Affective Recognition in Dynamic and Interactive Virtual Environments

Stone, Robert

DOI:
10.1109/TAFFC.2017.2764896

License:
Other (please specify with Rights Statement)

Document Version
Peer reviewed version

Citation for published version (Harvard):
Stone, R 2017, 'Affective Recognition in Dynamic and Interactive Virtual Environments', *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2017.2764896

Affective Recognition in Dynamic and Interactive Virtual Environments

Journal:	*Transactions on Affective Computing*
Manuscript ID	TAFFC-2016-11-0206
Manuscript Type:	Survey
Keywords:	Virtual Reality, Affective Computing, Affective VR, Emotion-based affective physiological database

Affective Recognition in Dynamic and Interactive Virtual Environments

Mohammadhossein Moghimi[1], Prof. Robert Stone[2] and Dr. Pia Rotshtein[3]
[1,2] School of Electronic, Electrical and Systems Engineering, University of Birmingham
[3] School of Psychology, University of Birmingham
[1]m.moghimi@pgr.bham.ac.uk, [2]r.j.stone@bham.ac.uk, [3]p.rotshtein@bham.ac.uk

Abstract – The past decade has witnessed a significant increase in interest in human emotional behaviours in the future of interactive multimodal computing. Although much consideration has been given to non-interactive affective stimuli (e.g. images and videos), the recognition of emotions within interactive virtual environments has not received an equal level of attention. In the present study, a psychophysiological database, cataloguing the EEG, GSR and heart rate of 30 participants, exposed to an affective virtual environment, has been constructed. 743 features were extracted from the physiological signals. Then, by employing a feature selection technique, the dimensionality of the feature space was reduced to a smaller subset, containing only 30 features. Four classification techniques (KNN, SVM, Discriminant Analysis (DA) and Classification Tree) were employed to classify the affective psychophysiological database into four Affective Clusters (derived from a Valence-Arousal space) and eight Emotion Labels. By employing cross-validation techniques, the performances of more than a quarter of a million different classification settings (various window lengths, classifier settings, etc.) were investigated. The results suggested that the physiological signals could be employed to classify emotional experiences, with high precision. The KNN and SVM outperformed both Classification Tree and DA classifiers; with 97.01% and 92.84% mean accuracies, respectively.

Keywords – Virtual Reality, Affective Computing, Affective VR, Emotion-based affective physiological database

1. Introduction

The recent "resurrection" of interest in Virtual Reality (VR), spurred on by the emergence of new interface and gaming technologies from international crowd-funding communities has, once again, stimulated interest in the quest for true "immersion" or the generation of a believable sense of "presence" in computer-generated worlds. Human-Computer Interaction (HCI) system designers have, in their quest to deliver compelling end user experiences of immersion, introduced several multi-dimensional input/output devices (aiming to provide user-friendly, intuitive techniques and styles of interaction with real-time 3D worlds[1]. However, the one area of HCI research that, arguably, is best placed to achieve immersion in the future, is that which strives towards establishing direct communication between a computer system and the human brain and has, until recently, been treated as science fiction (referencing such popular films as *The Matrix* and *Pacific Rim*[2,3]). In 2006, Cairns suggested that true "immersion" may only ever be achieved through the use of advanced brain-computer interfaces [**1**]. However, until that day arrives, it is important to understand in advance, how it may be possible to measure and, indeed, influence human engagement and emotional connectivity with virtual worlds using psychophysiological techniques.

Brain-Computer Interaction (BCI) systems attempt to improve human-computer interaction and increase the sense of immersion by interfacing directly with the human nervous system (both the central and autonomic nervous systems) and, thus, removing the artificial barriers to intuitive interaction afforded by conventional input-display techniques. These new interface channels can have the potential to introduce a large number of new communication techniques in advanced HCI systems, and may be able to improve the interaction process considerably (e.g. translating imaginary movements to virtual actions, improving levels of concentration, affecting emotional states, and so on). So far the interaction process has been mostly based on conventional methods, in that computer users typically use physical interaction devices to see, hear, act, sense haptic or olfactory stimuli and in some cases even talk to the system. The near-term goal of BCI systems, as an extension to these conventional systems (as opposed to a replacement, which is a longer-term aspiration), would be to translate human thoughts and emotions by direct connection to the human brain and use this information as a new modality channel for HCI systems [**2**]. To date, researchers have concentrated on the introduction of direct brain-computer interaction into HCI systems and virtual realities (VR) [4], with just some examples cited in the literature including the control, eliciting and measurement of depression and sadness within a virtual park [**3**], game adaptation according to stress levels [**4**], and using brain waves to either navigate through a virtual environment [**5**] or to control the balancing behaviours of a virtual avatar [**6**].

[1] For example various types of data input controllers, multifunctional touch panels
[2] https://en.wikipedia.org/wiki/The_Matrix
[3] https://en.wikipedia.org/wiki/Pacific_Rim_(film)

[4] Also known as *Neurogaming* [**1**]

One of the sub-categories of research into BCI systems is described as *affective computing*. Affective computing, in essence, records psychophysiological signals from the users, to enable the BCI system to extract data of relevance to their emotional and cognitive states. This new input channel could provide several features for advanced HCI systems, attempting to support the generation of believable immersive experiences. As an illustration, the system could use this information to adapt itself to the user's emotions and, by doing so, increase his/her performance and immersion levels, during the interaction process.

Turning briefly to the field of VR and the relevance of issues of affect, to date, researchers have studied the implementation of virtual realities in many different areas. As well as entertainment, VRs and their so-called "serious games" counterparts have been used for training purposes [**7**], [**8**], [**9**], pain distraction [**10**], [**11**], rehabilitation régimes [**12**], [**13**] and emotional disorder therapy [**14**], [**15**], to mention but a handful of applications. The focus of all these studies has been to engage human users in an interactive virtual environment, and to increase their sense of presence and immersion within them, thereby effectively delivering new skills, knowledge or in some cases, acting as a form of clinical distraction. In 2006 Joels suggested that changes in the excitement level (depending on pleasurable or dis-pleasurable condition), affects the learning and memory process. He proposed that memory performance changes (either improvements or impairments) are highly dependent on the time and context of the emotional experience [**16**]. Therefore, the recognition of users' emotions, when exposed to virtual realities, and controlling their affective experiences within the virtual environments (regardless of their purpose) can be as important as the VR's contextual outcome.

In [**17**] we conceptualised, designed and evaluated an *Affective Virtual Reality* (Affective VR), capable of evoking various emotional experiences on the part of the human user. In the present study, by employing the designed Affective VR, an affective computing system was conceptualised, designed and evaluated. To do this, the relationship between psychophysiological signals and human emotions, evoked through the designed Affective VR (presented in [**17**]), has been the focus of investigation. To support this research:

1. A psychophysiological experiment has been conducted, in which simple Electroencephalography (EEG), galvanic skin response (GSR) and Heart Rate signals of 30 participants have been recorded whilst playing the most powerful affective games, identified in [**17**]. These physiological signals provided a comprehensive database for further analysis and for the construction and evaluation of an Affective Recognition system (Section 2 of the present paper).

2. A number of overlapping windows of the raw physiological signals were separated for the feature extraction process. 743 physiological features were identified and extracted from the recorded database (Section 3 of the present paper).

3. By employing a feature selection technique, a small number of the most optimum features have been identified, to reduce the dimensionality of the database to a smaller subspace, to be used in the emotion classification process (Section 3.4 of the present paper).

4. By employing four classification algorithms (K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Discriminant Analysis (DA) and Classification Tree), the performances of a quarter of a million different classification settings (various window lengths and types, different number of features, etc.) have been evaluated and compared, using a 10-fold cross validation (Sections 4 and 5 of the present paper).

2. Psychophysiological Database Construction

2.1. Material

2.1.1. Affective Virtual Reality

To construct the psychophysiological database, the designed and evaluated Affective VR, presented in [**17**], has been used as the source of emotional stimuli. The Affective VR was based on a speedboat simulation[5] (Figure 1) acting as the background scenario. A number of parameters (called *affective incidents)* were implemented in the VR to change the affective power of the environment, within the Circumplex of Affect, presented by Russell in 1980s [**18**]. As an illustration, participants were challenged by driving the boat and collecting scores 'freely', in a 'minefield' or whilst 'being targeted by torpedoes', in various experimental setups, such as, coloured images, black and white or inverse black and white screens, using a mouse, or joystick with or without simple force feedback (for more details, refer to [**17**]).

As presented in [**17**], the environment was evaluated as effective in the manipulation of participants' emotions, in terms changes of locations of those emotions within the Circumplex of Affect. The duration of each game could vary between 90 to 300 seconds, depending on the performance or determination of the participants. There were some games with time limitations, which had to be completed as soon as possible depending on the participants'

[5] A short video of this simulation can be viewed at: https://www.youtube.com/watch?v=pqn-X1Z5AoM

performance. However, there were also those games that allowed the participants to drive freely within the environment, with no time limitations (no game was allowed to be longer than five minutes, and would be terminated if a participant spent longer than five minutes).

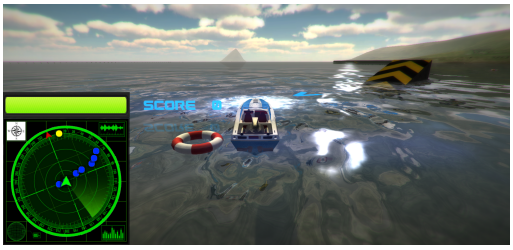

Figure 1 – Speedboat Simulation Environment

In the present study, the two most powerful affective games, in each of the four Affective Clusters[6] introduced in [**19**], have been identified using the Cosine Similarity Algorithm [**20**] as implemented in [**17**]. As a result of this analysis, the eight most powerful affective games (those, which have the highest probability of driving the emotional experience of the participants toward all affective clusters) have been identified. Following the identification of the most powerful affective games, two '*neutral games*' were added in the experiment (the neutral game from [**17**], plus the game close to (0, 0, 0) with the highest standard deviation). Therefore, overall, 10 affective games have been identified for presentation to the participants in the present experiment.

2.1.2. Participant Selection

One of the most important challenges of designing any affective psychophysiological database is the minimization of variability between participants, in each individual affective session, whilst maximizing the variability between sessions' experiences. This is due to the fact that, in any human-cantered experiment, minimum variability between participants' experiences, in a single affective session is an extremely important issue. Any acceptable analysis, dealing with either affects or physiological databases, should, intuitively, be based on changes in emotional experiences, due to **different environments**, rather than different **personal experiences**.

As discussed in [**17**], the Multi-variant Analyses of Variance (MANOVA) highlighted significant differences between the four participant groups (male gamers, male non-gamers, female gamers and female non-gamers). According to the results presented in [**17**], male gamers, male non-gamers and female gamers show marked similarities in their affective experiences, when compared to female non-gamers. Therefore, in the present study, in order to minimise *between participants* variability, it was decided to recruit only **male and female gamers** in the experiment. Recruiting the male non-gamers (despite their highest level of similarity compared to the others) in the experiment would disturb the comparison, as no female non-gamers were to be recruited. Therefore, 30 gamer participants of both genders (15 of each) were recruited to take part in this experiment (mean age=22.76). Each participant received a £10 gift voucher at the end of each experiment. The study was reviewed and approved by the University of Birmingham's Ethical Review Committee (Ethical Reference Number: ERN_13-1157).

2.1.3. Physiological Signal Recording

As discussed in [**19**], the majority of studies have employed EEG, Heart Rate and GSR signals to perform affective analysis and recognition. Therefore, in the present study it was decided to record data using these three techniques, for the purposes of supporting the psychophysiological database construction process. Participants were required to wear an EPOC EMOTIV[7] headset to record EEG signals, as well as Shimmer+ wearable sensor technologies[8] to record GSR and heart rate activities. The EPOC EMOTIVE records the EEG signals, with a 128Hz sampling frequency, from 14 channels[9]. The electrodes are arranged according to the 10-20 EEG system. The GSR and heart rate data are also recorded using the Shimmer+ wearable sensor technologies, with a 512Hz sampling frequency.

