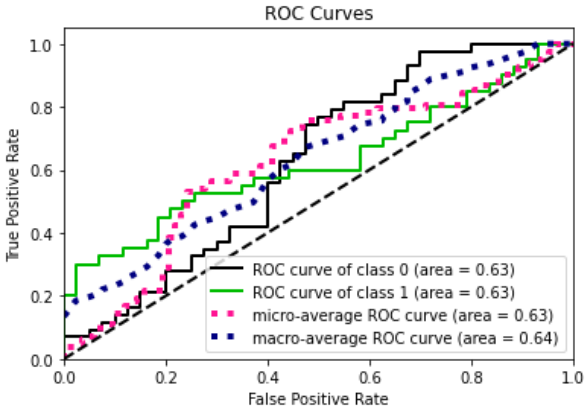
The ROC's area under the curves (AUC) of the best models is calculated to confirm that the best-developed model for each data is not over-fit or under-fit. The ROC curves are generated using the model trained and built's true positive rate vs. false positive rate. The ROCs in Figure 5 show that the trained Extra Tree model using the large SWELL dataset gives good performance for detecting stress levels in an individual (Figures 5-g and 5-h). The AUC value is 1 for both feature selection method. However, the models developed using the DREAMER dataset also give acceptable models as their AUC is greater than 0.5. The AUC value is 0.63, 0.63, and 0.54, for valence, arousal and dominance, respectively, using EFS feature selection method (Figures 5-a, c). AUC value for the model with Pearson's Correlation Matrix feature selection method is 0.53, 0.63 and 0.65 for valence, arousal and dominance, respectively (Figures 5 d and 5-f). The EFS feature selection tends to provide a better AUC value than Pearson's Correlation Matrix feature selection method.



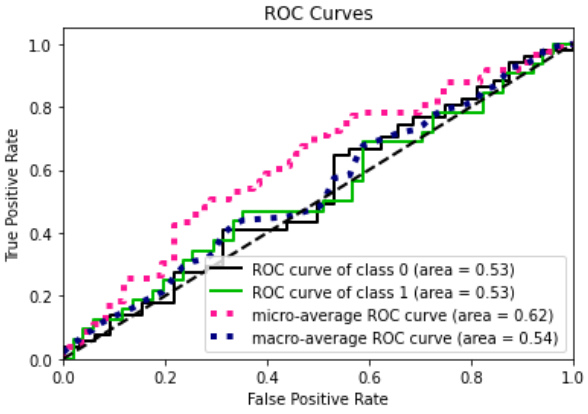
(a)



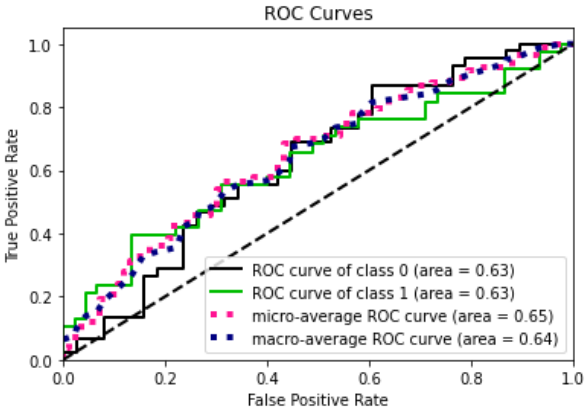
(b)



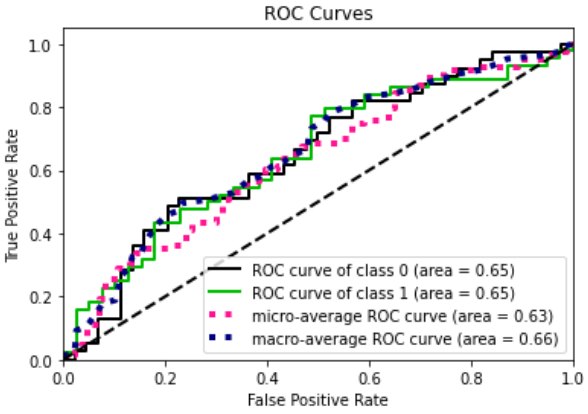
(c)



