Evolving Systems manuscript No.
(will be inserted by the editor)

Affect Detection from Non-stationary Physiological Data using Ensemble Classifiers

Omar AlZoubi · Davide Fossati · Sidney K. D'Mello · Rafael A. Calvo

Received: date / Accepted: date

Abstract Affect detection from physiological signals has received considerable attention. One challenge is that physiological measures exhibit considerable variations over time, making classification of future data difficult. The present study addresses this issue by providing insights on how diagnostic physiological features of affect change over time. Affective physiological data (Electrocardiogram, Electromyogram, Skin Conductivity, and Respiration) was collected from four participants over five sessions each. Classification performance of a number of training strategies, under different conditions of features selection and engineering, were compared using an adaptive classifier ensemble algorithm. Analysis of the performance of individual physiological channels for affect detection is also provided. The key result is that using pooled features set for affect detection is more accurate than using day-specific features. A decision fusion strategy which combines decisions from classifiers trained on individual channels data outperformed a features fusion strategy. Results also show that the performance of the ensemble is affected by the choice of the base classifier and the *alpha* factor used to update the member classifiers of the ensemble. Finally, the corrugator and zygomatic facial EMGs were found to be more reliable measures for detecting the valence component of affect compared to other channels.

Keywords Affect · emotion · classifier ensembles · physiological · non-Stationary

1 Introduction

There is increased motivation in using physiological signals in affect detection systems that detect either discrete emotional categories or affective dimensions of valence and arousal (AlZoubi et al, 2011; Kim and André, 2008; Picard et al, 2001). Physiological responses such as facial muscle activity, skin conductivity, heart activity, and respiration have all been considered as potential markers for recognizing affective states (Whang and Lim, 2008). Despite high classification rates achieved under laboratory conditions (Kim et al, 2004; Lichtenstein et al, 2008), the changing nature of physiological signals introduces significant challenges when one moves from the lab and into the real world (AlZoubi et al, 2012; Plarre et al, 2011). In particular, physiological data is expected to exhibit daily variations or non-stationarities (AlZoubi et al, 2011; Picard et al, 2001), which introduce difficulties for building effective classification models on future data (i.e., signals generated by the same individual across time). The ability to integrate automatic affect detection capabilities in computer systems depends largely on the underlying models of affect, and how these models can adapt to the changing nature of physiological data.

O. AlZoubi
Computer Science
Carnegie Mellon University
E-mail: oalzoubi@cmu.edu

D. Fossati
Computer Science
Carnegie Mellon University
E-mail: dfossati@cmu.edu

S.K. D'Mello
Department of Computer Science
The University of Notre Dame
Notre Dame, IN, USA
E-mail: sdmello@nd.edu

R.A. Calvo
School of Electrical and Information Engineering
The University of Sydney
Sydney, Australia
E-mail: Rafael.Calvo@sydney.edu.au

Previous research has shown that affective physiological data exhibited daily variations (AlZoubi et al, 2011; Picard et al, 2001). It was found that physiological data for a given emotion on a particular day (day-data) yielded a higher clustering cohesion or tightness compared to data for the same emotion across multiple days. This phenomenon can be attributed to a number of factors such as: 1) mood changes across days; 2) electrodes drift; 3) changes in electrode impedance; and 4) modulations by other mental states such as attention and motivation (Picard et al, 2001). Non-stationarity indicates that signals change their statistical characteristics (e.g., means, standard deviation) as a function of time, which then propagates to features values extracted from the signals over time.

Non-stationarities of physiological signals represent a major problem for building reliable classification models that span multiple days. Most classification methods assume that training data is obtained from a stationary distribution (Last, 2002). However, this assumption of stationarity is routinely violated in real-world contexts. According to Kuncheva (2004a), every real-world classification system should have a mechanism to adapt to time-varying changes. In order to address this issue, this study utilizes an adaptive ensemble classification approach - discussed in more detail in Section 3.