A program was developed to function in parallel with the game, to establish a connection with the Shimmer+ and EPOC EMOTIV devices (through the software development tool kits (SDKs) provided by the manufacturers), as soon as each game was started. As well as signal recording, the program performed time synchronisation, to align the outputs received from the devices, prior to data storage. Furthermore, a wireless real-time monitoring application (tablet-based) was implemented, to enable the experimental supervisor to monitor the software functionality and signal qualities of the recording devices, without any distraction to the participants' experiences.

2.2. Method

The experiment was performed in a quiet room. All participants were provided with a 32-inch Samsung HD LCD display, a Microsoft

[6] (1) Positive Valence and Low Arousal *(PVLA)* – (2) Positive Valence and High Positive Arousal *(PVHPA)* – (3) Negative Valence and Positive Arousal *(NVPA)* – (4) Negative Valence and Negative Arousal *(NVNA)*.

[7] http://emotiv.com/

[8] http://www.shimmersensing.com/

[9] AF3, AF4, F3, F4, F7, F8, FC5, FC6, T7, T8, P7, P8, O1 and O2, while P3 (Common Mode Sense – CMS) and P4 (Driven Right Leg – DRL) are used as the reference channels.

Wireless Mouse 5000, a Logitech Wingman 3D force feedback joystick and Sennheiser earphones. Each experiment commenced with a training session to prepare the participants for every possible incident within the games (as presented in [**17**]). The training introduced the game environment to the participants and served to reduce any element of surprise in the games. After the participants had completed the training session, they progressed to the two neutral games, followed by the other eight in a random order. At the end of each game the participants were instructed to self-assess their average emotional experience, based on both dimensional (Valence, Arousal and Dominance) and categorical (according to eight Emotion Labels: Relaxed, Content, Happy, Excited, Angry, Afraid, Sad and Bored) models of affect (as presented in [**17**]). The participants were given a 5- to 15-minute break, after playing the first five games, in order to reduce the fatigue factor caused by wearing the physiological sensing equipment. On average, each game lasted for three minutes, and the complete experiment took approximately 1.5 hours.

2.3. Results (Psychophysiological Database)

Of a possible total of 300 affective sessions, 290 were recorded, as 10 sessions were not attended by participants. During the affective sessions, the raw EEG signals from all 14 channels were recorded. Furthermore, the signal quality of each EEG channel was available from the EPOC EMOTIV headset[10] and was therefore recorded alongside the raw channel data. The raw Photoplethysmogram (PPG) output was recorded by the Shimmer+ device, mounted on the participant's index finger. During this recording a location of the skin is illuminated, and then the changes in light reflection are recorded. The alternating current component of the PPG signal relates to the blood pulse pressure. The Shimmer+ software uses the estimation techniques introduced in [**21**] to approximate the heart rate of the participants using the PPG signal. Moreover, the GSR signal was also recorded using two finger straps mounted on the middle and ring fingers. These raw data sources were synchronised according to the master clock of the main system and stored in Microsoft Excel files during the run-time of the experiment. The emotional ratings of the participants were recorded and stored separately at the end of each game.

3. Feature Matrix Construction

In this study, all pre-processing, windowing and data analyses have been implemented using MATLAB software (version R2015b).

[10] According to five classes; good, fair, poor, very bad and disconnected.

3.1. Pre-Processing

3.1.1. Filtering

As discussed in [**19**], majority of affective recognition studies apply no filtering technique to heart rate and GSR signals recorded from the participants. Therefore in this study, we also decided to use the heart rate and GSR signals without applying any filter. However, and as discussed in [**19**], majority of previous studies employed a band-pass filter (majority using 4Hz to 45Hz) in their EEG filtering process. Eye-blink artefacts are typically observed in frequencies lower than 4Hz, as humans rarely blink more than 4 times a second. In addition, high-frequency rhythms in the brain (Gamma range) can be observed from 30Hz up to 45Hz [**22**]. Therefore, in the present study, a 5th order Butterworth band-pass filter was applied to the raw EEG signals, whilst the lower-band was set to 4Hz and the upper-band was fixed at 45Hz.

As discussed in [**19**], majority of artefact removal techniques were performed either by the use of Electrooculography (EOG) signals, or by using computationally expensive EEG artefact removal algorithms, such as those based on Least Mean Square (LMS) and Blind Source Separation (BSS). Due to the absence of EOG signal recording in this study, and the high computational expense of other artefact removal algorithms, it was decided to apply **NO** EEG artefact removal technique.

3.1.2. Normalisation

As discussed in [**19**], almost half of the reviewed studies employed a normalisation technique. However, half of the studies, which employed normalisation techniques, normalised the recorded **raw signals**, while the rest normalised the **extracted features**. Normalisation can improve the accuracy of regression or classification techniques, as it can minimise the between-participants differences, which could be present in any physiological measurements (e.g. average heart rate can vary between people, different skull thicknesses could change the power of the EEG signals [**23**], etc.). However, any normalisation technique needs to be calibrated according to a recorded dataset[11] [**19**]. Implementation of normalisation techniques can be considered as an advantage for offline classifiers, as the required features or raw signals would be extracted from the pre-recorded raw signals, before the classification process. This can provide all of the required data for the normalisation technique for the calibration process. Whereas in online classifiers the required features need to be extracted from the progressing raw signals, while the classification

[11] Such as minimum or maximum values of the signal in min-max normalisation technique, α and β in log-transformation algorithm, etc.

is in progress. This issue can turn the normalisation leverage to a disadvantage, as the normalisation calibration process could perform inadequately in the absence of the entire required dataset.

In the present study, we decided to avoid feature normalisation techniques. However, before performing a spectral analysis on the raw signals (Fast Fourier Transform (FFT) – refer to Section 3.3.1), we applied the ***z*-score normalisation technique** [**19**] within EEG channels, to standardise the spectral analysis. As an illustration, the thickness of one part of a participants' head may vary from another, or the channels' signal contact quality[12] may slightly vary. These small changes could result in considerable variation between the signal powers of the EEG channels [**23**]. Therefore, applying normalisation on the raw signals can standardize the power spectrum comparison between channels, as the overall power of each normalised channel could be between around -3 and 3 [**19**], regardless of the corresponding skull thickness and signal contact quality.

3.2. Windowing

To extract the affective features, a portion (called a *Window*) of the corresponding physiological signal is extracted and analysed. Any affective feature, extracted from this portion of the physiological signal, has to be able to be confidently tagged by a specific emotional experience. The emotionally labelled affective features, extracted from this period, are employed as a single observation, within the affective database, for the emotion recognition training process. As discussed in [**19**], various window lengths have been used in different studies. However, they have been either shorter, or longer than, or equal to the stimuli length. Moreover, according to [**19**], different affective stimuli with various durations have been employed in the studies reviewed by the authors and, as a consequence, it has not been possible to use the available literature to define the most appropriate window length, with respect to the affective stimuli duration.

3.2.1. Window Length

As the majority of the studies have employed window lengths that are either equal to, or shorter than the stimuli duration [**19**], we decided to disregard the post- or pre-stimuli duration and, thus, to avoid any window length longer than the stimuli durations. To assess the efficiency of different window lengths, we selected 28 arbitrary durations according to two different algorithms:

1. **Fixed Duration:** In this method, the duration of windows would be a **fixed value** in all sessions, for all participants, regardless of the stimuli duration. In this method we arbitrarily selected the following (17) fixed window durations: 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 100, 150 and 200 seconds (if applicable; i.e. the window length is not longer than the game duration).
2. **Relative Duration:** In this method, the duration of the windows would be calculated in every session, independently, according to the session duration. To do so, a global **relative value** was selected (as a percentage of the stimuli duration), to enable the system to calculate the window length according to the duration of the stimuli. In this method we arbitrarily selected the following window (11) relative durations: 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 60%, 80% and 100%.

All relative durations would be shorter than the stimuli duration, except the 100% window length, which behave as a window with the entire stimuli duration. Therefore, in this study, both windowing techniques, with durations equal to and shorter than stimuli length, have been implemented and evaluated.

3.2.2. Window Type

To perform spectral analysis on the signals, we employed a Fast Fourier Transform (FFT) technique (Section 3.3.1). One of the hypotheses of the FFT analysis technique is the periodicity of the target signal [**24**]. However, the recorded physiological signals are not periodic waves. Applying FFT on non-periodic signals would cause a *Spectral Leakage* effect, which results in non-zero spectral powers in high frequencies, which may not belong to the original signal [**25**]. To eliminate this effect, **weighting window functions** can be applied to the signal before FFT analysis takes place [**25**]. If the weighting window function is made of N elements, there are N coefficient weights (called W vector) for all corresponding elements within the window. In the present study, two weighting window functions are implemented and evaluated:

1. **Hamming Window:** This window multiplies the values closer to the centre of the window by a coefficient close to one, and the values closer to the edges by a weight closer to zero. Equation 1 presents the Hamming window coefficients formula [**25**].

$$W(n) = 0.54 + 0.46\,cos\left(2\pi\frac{n}{N}\right), 0 \le n \le N$$

Equation 1 – Hamming Window Coefficient Formula

2. **Tukey Window:** This window is similar to a rounded-edge rectangular

[12] The signal quality of the channels depends on the wetness of the EEG electrodes foams. Although the appropriate preparations have been conducted, at the beginning of each experiment, to ensure proper channels' signal quality; it cannot be stated, with confidence, that all channels are recorded with exactly equal contact quality (i.e. exactly equal wetness).

window, with a parameter r, which can be tuned between 0 and N. Equation 2 presents the Tukey window coefficients formula [**25**]. In the present study, we fixed the value of r at $\frac{N}{2}$.

$$W(n) = \begin{cases} 0.5\left(1 + cos\left(\frac{2\pi}{r}\left[n - \frac{r}{2}\right]\right)\right), & 0 \le n \le \frac{r}{2} \\ 1, & \frac{r}{2} \le n \le N - \frac{r}{2} \\ 0.5\left(1 + cos\left(\frac{2\pi}{r}\left[n - N + \frac{r}{2}\right]\right)\right), & N - \frac{r}{2} \le n \le N \end{cases}$$

Equation 2 – Tukey Window Coefficient Calculation Formula

3.2.3. *Windows Overlap*

As explained above (Section 3.2.2), the signal values are attenuated, due to the window coefficient weights, before spectral analysis[13]. Therefore, by applying non-overlapped windows, almost 50% and 30% of the signal values, passed through Hamming and Tukey windows, respectively, would be attenuated by 50%. Consequently, this significant attenuation could result in considerable database signal loss. To resolve this issue, overlapping windows are employed to share the attenuated signal points with other windows.

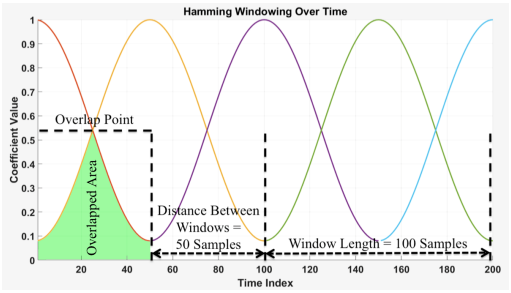

Figure 2 – 50% Overlapped, 100-Sample Hamming Windows, Over a 200-Sample Signal

Figure 2 presents five Hamming windows, with 50% overlap, with 100 samples, over 200 data points. The overlapping areas are the most attenuated data points in the entire windowed signal. In the present study it was decided to avoid any maximum attenuation, larger than (around) 5%, at all signal points. To achieve that, the Hamming window was shifted every $\frac{N}{6}$ samples, to create 83.34% overlap, with about 5.5% maximum attenuation; and the Tukey window was shifted every $\frac{N}{2}$ samples, to create 50% overlap, with about 0.5% maximum attenuation.

3.2.4. *Windowing Parameters*

As discussed in Sections 3.2.1, 3.2.2 and 3.2.3, there are two parameters in the windowing process; (1) **Duration** and (2) **Type**. For duration, there are 28 arbitrary choices, which have been implemented in this study; (I) fixed (we used 17 arbitrary lengths) and (II) relative (we used 11 arbitrary values). There are two window types that have been used in this study; (I) Hamming and (II) Tukey windows. Therefore, the combinations can create 56 different windowing processes ($2\times(17+11)=56$). The optimum selection of these parameters is investigated in Section 5.1.