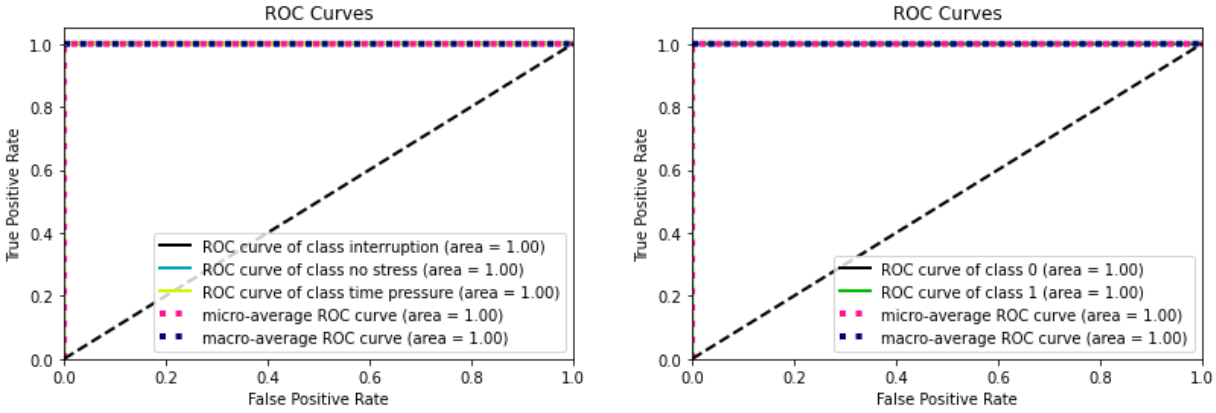
(d)



(e)



(f)



(g) (h)

Figure 5. ROC curve (a) DREAMER Dataset for Valence{0,1} with ESF Selection approach (b) DREAMER Dataset for Arousal{0,1} with ESF Selection approach (c) DREAMER Dataset for Dominance{0,1} with ESF Selection approach (d) DREAMER Dataset for Valence{0,1} with Pearson's Correlation Matrix feature Selection approach (e) DREAMER Dataset for Arousal{0,1} with Pearson's Correlation Matrix feature Selection approach (f) DREAMER Dataset for Dominance{0,1} with Pearson's Correlation Matrix feature Selection approach (g) SWELL dataset for Perceived stress{0,1} with ESF Selection approach (h) SWELL dataset for perceived stress with Pearson's Correlation Matrix feature selection (Class ratio {0,1} = {11,9}).

4-3- Average Rankings of Friedman Test

Here, the average ranks of the ML models obtained by applying the Friedman procedure output are presented in Table 4. This helps evaluate the best technique for ERS further. According to output, ET got the first position with 3.2083, and next, MLP got the second position. The last position was Adaboost, with a 4.7917 value.

Table 4. Ranking obtained for each trained ML algorithm

Algorithm	Ranking	Position
MLP	3.6667	2
SVM	4.2917	5
DT	3.8333	4
LR	4.4167	6
GBDT	3.7917	3
Extra trees	3.2083	1
Adaboost	4.7917	7

Friedman statistic considering reduction performance (distributed according to chi-square with 6 degrees of freedom: 4.357143) P-value computed by the Friedman Test: 0.6284651026747108. The Friedman statistical value obtained is 0.628, which is less than 12.59. Therefore, the null hypothesis is accepted. Extra tree classifiers are on par with and as good as the other methods.

5- Conclusion

In this study, the main aim is to use only one modality, the ECG, to recognize an individual's emotional state. Two benchmark datasets are used in this research: DREAMER and SWELL. Seven machine learning models are trained and tested. EFS technique and Pearson correlation are used to select the features to enhance the performances of the trained models. The results show that the Extra Tree classifier is able to achieve the best performance for SWELL, while for DREAMER dataset, the classifier depends on the emotion dimension. However, the accuracy achieved for SWELL data, which size is larger, is better than DREAMER. This shows the importance of dataset size in building a good ERS. The adoption of EFS improves almost all predictive emotion recognition models for DREAMER and SWELL data. The contribution of EFS is more significantly observed for small datasets. The Pearson's correlation, on the other hand, only improves the performance of some of the tested ML. This suggests that Pearson's correlation is not a suitable feature selection method for an emotion recognition system. Overall, the findings show that the selection of a classifier and feature selection method is a problem-dependent issue.

Additionally, the size of the data plays an essential role in building a good emotion recognition system. Hence, this issue needs to be addressed when building an emotion recognition system. There are a few drawbacks in this research work that have to be revised to improve the results. The first and foremost barrier of this study is the small dataset,

leading to the inefficacy of learning algorithms. Despite the popularity of deep learning, it cannot deliver efficient performance due to the lack of a large dataset. Hence, the machine learning approach was considered for this research work. In the future, with a larger dataset, the machine learning models can be replaced by deep learning on highly configured workstations for better prediction.

6- Declarations

6-1- Author Contributions

Conceptualization, N.A.A.A. and C.M.T.K.; methodology, C.M.T.K.; validation, J.E.R., and S.W.B.N; writing—original draft preparation, C.M.T.K.; writing—review and editing, N.A.A.A., J.E.R., P.R. and S.W.B.N; visualization, C.M.T.K.; supervision, N.A.A.A.; project administration, N.A.A.A.; funding acquisition, N.A.A.A. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6-3- Funding

This project is funded by TM Research & Development Grant (RDTC/190988) which is awarded to the Multimedia University.