Understanding the nature of non-stationarities in physiological signals is essential for developing reliable affect detection systems that can be deployed in real-world affective computing applications. There is a critical need for basic research on how physiological signals vary over time. This research contributes to this goal by addressing two fundamental issues. First, we study temporal changes to diagnostic physiological features collected from four participants over five recording sessions. The non-stationarities in physiological data might indicate that diagnostic features of affect may vary from one day/session to another. We test this issue and evaluate an adaptive ensemble classification approach that can potentially handle non-stationarities in affective physiological data. The performance of the ensemble was tested under different conditions of features engineering and selection. Second, we test the performance and reliability of individual physiological channels for affect detection over the span of multiple day recordings. Our results show that a decision fusion strategy which combines decisions from classifiers trained on individual channels data outperformed a features fusion strategy. The results also show that the choice of the base classier, and the *alpha* factor used to update the member classifiers of the ensemble have an effect on the performance of the ensemble. The corrugator and zygomatic facial EMGs were found to be more reliable measures for detecting valence than arousal compared to other channels.

The remainder of this paper is organized as follows. Section 2 gives an overview on affect detection using physiological data. Section 3 describes the procedure of collecting affective physiological data and the computational methods employed for features extraction and classification. Section 4 presents our results, while Section. 5 provides discussions and avenues for future work.

2 Background and related work

2.1 Physiological-based affect detection

According to Lazarus (1991) people adapt to their environment and to emotional stimuli via autonomic nervous system (ANS) responses. Therefore, patterns of ANS activity should be correlated with particular emotional states. For example, research has revealed consistent changes in facial electromyogram (EMG), particularly the corrugator muscle, in response to pleasant or unpleasant stimuli (Lee et al, 2009). Similarly, electrocardiogram (ECG) features, such as heart rate (HR) and HR variability, can both be used as indicators of valence and arousal (van den Broek et al, 2009). Skin Conductivity (SC) has been traditionally considered to be an index of arousal (Levenson, 1992). Respiratory patterns may reflect and distinguish between emotional states such as calmness versus excitement (Allanson and Fairclough, 2004).

Research on emotion has typically relied on a set of discrete emotional prototypes or basic emotions (e.g., happiness, sadness) (Ekman, 1992). As opposed to the existence of discrete basic emotions, Russell (1980) suggested that affective experience is best described in the two-dimensional space of valence and arousal. The arousal dimension ranges from highly deactivated to highly activated (or sleepy to active), and the valence dimension from highly unpleasant to highly pleasant. For example, happiness is considered to have a positive valence and high arousal. On the other hand, sadness has a negative valence and low arousal. According to Ekman (1994), emotions are short-lived, ranging from seconds to minutes at most.

Recent research has utilized physiological signals for affect detection of both arousal and valence. Kim and André (2008) used physiological signals to detect levels of valence and arousal during music listening. They recorded ECG, facial EMG, SC, and respiration (RSP) from three participants. Using an LDA classifier, they achieved an 89% classification accuracy for high/low valence, and 77% for high/low arousal. Similarly, Lichtenstein et al (2008) recorded physiological data (ECG, SC, EMG, RSP, and skin temperature) while 41 participants watched emotionally charged films. They were able to detect high/low arousal with 82% accuracy and 72% for high/low valence using a support vector machine (SVM) classifier. Likewise, Picard et al (2001) recorded physiological data over a period of 20 days from one participant (an actor who was asked to self-elicit a set of eight emotions). They faced the problem of degraded classification performance when data from multiple days were combined. They attempted to address the problem of day variation by including day information as additional classification features; however, this did not yield a significant improvement in accuracy.

Picard et al (2001) attempted to address the problem of daily variations in physiological data by including day information as classification features. However, they reported insignificant increase in classification accuracy. Similarly Vyzas and Picard (1998) found that the underlying mood appears to change the features values for all emotions. However, it had less of an effect on the relative inter-relations among emotions. Therefore, they emphasized the need for a real-time emotion detection system that can adapt to a person's underlying mood.