3.3. Training Features Matrix

The training features matrix is an n by m matrix, where n represents the number of observations (windows) and m signifies the number of features. Each row represents m features ($F_i = [f_{i1} \ \dots \ f_{im}]_{1\times m}$, i^{th} observation), extracted from a single window, with observed output, y_i. Equation 3 presents the relationship between the training feature matrix and the predicted outputs. Function g (classifier, regression function, etc.) predicts $\hat{y}_i$, at any given point, given the corresponding features matrix F_i [**26**]. To construct the training affective features matrix, four steps have been followed in the present study:

1. The essential affective features, for all physiological measurements have been identified and extracted, for each window (F_i - in total 743 features per window –Sections 3.3.2, 3.3.3, 3.3.4 and 3.3.5).
2. Each row of the features matrix is time-stamped with its corresponding window centre time.
3. Each row of the features matrix is tagged by the affective rating, self-reported by the participant at the end of each game (Section 0).
4. The defected rows of the features matrix have been identified and deleted from the final training features matrix (Section 3.3.7).

$$g\left\{\begin{bmatrix} f_{11} & \cdots & f_{1m} \\ \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nm} \end{bmatrix}_{n\times m}\right\} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}_{n\times 1}$$

Equation 3 – Training Features matrix Vs. Outputs Function

3.3.1. *Spectral Analysis*

In this study, the Fast Fourier Transform (FFT) technique [**24**] was employed, for the extraction process of all spectral features. To extract any given frequency bandwidth power, four calculation techniques have been implemented.

1. **Power Summation:** In this technique, the power of all frequency samples within the required frequency bandwidth (from frequency fr_1 to fr_2), are added, to generate the overall power summation (Equation 4).

[13] 50% of the signal values in Tukey window, with $r = \frac{N}{2}$, are attenuated with values smaller than 1, and only 1 point (window centre value) in the Hamming window would not be attenuated.

This measurement derives the simple accumulative power, within a particular frequency bandwidth.

$$Power\ Summation = \sum_{fr=fr_1}^{fr_2} Power_{fr}$$

Equation 4 – Power Summation Equation

2. Power Ratio: In this technique the squared power of all frequency samples within the required bandwidth (from frequency fr_1 to fr_2) are added, and then divided by the accumulated squared power of all frequency bandwidths in the signal (from frequency fr_{min} to fr_{max}), to generate the overall power ratio (Equation 5). This measurement derives a fractional power unit, within a particular frequency bandwidth, with respect to all other bandwidths.

$$Power\ Ratio = \frac{\sum_{fr=fr_1}^{fr_2}(Power_{fr})^2}{\sum_{fr=fr_{min}}^{fr_{max}}(Power_{fr})^2}$$

Equation 5 – Power Ratio Equation

3. RMS Power: In this technique the Root Mean Square (RMS) power of all frequency samples within the required bandwidth (from frequency fr_1 to fr_2) is calculated (Equation 6). N is the number of frequency samples, available within the corresponding bandwidth.

$$RMS\ Power = \sqrt{\frac{\sum_{fr=fr_1}^{fr_2}(Power_{fr})^2}{N}}$$

Equation 6 – RMS Power Equation

4. RMS Power Ratio (db): In this technique, the logarithmic measure of the Root Mean Square (RMS) power is calculated. To do so, the RMS power of the required frequency bandwidth (from frequency fr_1 to fr_2) is calculated. Then, the logarithmic measure of this value, divided by the RMS power of all frequency bandwidths (from frequency fr_{min} to fr_{max}), is calculated. N and M are the number of frequency samples, available within the corresponding bandwidth and all bandwidths, respectively (Equation 7). The unit in this measurement is the decibel (db). This measurement derives a normalised power unit of the RMS power of a frequency bandwidth, respect to the RMS power of all frequency bandwidths[14].

$$RMS\ Power\ Ratio = 10log_{10}\left(\frac{\sqrt{\frac{\sum_{fr=fr_1}^{fr_2}(Power_{fr})^2}{N}}}{\sqrt{\frac{\sum_{fr=fr_{min}}^{fr_{max}}(Power_{fr})^2}{M}}}\right)$$

Equation 7 – RMS Power Ratio (db) Equation

3.3.2. *Participant-Related Features*

In total, three features, related to the participant, have been extracted, in each window. These are the **gender** (male vs. female), **hand preference** (right *vs*. left handed) and **age** (four classes: 12-18, 18-24, 24-30 and 30-40 years old), each of which has been recorded within the features matrix.

3.3.3. *EEG Features*

Table 1 presents all features extracted from the EEG signals. 13 out of 14 EEG features, presented in [**19**], have been extracted from the EEG signal within each window. Only the event-related potentials (EPR), reviewed in [**19**], have not been employed in this study. This was due to the fact that the affective stimulations were presented during the game period, whereas EPR features need to be extracted according to specific stimuli presentation instances. In addition to the 13 EEG features introduced in [**19**], the $Alpha - Beta\ Ratio$ measurement, presented in Equation 8, has been implemented in this study. According to [**27**], Alpha waves can indicate a relaxed awareness, without any attention or concentration, whereas Beta waves can be associated to active thinking, active attention or solving concrete problems. Therefore, this ratio can indicate an "attention measure" in a location of the brain (large $Alpha - Beta\ Ratio$ indicates high alpha activities and lower beta activations, signifying lower attention and concentration, and vice versa).

$$Alpha - Beta\ Ratio = \frac{Beta\ Bandwidth\ Power}{Alpha\ Bandwidth\ Power}$$

Equation 8 – Alpha-Beta Ratio Equation

All spectral powers, in all frequency ranges (theta, slow-alpha, alpha, beta and gamma), have been extracted four times, using one of the power calculation formulae presented in Section 3.3.1 (Equation 4, Equation 5, Equation 6 and Equation 7). Therefore all related measurements (e.g. asymmetric ratio[15], EEG_w[16], etc.) have been calculated four times, using all four power calculation formulas. Consequently, in total, 707

[14] The Log-Transformation, discussed in [**19**], while $\beta = 10$ and $\alpha = -10log_{10}\left(\sqrt{\frac{\sum_{fr=fr_{min}}^{fr_{max}}(Power_{fr})^2}{M}}\right)$ as a reference point.

[15] Asymmetric Spectral Power Density $= \frac{P_{left}-P_{right}}{P_{left}+P_{right}}$, While "P" is the spectral power in either "Alpha" or "Slow-Alpha" frequency rhythms [**45**], [**46**].

[16] $EEG_w = \log\left(\frac{\sum_{i=1}^{N}\beta_i}{\sum_{i=1}^{N}(\theta_i+\alpha_i)}\right)$, While "N" is the number of channels. θ, α and β are Theta, Alpha and Beta frequency rhythms, respectively [**29**].

features have been extracted from all fourteen single and seven paired channels of EEG signals.

3.3.4. GSR Features

Table 3 presents all features extracted from the GSR signals. All GSR features, presented in [**19**], have been extracted from the raw GSR signal, within each window. In addition, three extra features have been introduced in the present study and have been extracted from the GSR signal: (1) mean of the positive values in the GSR first derivative, as well as negative values (employed by [**28**], [**29**], [**30**]) have been extracted; (2) mean of the GSR first derivative's peak values (local maxima), as well as the average peaks of the original signal (employed by [**28**], [**30**]) have been recorded; (3) the GSR fluctuation frequency has also been extracted, using Equation 9 ($X'(i)$ is the i^{th} element of the first derivative of the GSR signal; and $sign(a)$ presents the sign function). The fluctuation frequency signifies the number of times the signal changes direction (i.e. increase to decrease and vice versa).

$$Fluctuation\ Frequency = \sum_{i=0}^{N-1} [sign(X'(i)) \neq sign(X'(i+1))]$$

Equation 9 – Signal Fluctuation Frequency Equation

The spectral power has been extracted four times, each time by using one of the power calculation formulas presented in Section 3.3.1 (Equation 4, Equation 5, Equation 6 and Equation 7). Consequently, in total, 14 features have been extracted from the raw GSR signals.

3.3.5. Heart Rate Features

Table 2 presents all features extracted from the heart rate signals. Seven out of eight heart rate features, presented in [**19**], have been extracted from the raw heart rate signal, within each window. The sampling frequency of the Shimmer+ device (512Hz) did not provide the appropriate frequency resolution required for extracting low frequency spectral power[17] from the heart rate signal. Six additional features have also been implemented in this study and were extracted from the heart rate signal: (1) minimum and (2) maximum heart rate values within a window are extracted; (3, 4) mean of both positive and negative values of the first derivative of the heart rate signals are also recorded; (5) the mean of the peak values (local maxima) of the first derivative of the heart rate is extracted from each window; (6) the heart rate fluctuation frequency is also obtained, using the algorithm presented in Equation 9 (Section 3.3.4).

The spectral powers have been extracted four times, using one of the power calculation formulae presented in Section 3.3.1 (Equation 4, Equation 5, Equation 6 and Equation 7). Consequently, in total, 22 features have been extracted from the raw Heart Rate signals.

3.3.6. Affective Tagging

Each affective feature vector F_i (each row of the features matrix, extracted from a single window) has to be tagged by the corresponding emotion (y_i), which has been experienced by the participant. As discussed in Section 2.2, participants were asked to self-assess their average emotional experience during each game, using both dimensional and categorical models. The dimensional ratings (using Valence, Arousal and Dominance axes) are converted into one of the four Affective Clusters[18] (PVLA, PVHA, NVPA and NVNA – for more information refer to [**19**]). As the self-assessments are conducted at the end of each game, rather than continuously during the gameplay, the below hypothesis has been presented in this study. According to this hypothesis, we divided the emotional experience of the participants, during a single session (game), into two affective periods:

1. **'Emotion Build-Up' Period:** This period occurs during the first part of each game. Within this period, the emotional experience of the participant can be unpredictable, as it can be representative of a residual state from a previous game or some other pre-cognitive state.
2. **'Emotion Persistence' Period:** This period occurs during the last part of each game. Within this period, the emotional experience of the participant has been influenced by the current game, and can be (reasonably) confidently labelled by an affective cluster or label. This means that all emotional experience variations within this period are considered as minimal. This also means that the affective experience of the participants within this period is always close to the average affective label and cluster, reported by the participants at the end of the game.

Then, we hypothesised that the first 30% duration of each game constitutes the Emotion Build-Up Period, while the last 70% can be considered as the Emotion Persistence Period. As the emotional experience of the participants can be unpredictable during the Emotion Build-Up period, all windows, which have a centre time-stamp within the first 30% period of the each game, have been deleted from the features matrix. Then, all windows, which have a centre time-stamp within the last 70% period of the each game, have been tagged by the Affective Cluster and Emotion Label reported by the participants, at the end of that game.

[17] 0.01Hz to 0.04Hz as presented in [**28**], [**30**], [**47**], [**48**], [**49**], [**50**].

[18] The cluster is determined according the dimensional ratings of the participants at the end of each game and the cluster boundaries, presented in Section [**19**].