6-4- Acknowledgements

The authors want to thank those who involved in this experiment directly or indirectly. Especially our emotion recognition research team in the Centre for Engineering Computational Intelligence, Multimedia University.

6-5- Institutional Review Board Statement

Not applicable.

6-6- Informed Consent Statement

Not applicable.

6-7- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Picard, R. W. (1997). Affective Computing Cambridge. MIT Press, Cambridge, Massachusetts, United States. doi:10.1037/e526112012-054.

[2] Izard, C. E. (2009). Emotion theory and research: Highlights, unanswered questions, and emerging issues. Annual Review of Psychology, 60(60), 1–25. doi:10.1146/annurev.psych.60.110707.163539.

[3] D'Mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. ACM Computing Surveys, 47(3). doi:10.1145/2682899.

[4] Turabzadeh, S., Meng, H., Swash, R., Pleva, M., & Juhar, J. (2018). Facial Expression Emotion Detection for Real-Time Embedded Systems. Technologies, 6(1), 17. doi:10.3390/technologies6010017.

[5] Konar, A., & Chakraborty, A. (2015). Emotion recognition: A pattern analysis approach. John Wiley & Sons, Hoboken, United States. doi:10.1002/9781118910566.

[6] Ali, M., Al Machot, F., Mosa, A. H., & Kyamakya, K. (2016). A novel EEG-based emotion recognition approach for e-healthcare applications. Proceedings of the 31st Annual ACM Symposium on Applied Computing. doi:10.1145/2851613.2851916.

[7] Bhise, P. R., Kulkarni, S. B., & Aldhaheri, T. A. (2020). Brain Computer Interface based EEG for Emotion Recognition System: A Systematic Review. 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). doi:10.1109/icimia48430.2020.9074921.

[8] Kumar, A., Garg, N., & Kaur, G. (2019). An emotion recognition based on physiological signals. International Journal of Innovative Technology and Exploring Engineering, 8(9 Special Issue), 335–341. doi:10.35940/ijitee.I1054.0789S19.

[9] Dissanayake, T., Rajapaksha, Y., Ragel, R., & Nawinne, I. (2019). An ensemble learning approach for electrocardiogram sensor based human emotion recognition. Sensors (Switzerland), 19(20), 4495. doi:10.3390/s19204495.

[10] Ayata, D., Yaslan, Y., & Kamasak, M. E. (2020). Emotion Recognition from Multimodal Physiological Signals for Emotion Aware Healthcare Systems. Journal of Medical and Biological Engineering, 40(2), 149–157. doi:10.1007/s40846-019-00505-7.

[11] Li, Z., Tian, X., Shu, L., Xu, X., Hu, B. (2018). Emotion Recognition from EEG Using RASM and LSTM. Internet Multimedia Computing and Service. ICIMCS 2017, Communications in Computer and Information Science, 819. Springer, Singapore. doi:10.1007/978-981-10-8530-7_30.

[12] Nikolova, D., Petkova, P., Manolova, A., & Georgieva, P. (2018). ECG-based emotion recognition: Overview of methods and applications. ANNA'18; Advances in Neural Networks and Applications 2018, 15-17 September, 2018, St. Konstantin and Elena Resort, Bulgaria.

[13] Katsis, C. D., Katertsidis, N., Ganiatsas, G., & Fotiadis, D. I. (2008). Toward emotion recognition in car-racing drivers: A biosignal processing approach. IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, 38(3), 502–512. doi:10.1109/TSMCA.2008.918624.

[14] Tivatansakul, S., & Ohkura, M. (2013). Healthcare System Focusing on Emotional Aspects Using Augmented Reality - Implementation of Breathing Control Application in Relaxation Service. 2013 International Conference on Biometrics and Kansei Engineering. doi:10.1109/icbake.2013.43.

[15] Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., & Wróbel, M. R. (2014). Emotion Recognition and Its Applications. Human-Computer Systems Interaction: Backgrounds and Applications 3, 51–62. doi:10.1007/978-3-319-08491-6_5.

[16] Koldijk, S., Neerincx, M. A., & Kraaij, W. (2018). Detecting Work Stress in Offices by Combining Unobtrusive Sensors. IEEE Transactions on Affective Computing, 9(2), 227–239. doi:10.1109/TAFFC.2016.2610975.

[17] Zhang, T., & Iyengar, V. S. (2002). Recommender Systems Using Linear Classifiers. Journal of Machine Learning Research, 2(3), 313–334. doi:10.1162/153244302760200641.

[18] Garrett, D., Peterson, D. A., Anderson, C. W., & Thaut, M. H. (2003). Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 11(2), 141–144. doi:10.1109/tnsre.2003.814441.