It is anticipated that hardware and environmental factors that affect physiological data can be mitigated with advances in sensors and physiological recording devices technology. However, users' factors cannot be easily alleviated. Applying an adaptive learning strategy could be a possible solution to address changes in physiological data. In the next section, we discuss in more detail the justification for using adaptive and ensemble classification for affect detection from physiological data.

2.2 Approaches to adaptive classification in changing environments

Many affective computing studies have relied on the use of traditional batch static classification techniques (Kim and André, 2008; Kim et al, 2004; Lichtenstein et al, 2008; Picard et al, 2001). These classification techniques learn a single model by examining a large collection of instances at one time. These techniques are based on the assumption that training and future testing data are obtained from a stationary distribution, therefore there is no updating mechanism to their underlining model. In real-world scenarios, data is collected over time, which may range from seconds to days to years. Therefore, changes in the data characteristics are inevitable (Nishida et al, 2005; Sayed-Mouchaweh and Lughofer, 2012). According to Cieslak and Chawla (2009), the existence of a one-true-model or a well-calibrated classifier that is able to map every unseen example correctly assumes that data comes from a stationary distribution. However, if the data distribution changes substantially and unpredictably, the one-true-model may become irrelevant when applied to future instances. In other words, a pattern discovered by a model from past data may not be valid on the newly acquired data (Last, 2002). It is widely acknowledged that humans learn in changing environments in a sequential manner by leveraging prior knowledge in new situations. Therefore, the ability to make human-like quick responses should be developed in machines to handle real-world problems of this nature. Adaptive classifiers promise to give machines this human-like capability (Angelov et al, 2010; Nishida et al, 2005). In contrast to traditional classification systems which require a large sample of training data and start learning from scratch, adaptive classifiers learn sequentially, as data comes in, through an update mechanize to their underlying model.

Ensemble learning is a promising approach for handling non-stationary data (Kuncheva, 2004a; Yue et al, 2007). The ensemble consists of a group of classifiers that learn from the incoming data, instead of a single classifier. The idea is to train each ensemble member on a different data segment with an unknown rate of shift in distribution (Muhlbaier and Polikar, 2007). The final output of the ensemble will depend on some defined rules (e.g., majority voting). In the mathematical classical bias/variance trade-off, classifier ensembles offer an extra degree of freedom, which allow to obtain solutions that would be difficult with a single classifier (Oza and Russell, 2001; Oza and Tumer, 2008). When time to make decisions is not the most important factor, but high accuracy is required, an ensemble would be a likely solution. (Kuncheva, 2004a). Efficient learning in changing environments requires a learning algorithm that can adapt quickly to a change in classification environment by adjusting its knowledge-base, and can utilize previously learned knowledge in situations where old contexts reappear (Kuncheva, 2004a; Widmer and Kubat, 1996).

It has been established that much is to be gained from combining classifiers if the classifiers are as independent as possible and are trained in different regions of features space as they will be able to provide complementary information (Duda et al, 2001; Polikar et al, 2001). Thus, the individual weakness or instability of each classifier can be effectively averaged out by the combination process, which may significantly improve generalization of the classification system (Polikar, 2006). In general, there are multiple design reasons to consider ensemble-based approaches (Jain et al, 2000; Kuncheva, 2004b; Polikar, 2006; Webb, 2002), including but not limited to the following:

- Different classifiers can be developed in different contexts/representation of the same problem. An example is person identification by their voice, face as well as handwriting.
- Different classifiers trained on the same data may show strong local and global differences when deciding decision boundaries between classes.
- Different classifiers can be trained to solve a problem that is too difficult, and the decision boundary that separates classes is too complex. Thereby, using a divide and conquer strategy to break the problem into smaller sub-problems.

In emotion detection research ensemble classifiers have been found to offer some enhancement in accuracy rates compared to single classifiers. For example, Kuncheva et al (2011) compared the performance of eight single classifiers and six ensemble methods for detecting negative and positive affective states from physiological data (EEG, EDA and pulse sensor). They found that ensemble methods outperformed single classifiers in all comparisons. This shows that ensemble methods have the potential for building more accurate and reliable automatic affect detection systems. Similarly, AlZoubi et al (2009) compared the performance of adaptive and static classification approaches for classifying 10 affective states from electroencephalogram (EEG) signals. They used an adaptive algorithm that updates its knowledge base based on most recent examples, and deleting the oldest examples. Results showed that adaptive classifiers outperformed the static versions of the classifiers.