Table 1 – Extracted EEG Features List – All Features Were Extracted From Each Window

Feature	Description	Feature	Description
14 Single Channels Theta Rhythms	4Hz to 8Hz Frequency Range Power	7 Paired Channels[19] Signal Quality	As Reported by EPOC EMOTIV, in 5 Classes
14 Single Channels Slow-Alpha Rhythms	8Hz to 10Hz Frequency Range Power	7 Asymmetric Power Ratio[15] Using Slow-Alpha Rhythms	7 Symmetric Channel Pairs
14 Single Channels Alpha Rhythms	8Hz to 13Hz Frequency Range Power	7 Asymmetric Power Ratio[15] Alpha Rhythms	7 Symmetric Channel Pairs
14 Single Channels Beta Rhythms	14Hz to 26Hz Frequency Range Power	Left Frontal EEG_w[16]	AF3, F3, F7 and FC5 Channels
14 Single Channels Gamma Rhythms	30Hz to 45Hz Frequency Range Power	Right Frontal EEG_w[16]	AF4, F4, F8 and FC6 Channels
14 Single Channels[20] Signal Quality	As Reported by EPOC EMOTIV, in 5 Classes	Left Parietal EEG_w[16]	P7 and O1 Channels
7 Paired Channels[21] Theta Rhythms	4Hz to 8Hz Frequency Range Power	Right Parietal EEG_w[16]	P8 and O2 Channels
7 Paired Channels[21] Slow-Alpha Rhythms	8Hz to 10Hz Frequency Range Power	Frontal EEG_w[16]	AF3, AF4, F3, F4, F7, F8, FC5 and FC6 Channels
7 Paired Channels[21] Alpha Rhythms	8Hz to 13Hz Frequency Range Power	Parietal EEG_w[16]	P7, P8, O1 and O2 Channels
7 Paired Channels[21] Beta Rhythms	14Hz to 26Hz Frequency Range Power	Overall EEG_w[16]	All 14 Channels
7 Paired Channels[21] Gamma Rhythms	30Hz to 45Hz Frequency Range Power	7 Signal Quality for all Measurements[22]	As Reported by EPOC EMOTIV, in 5 Classes

Table 2 – Extracted Heart Rate Features List – All Features Were Extracted From Each Window

Feature	Description	Feature	Description
Mean	Average of the Heart Rate Signal	Mean of the Negative Values in the First Derivative	Average of the Negative Values of the First Derivative of the Heart Rate Signal
Minimum	Minimum Value of the Heart Rate Signal	Mean of The First Derivative Peaks	Average of the Peak Values (Local Maximums) of the First Derivative of the Heart Rate Signal
Maximum	Maximum Value of the Heart Rate Signal	Heart Rate Medium Frequency Spectral Power	0.04Hz to 0.15Hz Frequency Range Power
Standard Deviation	Standard Deviation of the Heart Rate Signal	Heart Rate High Frequency Spectral Power	0.15Hz to 0.5Hz Frequency Range Power
Mean of The Peaks	Average of the Peak Values (Local Maximums) of the Heart Rate Signal	Heart Rate Spectral Power Ratio[23]	Fractional Ratio of the Heart Rate Medium and High Spectral Power
Mean of The First Derivative	Average of the First Derivative of the GSR Signal	Fluctuation Frequency	Frequency of the GSR Signal Direction Changing
Mean of the Positive Values in the First Derivative	Average of the Positive Values of the First Derivative of the GSR Signal	–	–

[19] Average signal quality of two symmetric channels, within the corresponding window
[20] Average channel's signal quality, within the corresponding window
[21] Voltage subtraction of two symmetric channels
[22] Average signal quality of all target channels, within the corresponding window
[23] $\text{Heart Rate Spectral Power Ratio} = \frac{\text{Medium Frequency Spectral Power}}{\text{High Frequency Spectral Power}}$

Table 3 – Extracted GSR Features List – All Features Were Extracted From Each Window

Feature	Description	Feature	Description
Mean	Average of the GSR Signal	Mean of the Positive Values in the First Derivative	Average of the Positive Values of the First Derivative of the GSR Signal
Minimum	Minimum Value of GSR Signal	Mean of the Negative Values in the First Derivative	Average of the Negative Values of the First Derivative of the GSR Signal
Maximum	Maximum Value of GSR Signal	Mean of The First Derivative Peaks	Average of the Peak Values (Local Maximums) of the First Derivative of the GSR Signal
Standard Deviation	Standard Deviation of the GSR Signal	GSR Low Frequency Spectral Power	0Hz to 2.4Hz Frequency Range Power
Mean of The Peaks	Average of the Peak Values (Local Maximums) of the GSR Signal	Fluctuation Frequency	Frequency of the GSR Signal Direction Changing
Mean of The First Derivative	Average of the First Derivative of the GSR Signal	–	–

3.3.7. Defective Data Removal

All windows exhibiting EEG signals with an average signal quality below "fair" (according to the EPOC EMOTIV signal quality classes[10]) have been removed from the features matrix. Furthermore, all windows exhibiting infinity or NAN (Not-A-Number) values have also been removed from the features matrix.

3.4. Features Selection

3.4.1. Minimal-Redundancy-Maximal-Relevance (mRMR)

As discussed in Section 3.3, a total of 743 features have been extracted from the raw physiological signals. To be able to perform emotion classification, the dimension of the features matrix has to be reduced to a subspace. This subspace has fewer features (labelled *Most Optimum Features* throughout the paper), while they can adequately capture the essence of the data [**26**]. To perform the feature selection, the minimal-redundancy-maximal-relevance (mRMR) technique has been employed. Consider the features matrix of $F \in \mathbb{R}^{N \times D}$, while N is the number of observations and D is the number of features. The mRMR algorithm finds the most optimum subset $F_S \in \mathbb{R}^{N \times d}$, such that $d \ll D$, and F_S can optimally characterise F [**31**].

The mRMR algorithm employs *Shannon's Entropy* [**32**] to identify those features, which are mutually exclusive with respect to each other (minimal redundancy), whilst remaining mutually inclusive with respect to the classification clusters (maximal relevance – Affective Clusters or Emotion Labels in this study) [**31**]. To perform the analysis, the database has to be discretised prior to the Shannon's Entropy calculations. Therefore, all features were discretised according to 3 classes (-1, 0 and 1), with respect to the features' mean and standard deviation values[24] [**31**].

3.4.2. Feature Selection Parameters

In the present study, 30 arbitrary values have been used as the number of required features (d – 1 to 30), each of which could be selected according to either Affective Clusters or Emotion Labels. Furthermore, the mRMR technique can produce various lists of 30 Most Optimum Features, according to different windowing techniques employed in the features matrix construction process (28 different window lengths for each windowing technique, Hamming and Tukey – Section 3.2.4). This combination can create 1680 different settings ($2 \times 28 \times 30 = 1680$), for classification according to either Affective Clusters or Emotion Labels[25].

[24] $Discretised\ F_i = \begin{cases} 1 & F_{ij} > mean(F_i) + std(F_i) \\ 0 & mean(F_i) - std(F_i) \le F_{ij} \le mean(F_i) + std(F_i) \\ -1 & F_{ij} < mean(F_i) - std(F_i) \end{cases}$, while "mean" and "std" are the average and standard deviation of F_i, respectively [**31**].

[25] E.g. (1) most optimum feature for the 2-second Hamming window, (2) 2 most optimum features for the 2-second Hamming window, … , (1680) 30 most optimum features for the 100% Tukey window.

4. Classification and Affect Recognition

4.1. Classification Techniques

As discussed in [**19**], Support Vector Machine (SVM) [**33**], Discriminant Analysis (DA) [**34**] and Classification Trees [**35**] have been employed by the majority of the studies, reviewed. In the present study, we evaluated the performance of these classifiers (SVM, DA and Classification Tree), plus the K-Nearest Neighbour (KNN) classifier [**36**], in the affect recognition process. All classifications and cross-validations have been implemented within MATLAB software (version R2015b), using the Statistics and Machine Learning Toolbox[26].

4.1.1. Support Vector Machine (SVM)

Support Vector Machines are supervised classification and regression methods, originally designed for binary classifications, but with the capability for extension to be implemented in multi-class and regression applications. SVM classifier is a Kernelized algorithm, which attempts to cluster a feature space according to a number of known labels, with maximum possible distance between the clusters' borders, by using a kernel function [**33**]. There are various Kernel functions that could be implemented in SVM classification algorithms [**37**]. In the present study, the Linear, 2nd Order Polynomial (Quadratic), 3rd Order Polynomial (Cubic) and Gaussian Kernel functions have been implemented and evaluated in the SVM classification. Twenty-four arbitrary Kernel Scales, for the Gaussian Kernel function, have been selected and evaluated in the cross-validation process (0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 1, 1.4, 2, 3, 4, 5 and 5.7).

4.1.2. Discriminant Analysis (DA)

Discriminant Analysis is a supervised discrimination (classification) algorithm, which categorises feature spaces into binary or multi-class clusters [**34**]. In the present study, Linear (LDA) and Quadratic (QDA) Discriminant Analyses have been implemented and evaluated in the affect recognition process.

4.1.3. Classification Trees

Classification Trees (also known as Decision Trees) are supervised classification algorithms which are defined by separating and partitioning a feature space, using multiple rules (called *Splits*), and defining a local model, into which feature spaces can be categorised as binary or multi-class clusters. The number of Splits in a tree can be defined as its complexity level. As an illustration, a tree with 50 Splits is five times more complex than a tree with only 10 Splits [**35**]. In the present study, twenty different arbitrary Splits numbers have been selected and evaluated in the cross-validation process (5 to 100 with step size of 5).

4.1.4. K-Nearest Neighbour (KNN)

K-Nearest Neighbour is a supervised classification algorithm, which categorises feature spaces into binary or multi-class clusters. To do so, the algorithm employs a training dataset to classify further data points according to the K closest data points in the training dataset (K nearest neighbours – using Euclidean distance[27]) [**36**]. In the present study, 30 different arbitrary K values have been implemented and evaluated in the cross-validation process (1 to 30).

4.2. Hyper-Parameters Tuning

As explained in Section 3.2.4, 56 different windowing settings can be applied (each of which with different length and type) to generate 56 various feature matrices. The feature selection algorithm, presented in Section 3.4, is capable of identifying different sets of the most optimum features, by employing different feature matrices constructed from various window lengths and types, to classify the physiological affective space. On the other hand, the feature selection algorithm, presented in Section 3.4, can generate 30 feature sets, each of which containing between 1 and 30 of the most optimum features. This will result in 1680 different training matrices for the classification process (Section 3.4.2). By employing the Affective Clusters (4 clusters – dimensional assessment) and Emotion Labels (8 labels – categorical assessment), the classifiers could categorise the physiological responses into four or eight classes, respectively (Section 2.1.1).

The performance of the classifiers, trained according to each training matrix, vary in terms of the classification accuracy. These variables are the *hyper-parameters* of the affect recognition system. Hyper-parameters are the restrictions of learning algorithms, which can be tuned prior to the training process, resulting in various classification performances [**38**]. The process, in which the best set of hyper-parameters, which can produce the best classification performance, is identified, is called *hyper-parameters tuning* [**38**]. In machine learning algorithms, cross-validation is mainly used as a measure in the hyper-parameters tuning process. However, various searching algorithms (e.g. grid or random search) are employed in addition, in order to identify the best set of hyper-parameters, which generate the best classification accuracy [**38**].

[26] http://uk.mathworks.com/help/stats/index.html

[27] $Euclidean\ Distance = \sqrt{\sum_{i=1}^{D}(f_{1i} - f_{2i})^2}$, while f_{1i} and f_{2i} are the i^{th} features of the 1st and 2nd points respectively. And "D" is the number of features, employed to present each point in the affective space.

It is almost impossible to train each classifier according to all possible hyper-parameter variations, as there are infinite combinations (considering all possible window lengths, rather than a finite number of arbitrary selections). Therefore, by selecting the arbitrary window lengths (Section 3.2.4), the number of required features (Section 3.4.2) and the classification settings (Section 4.1.1, Section 4.1.2, Section 4.1.3 and Section 4.1.4), the number of possible hyper-parameters tuning variations have reached to 132,720 [28] for each classification process (according to either Affective Clusters or Emotion Labels – 265,440 in total). According to [**38**], this small subset of the infinitely larger settings space may be sufficient in the hyper-parameters tuning process, as the majority of the hyper-parameters variations do no matter much, as only those which result in high accuracy actually matter. To assess the performance of different classifications settings, the accuracies of the classifiers have been estimated through a 10-Fold (random folding) Cross-Validation technique [**39**]. To train and cross-validate the performance of all 265,440 classifiers, a processing farm service (HPC[29]) was employed to speed up the process.

To perform the comparison, the scatter plots of the classification accuracies, for each setting, have been analysed. As an illustration, the scattered dots in Figure 3 define different classifiers with various settings. For example, if the classifier employs five features for the classification, different window types (Hamming *vs*. Tukey), window lengths (28 different window lengths), classifier settings (different K-value in KNN, etc.), and so on, can result in various accuracies (all scattered dots presented in a vertical manner for five features). However, as the best performing classifier in each setting has to be selected the setting, which generates the maximum classification accuracy is identified and highlighted (e.g. the line, highlighting the maximum values in Figure 3). The analyses of different hyper-parameters variations are presented in Sections 4.3, 4.4 and 4.5. The best performing classifiers (with the highest accuracy) have been identified and presented in Section 5.1.