[19] Davidson, R. J., Ekman, P., Saron, C. D., Senulis, J. A., & Friesen, W. V. (1990). Approach-Withdrawal and Cerebral Asymmetry: Emotional Expression and Brain Physiology I. Journal of Personality and Social Psychology, 58(2), 330–341. doi:10.1037/0022-3514.58.2.330.

[20] Yang, C., Lu, J., Wu, Q., & Chen, H. (2021). Research progress of speech emotion recognition based on discrete emotion model. Journal of Physics: Conference Series, 2010(1). doi:10.1088/1742-6596/2010/1/012110.

[21] Zhang, Z., Song, Y., Cui, L., Liu, X., & Zhu, T. (2016). Emotion recognition based on customized smart bracelet with built-in accelerometer. PeerJ, 2016(7), 1–14,. doi:10.7717/peerj.2258.

[22] Jang, E. H., Park, B. J., Park, M. S., Kim, S. H., & Sohn, J. H. (2015). Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. Journal of Physiological Anthropology, 34(1), 1–12. doi:10.1186/s40101-015-0063-5.

[23] Cho, D., Ham, J., Oh, J., Park, J., Kim, S., Lee, N. K., & Lee, B. (2017). Detection of stress levels from bio-signals measured in virtual reality environments using a kernel-based extreme learning machine. Sensors (Switzerland), 17(10). doi:10.3390/s17102435.

[24] Gravina, R., & Li, Q. (2019). Emotion-relevant activity recognition based on smart cushion using multi-sensor fusion. Information Fusion, 48, 1–10. doi:10.1016/j.inffus.2018.08.001.

[25] Katsigiannis, S., & Ramzan, N. (2018). 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, 22(1), 98–107. doi:10.1109/JBHI.2017.2688239.

[26] Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., & Kraaij, W. (2014). The SWELL Knowledge Work Dataset for Stress and User Modeling Research. Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey November 12 - 16. doi:10.1145/2663204.2663257.

[27] Goshvarpour, A., Abbasi, A., & Goshvarpour, A. (2017). An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. Biomedical Journal, 40(6), 355–368. doi:10.1016/j.bj.2017.11.001.

[28] Mehmood, R. M., & Lee, H. J. (2015). Exploration of Prominent Frequency Wave in EEG Signals from Brain Sensors Network. International Journal of Distributed Sensor Networks, 2015(386057). doi:10.1155/2015/386057.

[29] Chen, J., Hu, B., Wang, Y., Moore, P., Dai, Y., Feng, L., & Ding, Z. (2017). Subject-independent emotion recognition based on physiological signals: A three-stage decision method. BMC Medical Informatics and Decision Making, 17. doi:10.1186/s12911-017-0562-x.

[30] Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. H. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior Research Methods, 53(4), 1689–1696. doi:10.3758/s13428-020-01516-y.

[31] Shu, L., Yu, Y., Chen, W., Hua, H., Li, Q., Jin, J., & Xu, X. (2020). Wearable emotion recognition using heart rate data from a smart bracelet. Sensors (Switzerland), 20(3), 1–19. doi:10.3390/s20030718.

[32] Bulagang, A. F., Mountstephens, J., & Teo, J. (2021). Multiclass emotion prediction using heart rate and virtual reality stimuli. Journal of Big Data, 8(1). doi:10.1186/s40537-020-00401-x.

[33] El Attaoui, A., Hazmi, M., Jilbab, A., & Bourouhou, A. (2020). Wearable Wireless Sensors Network for ECG Telemonitoring Using Neural Network for Features Extraction. Wireless Personal Communications, 111(3), 1955–1976. doi:10.1007/s11277-019-06967-x.

[34] Fotiadou, E., Xu, M., van Erp, B., van Sloun, R. J. G., & Vullings, R. (2020). Deep Convolutional Long Short-Term Memory Network for Fetal Heart Rate Extraction. 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). doi:10.1109/embc44109.2020.9175442.

[35] Narendra, P. M., & Fukunaga, K. (1977). A Branch and Bound Algorithm for Feature Subset Selection. IEEE Transactions on Computers, C–26(9), 917–922. doi:10.1109/TC.1977.1674939.

[36] Kumar, M., & Rath, S. K. (2016). Feature Selection and Classification of Microarray Data Using Machine Learning Techniques. Emerging Trends in Applications and Infrastructures for Computational Biology, Bioinformatics, and Systems Biology, 213–242. doi:10.1016/b978-0-12-804203-8.00015-8.

[37] Ye, J., Chow, J.-H., Chen, J., & Zheng, Z. (2009). Stochastic gradient boosted distributed decision trees. Proceeding of the 18th ACM Conference on Information and Knowledge Management - CIKM '09. doi:10.1145/1645953.1646301.

[38] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. doi:10.1214/aos/1013203451.

[39] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3–42. doi:10.1007/s10994-006-6226-1.