Adaptive classification techniques have been used to handle non-stationarities in two close domains of study: speech recognition and Brain Computer Interfaces (BCI). For example, Maier-Hein et al (2005) implemented an adaptive approach to detect non-audible speech using seven EMG electrodes. They found that a key problem in surface EMG-based speech recognition result from electrodes repositioning between recording sessions, temperature changes in the environment, and skin characteristics of the speaker. In order to reduce the impact of these factors, they investigated a variety of signal normalization and model adaptation methods. An average word accuracy of 97.3% was achieved using seven EMG channels with the same electrode positions. The performance dropped to 76.2% after repositioning the electrodes, when no normalization or adaptation was performed. However, they were able to restore the recognition rate to 87.1% using adaptive classification methods.

Adaptive classification has also been employed in BCI research. BCI aims at giving the ability to control devices through mere thoughts by analyzing brain signals, such as Electroencephalogram (EEG). For example, Lowne et al (2010) compared the performance of a dynamic classification approach to a static classifier and a multilayer perceptron (MLP) classifier on an online BCI experiment. They used EEG data from eight participants during a wrist extension exercise; 20% of the data were labeled with true labels (movement, nonmovement). The three classifiers were then tested on the EEG data in a sequential manner (time ordered) to detect one of the two classes. The performance of the dynamic classifier was significantly higher than that of the static classifier and MLP. One important feature of the dynamic classifier is the active label requesting, which employs a probabilistic model and sets a threshold about the confidence of the predicted class label, if the confidence is low the classifier might issue a request for the true class label.

In summary, these studies show that classifier adaptation might be more suitable to handle non-stationary data. Our approach capitalizes on the advantages of both adaptive and ensemble classification techniques for classifying our affective physiological dataset.

2.3 Fusion techniques in multimodal affect detection

Affective information can be collected from multiple sources such as voice, facial expressions, and physiological signals. Therefore there is a need for techniques that combine and synthesize information from these multimodal sources. This process is referred to as information fusion. There are a number of fusion techniques associated with multimodal emotion detection, such as features level fusion and decision level fusion (Zeng et al, 2009). Features level fusion aims at integrating extracted features from each modality into one joint features vector. The issue with this approach is that features from different signals might have different temporal resolutions, which may require synchronization of the extracted features. On the other hand, decision level fusion aims at integrating asynchronous but temporally correlated modalities. Each modality is classified independently and the final decision is obtained by fusing the decisions of all the modalities based on some criteria such as averaging or voting. Designing an optimal strategy for decision level fusion is still an open research problem (Kim and André, 2006).

Chanel et al (2006) found that fusion provides more robust results when combining EEG and peripheral signals related to ANS responses such as ECG, EMG, SC and RSP. According to the authors, some participants had better scores with peripheral signals than with EEG and vice-versa. Similarly, Kim and André (2006) found that features level fusion provided the best results using physiological signals and voice modalities, noting that features level fusion is more appropriate when combining modalities with analogous characteristics. In this study we evaluate both fusion strategies.