4.3. Number of Features Evaluation

Figure 3 presents the performance of the classifiers with respect to the different number of features. As it can be obtained by the graph, the DA performance has not been changed considerably, by employing more or less features, whereas employing more features has increased the accuracy of the Classification Tree. The accuracy of both KNN and SVM classifiers, with respect to the number of employed features, follows a Sigmoid function[30] pattern. This means that their accuracies have increased by employing more features, with saturation occurring around 98%. As it can be seen in the graphs, the accuracy of KNN and SVM classifiers has increased around 0.6% by increasing the number of features from 20 to 30. By increasing the number of features, the complexity of the classifier grows, which consequently increases the classifier's processing and timing expense. Therefore, we decided to not to employ more than 30 features in the classification process.

4.4. Windowing Settings Evaluation

As discussed in Section 3.2.4, there are two tuning parameters for the windowing process; window type (Hamming *vs*. Tukey) and length (17 fixed *vs*. 11 relative). As shown in Figure 4, the performances of KNN, SVM and Classification Tree are slightly better when using the Hamming window, compared to the Tukey window. The DA classifier performance did not change considerably by employing either the Hamming or Tukey window.

Figure 5 presents the classification accuracies of the classifiers, respect to different **fixed** window lengths. Figure 6, on the other hand, presents the classification accuracies of the classifiers, with respect to different **relative** window lengths. As it can be seen in the graphs the performance of KNN, SVM and Classification Tree classifiers have been increased by employing shorter windows (in both fixed and relative windowing techniques). However, the DA performance has not been changed considerably by using different window lengths.

4.5. Classifiers Settings Evaluation

As discussed in Section 4.1, each classifier can be tuned according to a parameter. As can be seen in Figure 7, the performance of the KNN classifier is slightly attenuated, whilst "K" is increased. This means that the KNN classifier performs better when considering fewer neighbours in the affective space, in its attempts to classify the affective features. According to this analysis the 1st Nearest Neighbour (K=1) has the highest accuracy, compared to other "K" values. In contrast, as shown in Figure 7, the accuracy of the Classification Tree is boosted, while the number of Splits is increased. This means that the Classification Tree performs better, if more conditions are defined and a more complex tree is generated. According to the analysis, the Classification Tree has its maximum accuracy if 100 Splits are generated in the Tree.

[28] $(1680\times27)_{SVM} + (1680\times2)_{DA} + (1680\times20)_{Classification\ Tree} + (1680\times30)_{KNN} = 132{,}720$

[29] High Performance Computing (HPC) services provided by Queen Marry University of London (QMUL): http://docs.hpc.qmul.ac.uk/

[30] $sigm(x) = \frac{e^x}{1+e^x}$, The term "sigmoid" means S-shaped [**51**].

On the other hand, the performance of the DA classifier is attenuated if a 2nd order polynomial (quadratic instead of linear) function is employed.

Figure 8 presents the performance of the SVM classifier, according to four different Kernel functions; linear, 2nd order polynomial (quadratic), 3rd order polynomial (cubic) and Gaussian function with 24 different Kernel scales. As illustrated by the graph, the performance of the SVM classifier is boosted when a higher order non-linear Kernel function is employed. The Gaussian Kernel function with relatively large kernel scales (either 2 or 3) performed better than the Linear and Quadratic Kernel functions. Although the Cubic Kernel performance was very similar to the Gaussian function, the best performing classifiers (Section 5.1) employed the Gaussian Kernel.

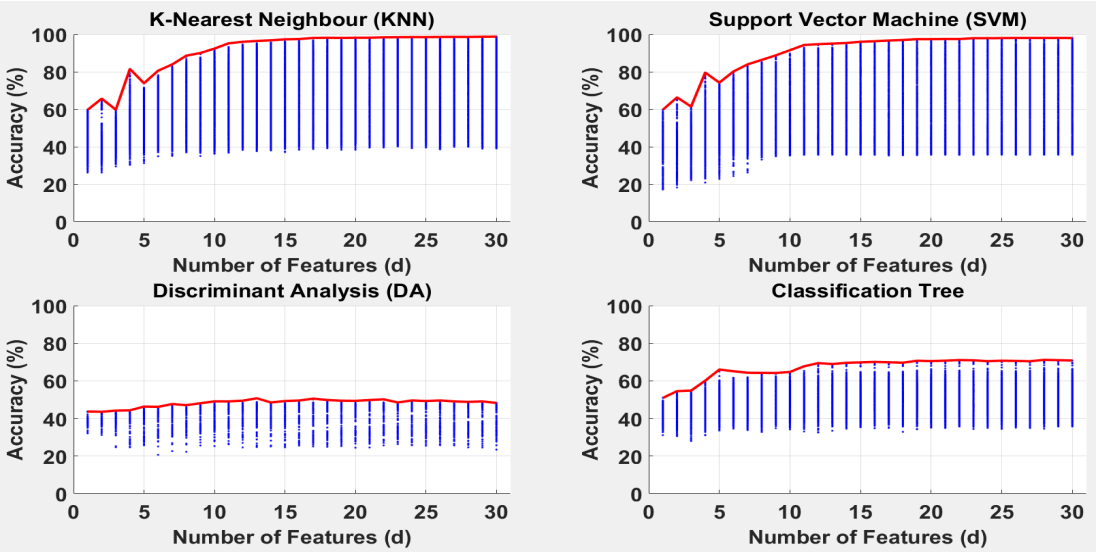

Figure 3 – Classifiers Performance Respect to Different Number of Features, According to Affective Clusters

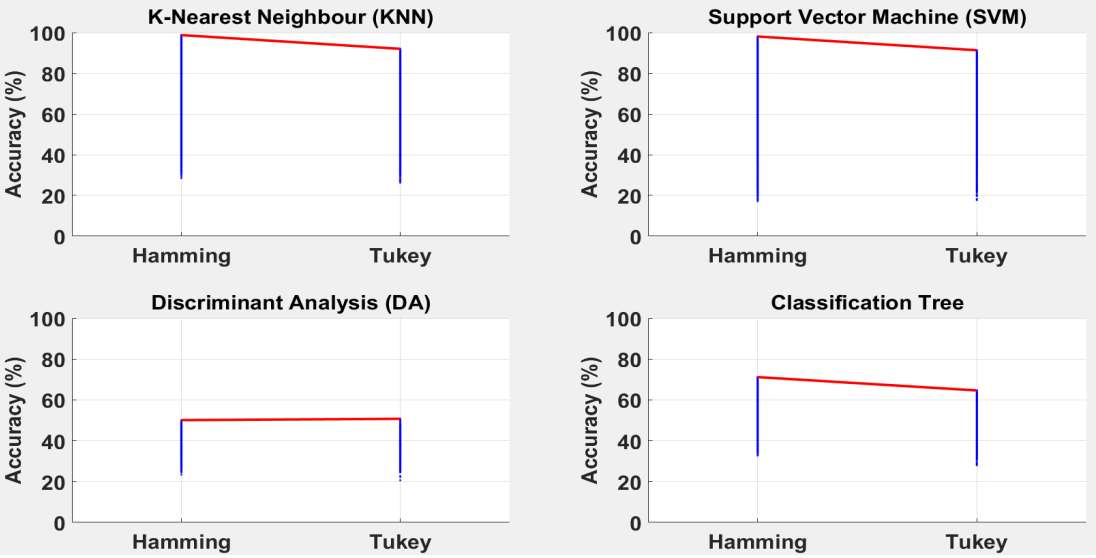

Figure 4 – Window Type Vs. Classifiers Accuracy, According to Affective Clusters

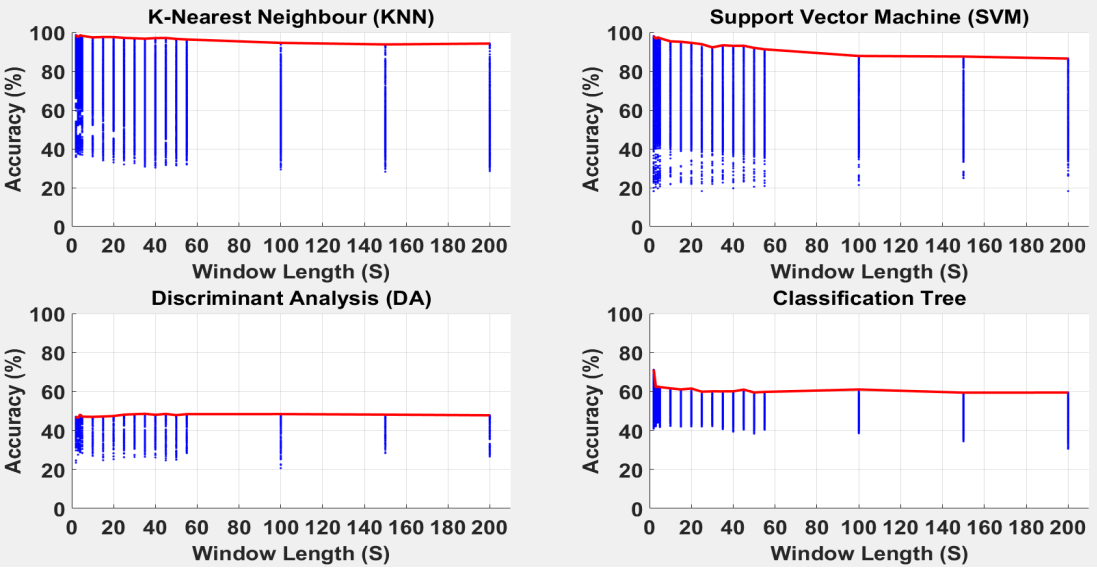

Figure 5 – Window Fixed Length Vs. Classifiers Accuracy, According to Affective Clusters

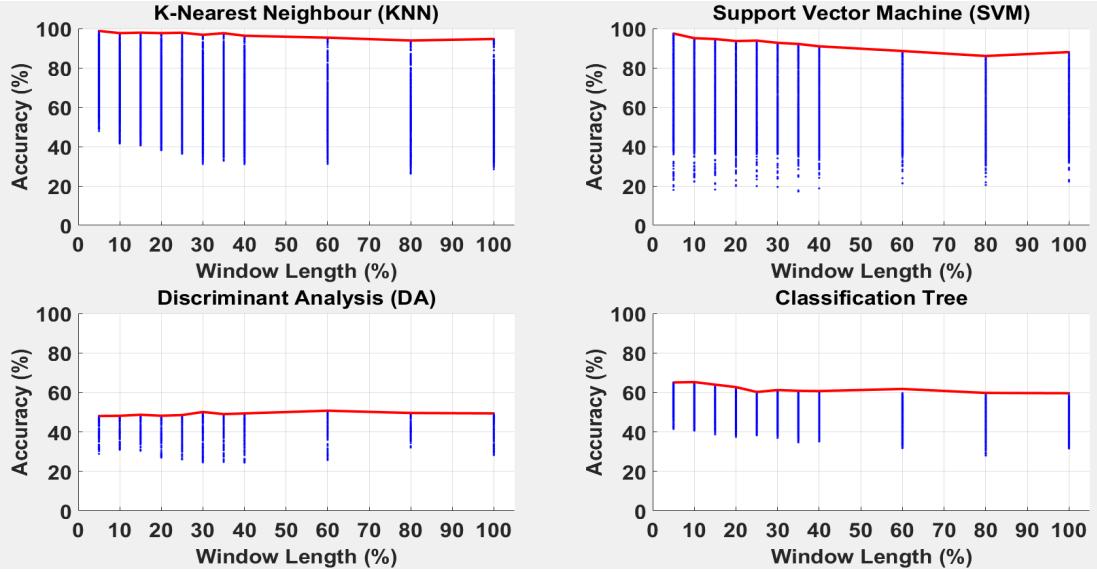

Figure 6 – Window Relative Length Vs. Classifiers Accuracy, According to Affective Clusters

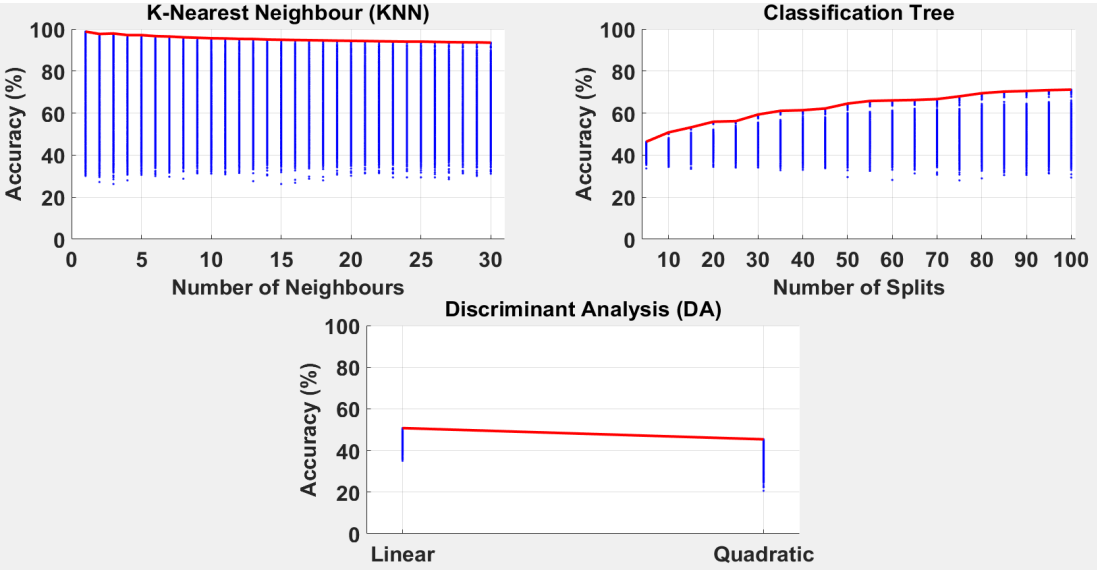

Figure 7 – KNN, Classification Tree and DA Classifiers Settings *vs.* Accuracy, According to Affective Clusters

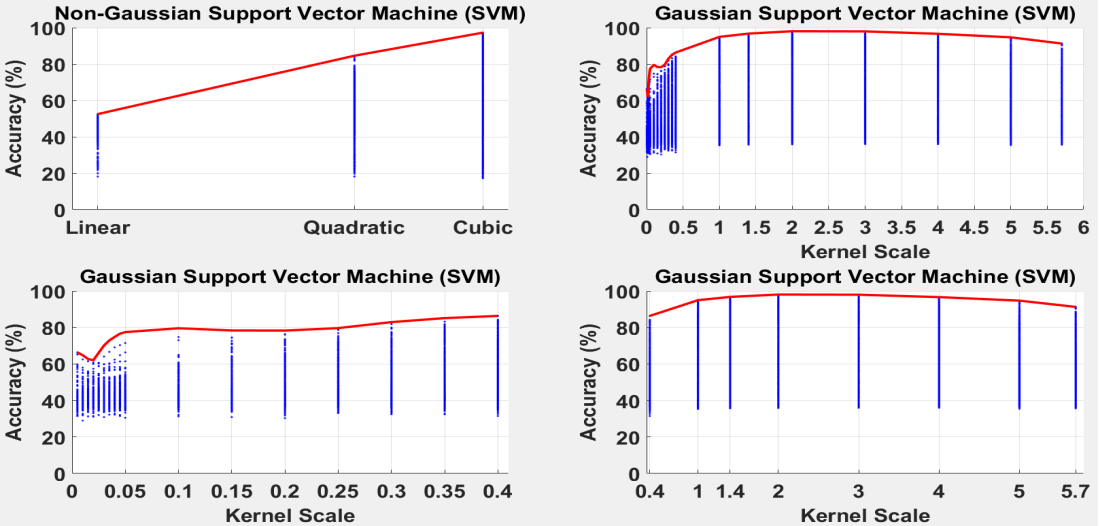

Figure 8 – SVM Classifier Settings Vs. Accuracy, According to Affective Clusters – The Right Top Graph Presents the *Kernel* Scale Between 0.005 and 5.7 – The Bottom Graphs Presents Magnified Versions, with *Kernel* Scale Between 0.005 and 0.4 (Left) and 0.4 and 5.7 (Right).

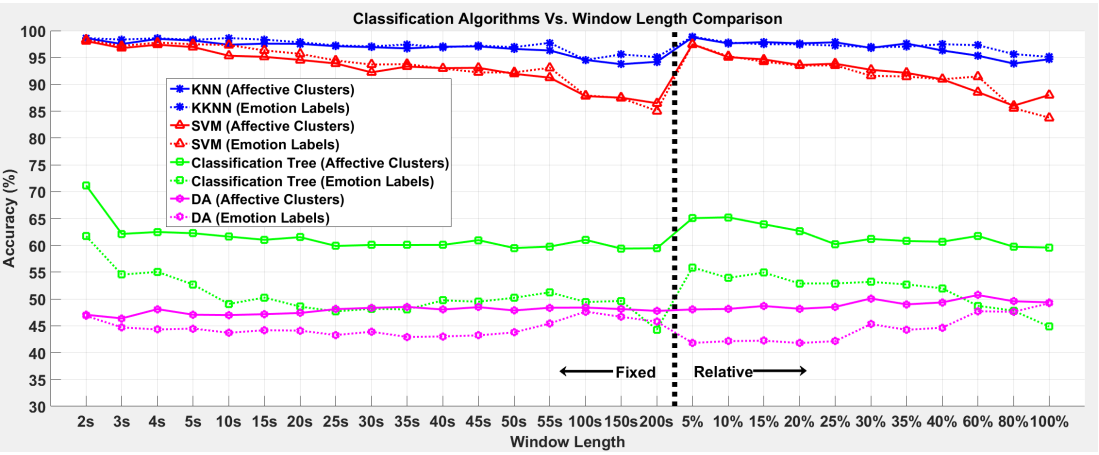

Figure 9 – Classification Methods Comparison, According to Affective Clusters and Emotion Labels – The Horizontal Axes Presents 28 Different Window Lengths; 17 Fixed (Left Side of the Vertical Dashed Line) and 11 Relative (Right Side of the Vertical Dashed Line)

5. Discussion

5.1. Best Performing Classifiers

To be able to compare the performance of all classification techniques, the best performing classifier setting (e.g. K value in KNN, etc.), for each window length, has been identified. As a result, 28 settings for each classification technique (KNN, SVM, DA and classification tree) have been identified. Figure 9 presents the best classification accuracy, for each classifier, in each window length. The horizontal axis of the figure presents 28 different window lengths; 17 Fixed (left side of the vertical dashed line) and 11 Relative (right side of the vertical dashed line). An Analysis of Variance (ANOVA)[31] showed that the different windowing techniques (fixed *vs.* relative) is not a significant factor in changing the classifications performances ($P_{Windowing}$ = 0.691). However, the performances of different classification techniques are significantly different, in terms of their classification accuracy ($P_{Classification}$ < 0.001). The KNN (96.78% (±1.42%) and 97.24% (±1.14%) mean accuracy, for Affective Clusters and Emotion Labels respectively, across all window lengths) and SVM (92.77% (±3.42%) and 92.91% (±3.97%) mean accuracy, for Affective Clusters and Emotion Labels respectively, across all window lengths) perform better than the Classification Tree (61.54% (±2.46%) and 51.06% (±3.62%) mean accuracy, for Affective Clusters and Emotion Labels respectively, across all window lengths). The DA classifier performs worst than the other three, with 48.28% (±0.96%) and 44.53% (±1.98%) mean accuracy, for Affective Clusters and Emotion Labels respectively, across all window lengths.

[31] Classifiers accuracy is considered as the dependent variables, while relative vs. fixed windowing technique and different classifiers as the independent parameters.

As there is a high and significant negative correlation between the window durations and the accuracies of the KNN and SVM classifiers ($r_{Fixed}(68) = -0.69$ and $r_{Relative}(44) = -0.61$, P<0.001 – for both Affective Clusters and Emotion Labels), one could conclude that, by decreasing the window length, the accuracy of the KNN and SVM classifiers can be improved. In the fixed windowing technique, the KNN classifier achieved its best performance by the 2-second window with 98.61% accuracy, for both Affective Clusters and Emotion Labels. The SVM classifier achieved its best performance using the same 2-second window, with 98.06% and 98.29% accuracy, for Affective Clusters and Emotion Labels, respectively. In the relative windowing technique on the other hand, the KNN and SVM classifiers achieved their best performance by the 5% window with 98.7% and 97.47% accuracy, for Affective Clusters and Emotion Labels, respectively.

5.2. Most Optimum Features

As discussed in Section 3.4, 743 features were extracted from the recorded physiological signals, while no more than the 30 most optimum features were selected, for the classification process, using the mRMR algorithm. There are 56 different windowing techniques (different length and type – Section 3.2.4), which result in 56 various sets of most optimum features. Furthermore, depending on the classification class (either Affective Clusters of Emotion Labels) provided for the mRMR algorithm, different sets of most optimum features could be nominated (Section 3.4). The mRMR algorithm guarantees to find the *d* most optimum features, which have minimum mutual information amongst each other (minimum redundancy), whilst maximum mutual information with respect to the classification classes (maximum relevance – Affective Clusters or Emotion Labels). Therefore, in total, 112 different sets of the most optimum features[32] (each of which containing 30 features – Section 3.4.2) have been identified by the mRMR algorithm.

On the other hand, to consider the performance of classifiers as well (according to the cross-validation accuracies), the sets of most optimum features, extracted from the Tukey-based windows, were excluded[33]. Therefore, 56 different sets of the most optimum features were analysed and the *unique features*, which are present in at least one of the 56 sets, were identified. This *Unique Most Optimum Features List* contains 250 features, out of the 743 features, recorded at the beginning of the process. Furthermore, by considering the best performing KNN and SVM classifiers[34] in each window, the features, which have not been employed in the classification process, have been excluded, and the number of features in the Unique Most Optimum Features List has been reduced to 230 features. This guarantees that these sets of optimum features are those, which not only preserve the maximum relevance and minimum redundancy aspect of the mRMR results, but also generate the best classification accuracies, according to both Affective Clusters and Emotion Labels.

Analysing this list can highlight the superiority of each feature against the others. To do so, the popularity of all features of the Unique Most Optimum Features List (containing the 230 features), within the 56 different sets of most optimum features[35], have been calculated. This value has been reported as a percentage, signifying the number of windows (among all 56 windows) that employed a particular feature. This analysis has not been used in the classification process, and has only been developed for appropriate presentation purposes. Table 4 presents the Unique Most Optimum Features List, grouped according to their measurement categories. The table presents the popularity percentages as well, and these signify the occurrence frequency ranges in which each feature group has been employed within a windowing and classification technique (among all 56 windows). Among participant-related features, 'Age' has been employed by all windows, as the most optimum feature, to classify the participants' emotional experiences (according to both Affective Clusters and Emotion Labels). This signifies the fact that the participants' age can provide a substantial amount of information, to classify their emotional experiences. This relationship has been addressed by other studies, as well [**40**], [**41**], [**42**], [**43**].

On the other hand, and as discussed in Section 3.3.1, four different calculation techniques have been employed in performing the spectral analysis. In Table 4 the rhythms measured with various techniques have been combined. As an illustration, the alpha rhythm asymmetric ratio for the AF3-AF4 paired electrodes has been employed by 68% of the windowing and classification techniques, whilst 17.5% measured the powers through Summation (Equation 4), 3% based on the Power Ratio (Equation 5), 1.5% according to the RMS Ratio db (Equation 7) and the remaining 46% through the RMS (Equation 6) technique. To compare the popularity percentages of the spectral

[32] 28 sets for Hamming and 28 sets for Tukey windows, for both Affective Clusters and Emotion Labels ($28 \times 4 = 112$).

[33] All classifiers with the highest accuracy (Section 5.1) employed Hamming window technique.

[34] The KNN and SVM classifiers outperformed the Classification Tree and DA classifiers, considerably (Section 5.1), so only these 2 were considered in this process.

[35] 28 Hamming windows for classification according to Affective Clusters and 28 according to the Emotion Labels.

measurements according to all 4 techniques (Equation 4, Equation 5, Equation 6 and Equation 7), an ANOVA analysis was conducted. The analysis showed that there is no statistical difference between the spectral power calculation techniques ($P_{Measurment\_Technique}$ = 0.542). One can conclude that, no spectral analysis technique is superior when compared to the others.