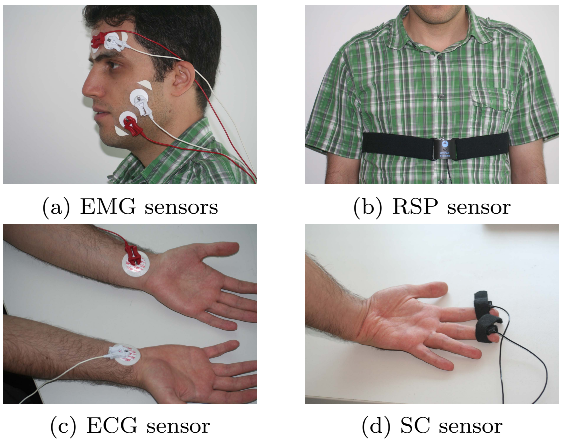
3 Measures, data and methods

3.1 Participants and measures

Participants were six students enrolled in an Australian University (five males and one female), between 24 and 39 years of age. Participants were paid for their participation in the study. The study was approved by the University of Sydney's Human Ethics Research Committee (HERC), and consents were obtained from participants prior to data collection. Physiological data included: Electrocardiogram (ECG), Skin Conductivity (SC), Electromyogram (EMG), and Respiration (RSP). The physiological signals were acquired using BIOPAC MP150 system and AcqKnowledge software with a sampling rate of 1000 Hz for all channels. The ECG signal was collected with two electrodes placed on both wrists. EMG was recorded from the corrugator (eyebrow) and zygomatic (cheek) facial muscles. The SC was recorded from the index and middle fingers of the non-dominant hand, and a respiration belt fixed around the participant chest was used to measure respiration activity. These sensors were non-invasive and caused minimal distress to participants. Figure 1 shows a participant with the sensors attached.

Physiological signals were filtered to remove environmental noise including baseline drifts, artefacts resulting from movements, and mains interference. The ECG signal was high pass filtered at 0.05 Hz and low pass at 35 Hz, with a notch filter applied through the recording device. The EMG signal was high pass filtered at 10 Hz to remove low frequency artefacts such eye movements, eye blinks and motion potentials, and low pass filtered at 500 Hz. The SC and RSP signals were high pass filtered at 0.05 Hz in order to remove slow drifts, and low pass filtered at 1 Hz in order to remove high frequency noise. Figure 2 shows a sample of a recorded signal.

The affect-inducing stimulus consisted of set of 400 images selected from the International Affective Picture System (IAPS) collection (Lang et al, 1995). The IAPS collection is designed to provide a set of normative emotional stimuli for the study of emotions and attention (Lang et al, 2005). The images were selected on the basis of their normative valence and arousal scores. The mean valence normed scores range from 1.40 to 8.34 (M = 5.03, SD = 1.77), and mean arousal normed scores range from 1.72 to 7.35 (M = 4.82, SD = 1.55), on a scale from 1 to 9. Figure 3 shows the distribution of IAPS images on the valence/arousal plane.



(a) EMG sensors (b) RSP sensor

(c) ECG sensor (d) SC sensor

Fig. 1 Sensors' placement.

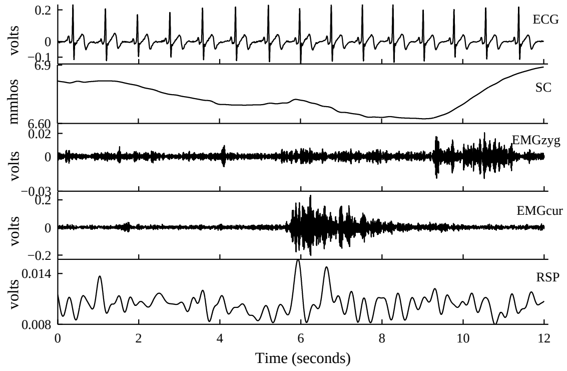


Fig. 2 Sample recorded signals.

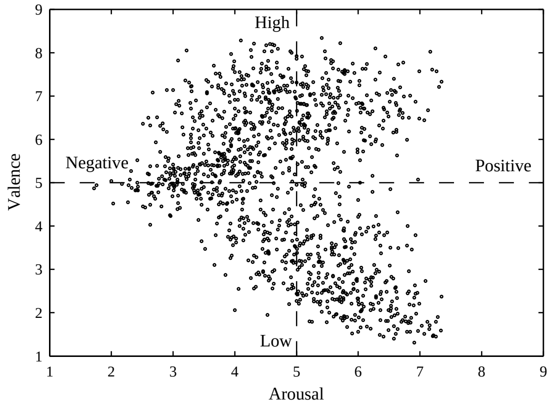


Fig. 3 The distribution of IAPS images into the valence-arousal plane.