Table 4 – Selected Features According to the mRMR algorithm and the Best Performing Classifiers

Feature Group	Detail			Mean Popularity Percentage (25^{th} – 75^{th} Percentiles)		
Age	Participants age according to 4 classes			100%		
GSR	Minimum	Mean	Fluctuation Frequency	100%	89.29%	67.86%
Heart Rate	Maximum			55.36%		
Alpha Rhythm Asymmetric Ratio	All 7 paired channels			43.82% (21.42% - 65.62%)		
Heart Rate	Fluctuation Frequency			39.29%		
Slow-Alpha Rhythm Asymmetric Ratio	All 7 paired channels			38.52% (25.44% - 54.91%)		
EEG Gamma Rhythms	6 paired channels (Excluding F7-F8) 8 single channels (Excluding AF3, F3, F4, F7, F8, P8)			33.93% (7.14% - 58.92%)		
Heart Rate	Mean of the Peaks			30.36%		
Hand Preference	Participants dominant hand			28.57%		
EEG_w	4 paired channels (Excluding F7-F8, FC5-FC6, O1-O2) 7 single channels (Excluding F3, F4, F7, F8, FC6, T8, P8)			28.57% (8.48% - 43.75%)		
GSR	Low Frequency Power			23.21%		
Heart Rate	Medium Frequency Power			23.21%		
EEG Theta Rhythm	5 paired channels (Excluding F3-F4, T7-T8) 3 single channels (AF4, P7, O2)			22.77% (6.25% - 26.78%)		
EEG Alpha Rhythm	5 paired channels (Excluding F3-F4, P7-P8) 8 single channels (Excluding AF3, AF4, F3, F4, T7, P8)			20.19% (8.48% - 26.78%)		
Alpha-Beta Ratio	All 7 paired channels 8 single channels (Excluding F3, F4, F8, FC6, T7, O1)			17.38% (9.37% - 19.64%)		
EEG Beta Rhythm	6 paired channels (Excluding T7-T8) 7 single channels (Excluding AF3, F3, F7, F8, FC5, T8, P8)			15.93% (3.57% - 30.35%)		
Heart Rate	Power Spectral Ratio			12.5%		
Gender	Male or Female			10.71%		
Heart Rate	Mean			10.71%		
EEG Slow-Alpha Rhythm	6 paired channels (Excluding FC5-FC6) 6 single channels (AF3, AF4, T8, P7, O1, O2)			10.71% (2.67% - 8.92%)		
GSR	Mean of the first derivative			3.57%		
Heart Rate	Minimum	High Frequency Power		3.57%	1.79%	

5.3. Affective Clusters vs. Emotion Labels Classification

Figure 10 presents the performance of KNN and SVM classifiers, while classifying the features space with respect to Affective Clusters and Emotion Labels. An Analysis of Variance (ANOVA)[36] showed that the performance of the classifiers, in categorising the emotions into either Affective Clusters or Emotion Labels is not

[36] Classifiers accuracy is considered as the dependent variables, while different classifiers and Affective Clusters vs. Emotion Labels classification technique as the independent parameters.

statistically different ($P_{Clusters}$ = 0.569). However, the performances of KNN and SVM classifiers are significantly different, in terms of their classification accuracy ($P_{Classification}$ < 0.001). On average, KNN (97.01% (±1.3%) mean accuracy across different windowing techniques, Affective Clusters and Emotion Labels) outperformed the SVM algorithm (92.84% (±3.67%) mean accuracy across different windowing, Affective Clusters and Emotion Labels) with around 4%.

As well as the classification accuracy, the F1-Score is another measure that could evaluate the performance of a classifier. Equation 10 presents the F1-Score formula, which is the harmonic mean of *Precision* (the fraction of the identified instances that are relevant), and *Recall* (the fraction of relevant instances that are identified) [**44**].

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Equation 10 – F1-Score Equation [**44**]

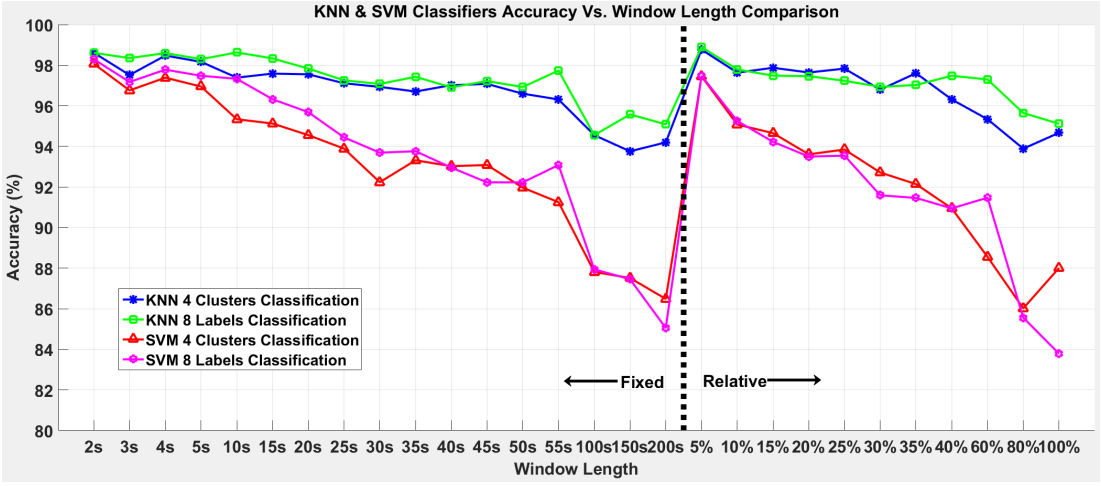

Figure 10 – Accuracy Comparison of KNN and SVM Classifiers, According to Both Affective Clusters and Emotion Labels

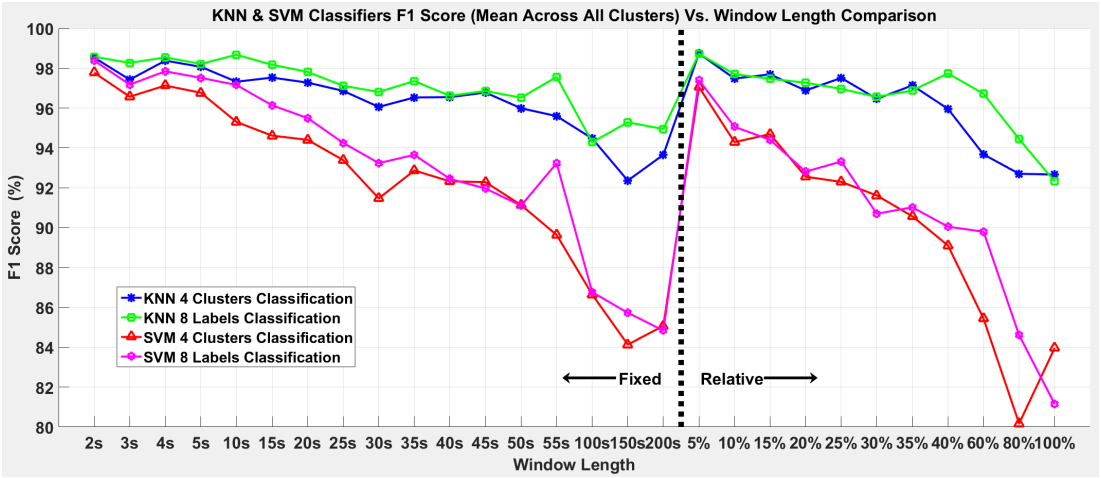

Figure 11 – KNN and SVM Mean F1-Score Across All Clusters

Figure 11 presents the KNN and SVM classification F1-Score, averaged across classes[37]. No game in the experiment has been able to evoke sadness on the part of the participants. Therefore, the classifiers, trained according to Emotion Labels, have not been able to classify any part of the features space into the "Sad" cluster. To be able to compare the performance of the classifiers (with respect to their F1-Score), an Analysis of Variance (ANOVA) has been conducted. The analysis highlighted a significant difference in F1-Score generated by different classifiers ($P_{Clasifier}$ < 0.001) and classification according to either Affective Clusters or Emotion Labels ($P_{Classification\text{-}Class}$ < 0.001). This means that the classifiers' F1-Score, in categorising the emotions into either Affective Clusters or Emotion Labels, is statistically different (unlike the accuracy analysis). Table 5 presents the mean F1-Scores for Affective Clusters, Emotion Labels and classifiers. On average, KNN outperformed the SVM classifier, while the classification according to Affective Clusters performed better, when compared to Emotion Labels.

[37] F1-Score is calculated within each class. Therefore in each windowing technique, 4 and 8 F1-Score for each (respectively) Affective Cluster, and Emotion Label, are calculated.

Table 5 – Mean F1-Scores Across Classifiers, Classes and All Windowing Techniques – (A - B) Presents the (A) 25th and (B) 75th Percentile

Emotion Label	Mean F1-Score (25th – 75th Percentiles)	Affective Cluster	Mean F1-Score (25th – 75th Percentiles)
Relaxed	95.50% (94.31% - 98.07%)	PVLA	94.73% (92.81% - 97.56%)
Content	95.06% (93.78% - 97.74%)		
Happy	94.12% (92.50% - 97.27%)	PVHPA	95.35% (94.34% - 97.34%)
Excited	95.61% (94.04% - 97.94%)		
Angry	94.40% (91.69% - 97.63%)	NVPA	94.82% (93.28% - 97.42%)
Afraid	93.11% (89.86% - 97.25%)		
Sad	Not Available	NVNA	90.76% (87.78% - 96.12%)
Bored	94.87% (93.64% - 97.59%)		

Emotion Labels Classifier	Mean F1-Score (25th – 75th Percentiles)	Affective Clusters Classifier	Mean F1-Score (25th – 75th Percentiles)
KNN	96.94% (96.23% - 98.09%)	KNN	96.29% (95.29% - 97.66%)
SVM	92.40% (89.81% - 96.19%)	SVM	91.54% (88.70% - 95.42%)

6. Conclusion

The human-computer interface has become one of the most important research topics in computer science since the introduction of the first computers (calculators) in the 17th century. As highly complex real-time systems, computers and their interfaces are undergoing an evolution on a hitherto unheard-of scale, in what has become a quest to ensure that they become synergistic, even symbiotic with their human users – transparent, usable, intuitive, sensitive and reactive. As a key part of this evolution, the field of Brain-Computer Interaction is introducing a new dimension to the human interaction process, by establishing direct human brain and computer connection, in an attempt to enhance this symbiosis as much as is technically and ethically possible.

This paper demonstrated the phases of designing, conceptualisation and evaluation of an affective computing system, implemented in virtual reality. The findings of this study suggested that the physiological signals, measured from central and autonomic nervous system, could be employed to classify emotional experiences. By employing a feature selection technique, the dimension of the extracted affective features matrix was reduced from 743 to only 30 most optimum features, for the classification process. By assessing the performance of 28 different windowing techniques, we concluded that there is no difference in employing either relative or fixed windowing techniques. Therefore, as the relative windowing technique cannot be implemented in real-time applications (as the duration of the stimuli cannot be determined until the end of the VR session), the fixed windowing technique could be a more appropriate and credible choice to adopt for real-time applications. However, the analysis suggested that the shorter window length could perform better in the classification process. Furthermore, by training more than a quarter of a million different classifiers (using KNN, SVM, DA and Classification Tree algorithms), the classification accuracies of the affective recognition system, under various settings, have been investigated. According to the cross-validation results, the KNN and SVM outperformed both Classification Tree and DA classifiers. However, the KNN classifier performed slightly better than the SVM in the classifying both Affective Clusters and Emotion Labels (5% higher accuracy and F1-Score).

The final motivation of this research is to implement the designed affective recognition system, into an *Adaptive Virtual Reality* (Adaptive VR) demonstration, capable of adapting its internal environment according to the human users' emotion. Such a development could have significant implications for the development of dynamic human-centred interface techniques, supporting efficient human-system communication styles in a wide range of real-world applications. For example, in command and control for military or counter-insurgency operations, it may be possible to endow multi-input situational awareness display systems with the capability to support end users' decision-making capabilities, generating responses and outcomes based on their instantaneous workload,

stress and emotional characteristics, as remote military incidents evolve and crucial tactical and strategic decisions need to be made. Also in the healthcare domain, where the successful use of Virtual Reality in the delivery of real and imaginary scenes to support patients' cognitive restoration or physical/mental rehabilitation depends significantly on their emotional status and their motivation to engage. These are but two applications domains where the complexity of the human perceptual, motor and cognitive subsystems are only now being given the academic and scientific attention they deserve in the future development of symbiotic, engaging and immersive interfaces.

References

[1] Paul Cairns, Anna Cox, Nadia Berthouze, Samira Dhoparee, and Charlene Jennett, "Quantifying the experience of immersion in games," in *Workshop on the Cognitive Science of Games and Gameplay*, Vancouver, 2006.

[2] Anton Nijholt, Bos, Danny Plass-Oude, and Boris Reuderink, "Turning shortcomings into challenges: Brain–computer interfaces for games," *Entertainment Computing 1*, vol. 1, no. 2, pp. 85–94, April 2009.

[3] Alejandro Rodríguez, Beatriz Rey, Miriam Clemente, Maja Wrzesien, and Mariano Alcañiz, "Expert Systems with Applications Assessing brain activations associated with emotional regulation during virtual reality mood induction procedures," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1699-1709, February 2015.

[4] Avinash Parnandi, Youngpyo Son, and Ricardo Gutierrez-Osuna, "A Control-Theoretic Approach to Adaptive Physiological Games," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Geneva, 2013, pp. 7-12.

[5] Günter Edlinger, Clemens Holzner, Christoph Guger, C. Groenegress, and Mel Slater, "Brain-Computer Interfaces for Goal orientated Control of a Virtual Smart Home Environment," in *Neural Engineering, 2009, NER '09. 4th International IEEE/EMBS*, Antalya, 2009, pp. 463-465.

[6] E. Lalor et al., "Brain Computer Interface based on the Steady-State VEP for Immersive Gaming Control," *EURASIP Journal on Applied Signal Processing*, vol. 49, no. 1, pp. 63-64, 2004.

[7] Gunnar Ahlberg et al., "Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies," *The American Journal of Surgery*, vol. 193, no. 6, pp. 797–804, June 2007.

[8] Michael Zyda, "From visual simulation to virtual reality to games," *Computer*, vol. 38, no. 9, pp. 25 - 32, September 2005.

[9] Neal E. Seymour et al., "Virtual Reality Training Improves Operating Room Performance," *Annals of surgery*, vol. 236, no. 4, pp. 458–464, October 2002.

[10] Nicole E Mahrer and Jeffrey I. Gold, "The Use of Virtual Reality for Pain Control: A Review," *Current pain and headache reports*, vol. 13, no. 2, pp. 100-109, April 2009.

[11] Hunter G. Hoffman, Jason N. Doctor, David R. Patterson, Gretchen J. Carrougher, and Thomas A. Furness, "Virtual reality as an adjunctive pain control during burn wound care in adolescent patients," *Pain*, vol. 18, no. 2, pp. 305–309, March 2000.

[12] A A Rizzo et al., "Virtual environments for the assessment of attention and memory processes: the virtual classroom and office," in *Proceedings of the Fourth ICDVRAT*, Veszprém, September 2002, pp. 3-12.

[13] David Jack et al., "Virtual Reality-Enhanced Stroke Rehabilitation David," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 9, no. 3, pp. 308-318, September 2001.

[14] Thomas D. Parsons and Albert A. Rizzo, "Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 39, no. 3, pp. 250-261, September 2008.

[15] Joann Difede et al., "Virtual Reality Exposure Therapy for the Treatment of Posttraumatic Stress Disorder Following," *The Journal of clinical psychiatry*, vol. 68, no. 11, pp. 1639-1647, November 2007.

[16] Marian Joels, Zhenwei Pu, Olof Wiegert, Melly S. Oitzl, and Harm J. Krugers, "Learning under stress: how does it work?," *Trends in Cognitive Sciences*, vol. 10, no. 4, pp. 152-158, April 2006.

[17] Mohammadhossein Moghimi, Robert J. Stone, Pia Rotshtein, and Neil Cooke, "Influencing Human Affective Responses to Dynamic Virtual Environments," *Teleoperators and Virtual Environments*, vol. 25, no. 2, pp. 81-107, November 2016.

[18] James A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980.

[19] Mohammadhossein Moghimi, Robert J. Stone, and Pia Rotshtine, "Affective Recognition for Multimedia Environments: A Review and Subjective Survey," *IEEE Transactions on Affective Computing*.

[20] Michael Steinbach, Vipin Kumar Pang-Ning Tan, "Introduction to Data Mining," in *Introduction to Data Mining*.: Addison-Wesley, 2005, pp. 65-73.

[21] T. Tamura, H. Miike K. Nakajima, "Monitoring of heart and respiratory rates by photoplethysmography using a digital filtering technique," *Medical Engineering & Physics*, vol. 18, no. 5, pp. 365–372, 1996.

[22] Saeid Sanei and Jonathon Chambers, "Brain Rhythms," in *EEG Signal Processing*. West Sussex: John Wiley & Sons, 2009, pp. 10-13.

[23] Mika Lehtinen, Kimmo Forsman, Jaakko Malmivuo, and Hannu Eskola, "Effects of skull and scalp thickness on EEG," *Medical & Biological Engineering & Computing*, vol. 34, no. 1, pp. 263-264, June 1996.

[24] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, "Chapter 12 - Fast Fourier Transform," in *Numerical recipes in Fortran (The art of scientific computing)*.: Cambridge University Press, 1992, vol. 1, pp. 490-529.

[25] FREDRIC J. HARRIS, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 68, no. 1, pp. 51-83, JANUARY 1978.

[26] Kevin P. Murphy, "Introduction," in *Machine Learning: A Probabilistic Perspective*.: MIT Press, 2012, vol. 1, pp. 2-12.

[27] Saeid Sanei and Jonathon Chambers, "Brain Rhythms," in *EEG Signal Processing*. West Sussex, England: John Wiley & Sons Ltd., 2009, pp. 10-13.

[28] Mohammad Soleymani, Sander Koelstra, Ioannis Patras, and Thierry Pun, "Continuous emotion detection in response to music videos," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, Santa Barbara, 2011, pp. 803 - 808.

[29] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun, "Emotion assessment from physiological signals for adaptation of game difficulty," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* , vol. 41, no. 6, pp. 1052-1063, November 2011.

[30] Sander Koelstra et al., "DEAP: a Database for Emotion Analysis Using Physiological Signals," *Affective Computing, IEE Transactions*, vol. 3, no. 1, pp. 18-31, January 2012.

[31] Hanchuan Peng, Fuhui Long, and Chris Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *1226 IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 27, no. 8, pp. 1226-1238, August 2005.

[32] Kevin P. Murphy, "Entropy," in *Machine Learning A Probabilistic Perspective*.: MIT Press, 2012, p. 57.

[33] Kevin P. Murphy, "Support Vector Machines," in *Machine Learning A Probabilistic Perspective*.: MIT Press, 2012, vol. 1, pp. 498-507.

[34] Kevin P. Murphy, "Gaussian Discriminant Analysis," in *Machine Learning A Probabilistic Perspective*.: MIT Press, 2012, vol. 1, pp. 103-112.

[35] Kevin P. Murphy, "Classification and Regression Trees (CART)," in *Machine Learning A Probabilistic Perspective*.: MIT Press, 2012, vol. 1, pp. 546-554.

[36] Kevin P. Murphy, "A Simple non-Parametric Classifier: K-Nearest Neighbors," in *Machine Learning A Probabilistic Perspective*.: MIT Press, 2012, vol. 1, pp. 16-18.

[37] Kevin P. Murphy, "Kernels," in *Machine Learning A Probabilistic Perspective*.: MIT Press, 2012, pp. 481-488.

[38] James Bergstra and Yoshua Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281-305, February 2012.

[39] Kevin P. Murphy, "Estimation the Risk Using Cross Validation," in *Machine Learning: A Probabilistic Perspective*.: MIT Press, 2012, p. 209.

[40] James J. Gross et al., "Emotion and aging: experience, expression, and control," *Psychology and aging*, vol. 12, no. 4, pp. 590-9, 1997.

[41] Christine Combain, Arnaud D'Argembeau, Martial Van der Linden, and Laurence Aldenhoff, "The effect of ageing on the recollection of emotional and neutral pictures," *Memory (Hove, England)*, vol. 12, no. 6, pp. 673-684, November 2004.

[42] Emery Schubert, "Locus of emotion: The effect of task order and age on emotion perceived and emotion felt in response to music," *Journal of music therapy*, vol. 44, no. 4, pp. 344-368, December 2007.

[43] Daniel K. Mroczek, "Age and Emotion in Adulthood," *Current Directions in Psychological Science*, vol. 10, no. 3, pp. 87-90, June 2001.

[44] Kevin P. Murphy, "F-Score," in *Machine*

Learning: A Probabilistic Perspective.: MIT Press, 2012, vol. 1, pp. 184-186.

[45] M. Murugappan et al., "Time-Frequency Analysis of EEG Signals for Human Emotion Detection," in *Springer-Verlag*, Berlin, 2008, pp. 262-265.

[46] M. Rizon, M. Murugappan, R. Nagarajan, and S. Yaacob, "Asymmetric Ratio and FCM based Salient Channel Selection for Human Emotion Detection Using EEG," *WSEAS Transactions on Signal Processing*, vol. 4, no. 10, pp. 596-603, October 2008.

[47] Jeremy N. Bailenson et al., "Real-time classification of evoked emotions using facial feature tracking and physiological responses," vol. 66, no. 5, pp. 303-317, May 2008.

[48] Davor Kukolja, Siniša Popović, Marko Horvat, Bernard Kovač, and Krešimir Ćosić, "Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications," *International Journal of Human Computer Studies*, vol. 72, no. 10-11, pp. 717-727, November 2014.

[49] Jonghwa Kim and Elisabeth Andre, "Emotion recognition based on physiological changes in music listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067-2083, December 2008.

[50] Mimma Nardelli, Gaetano Valenza, Alberto Greco, Antonio Lanata, and Enzo Pasquale Scilingo, "Recognizing emotions induced by affective sounds through heart rate variability," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 385-394, October 2015.

[51] Kevin P. Murphy, "Logistic Regression," in *Machine Learning A Probabilistic Perspective*.: MIT Press, 2012, vol. 1, pp. 21-22.

[52] Mohammadhossein Moghimi, Robert J. Stone, and Pia Rotshtine, "Part A:," *IEEE Transactions on Affective Computing*.

Mohammadhossein Moghimi is a teaching assistant and final-year PhD student in the Human Interface Technologies (HIT) research group at the School of Electronic, Electrical and System Engineering. His main research area is in implementation of Brain-Computer Interaction (BCI) in Human-Computer Interaction (HCI) domain, Virtual Reality (VR) and 3D environments. His interests today focus on the development of Emotion Recognition system, and their implementations within virtual highly interactive environments.

Robert (Bob) Stone is Professor of Interactive Multimedia Systems at the University of Birmingham, Director of the University's Human Interface Technologies Team, and an Academician of the Russian International Higher Education Academy of Science. A Chartered Psychologist and a Fellow of the Institute of Ergonomics and Human Factors, Bob has over 36 years of experience in human factors research and consultancy in both industry and academia. His interests today focus on the development of human-centred solutions for virtual reality, mixed reality and advanced human interface applications in defence, digital heritage and healthcare. His healthcare research focuses on future simulation and interactive technologies for post-trauma/post-operative physical and psychological rehabilitation, and future "mixed reality" pre-deployment training solutions for defence medics. Bob collaborates closely with the Royal Centre for Defence Medicine and the Queen Elizabeth Hospital in Birmingham, where he is also a member of the Hospital's Human Factors Faculty. The recipient of numerous national and international awards, 2017 will be Bob's 30th year of involvement in the VR community, courtesy of a career-changing experience in 1987 at NASA Ames in California – one that has driven his human factors research goals ever since.

Dr. Pia Rotshtein is a cognitive neuroscientist working in the School of Psychology, University of Birmingham. Part of her areas of interest includes understanding the mechanisms underlying the superior social skills of humans, including emotional and motivational processing.