Table 1 The four quadrants, and mean values for valence and arousal

Quadrant	Mean valence and arousal range
Positive-Valence/Low-Arousal	(valence-mean > 6.03 && arousal-mean < 5.47)
Positive-Valence/High-Arousal	(valence-mean > 6.03 && arousal-mean > 5.47)
Negative-Valence/High-Arousal	(valence-mean < 3.71 && arousal-mean > 5.47)
Negative-Valence/Low-Arousal	(valence-mean < 3.71 && arousal-mean < 5.47)

Images were selected from the four quadrants of the valence-arousal plane, described in Table 1. The idea was to select images from the extremes of both valence and arousal in order to maximize the differences of participants' physiological responses. The set of 400 images was then divided into 5 sets of 80 images each (20 images from each category). However, we consider the valence and arousal dimensions separately in the classification experiments described in Section 4.

Only four participants were able to complete the five recording sessions, which was due to the distressing nature of some IAPS images. Therefore, only data from these four participants were used in the current study. We note that even though the participant sample size is small, each participant was recorded over 5 sessions. This is consistent with the present goal of tracking variations within an individual rather than across individuals.

3.2 Experimental procedure

Physiological signals were continuously recorded while participants viewed a set of emotionally charged IAPS images. Each recording session lasted approximately 60 minutes. Emotional trials consisted of presenting each image for 12 seconds, followed by a 2 × 2 affective grid (Russell et al, 1989) that asked participants to rate their levels of valence (positive, negative), and arousal (low, high). The affective grid had four buttons projected on the four quadrants of the valence/arousal plane, each button representing one of the categories described in Table 1. Next, a blank screen was presented for 8 seconds to allow physiological activity to return to baseline neutral levels before a new image was presented. Five images were presented consecutively from each category in order to maintain a stable emotional state for that category. This protocol was designed to suit the intended goals of our study and is based on previous research (Bradley and Lang, 2007; Chanel et al, 2006). Each participant participated in five recording sessions, each separated by one week. A different set of images were presented for each session in order to prevent habituation effects. However, each set contained 20 images from each of the four categories described in Table 1.

3.3 Day-datasets

Day-datasets were constructed separately for the two affective measures of valence and arousal. Datasets were constructed for both IAPS-mapped categories and self-reports of participants. In total there were 80 datasets, (4 participants × 5 recording sessions × 2 affective measures (valence and arousal) × 2 ratings (IAPS and self-reports)), with 80 instances in each dataset. IAPS ratings datasets had a balanced distribution of classes with 40 instances for each class of positive/negative valence, and low/high arousal. On the other hand, self-reports had unbalanced distribution of classes. A down-sampling procedure, WEKA's SpreadSubsample which produces a random sub-sample of a dataset, was applied to obtain a balanced distribution of classes. Therefore, the baseline classification accuracy is (50%) for both types of data sets. It should be noted that we did not opt to use oversampling procedures, since they can introduce artificial patterns that may affect the reliability or interpretation of results.

3.4 Features extraction

The MATLAB Augsburg Biosignal Toolbox (AuBT) (Wagner et al, 2005) was used to extract features from the raw physiological data. A total of 214 statistical features (e.g. mean, median, standard deviation, maxima and minima) were extracted from the five physiological channels using window size of 12 seconds (the length of the emotional trial). The same statistical features were obtained for different transformations of the signals, including RSP rate, amplitude of the RSP signal, heart rate variability (HRV) and amplitude of the different segments of the QRS complex of the ECG signal. These same features were also computed from the first and second derivatives of the signals and their transformations. It is known that the temporal resolution for these autonomic measures vary in response to emotional stimuli. In general, SC responses (SCR) can be observed 1–3 seconds after stimulus presentations. EMG responses are substantially faster, however, the frequency of the muscle activity can be summed up over a period of time to indicate a change in behavior (Andreassi, 2007). ECG and respiration responses are considered slower, but we were constrained to use a window size of 12 seconds because this was the length of a single trial. However, estimating short term cardiac and respiratory patterns is common in psychophysiology research area (Kreibig, 2010). Overall, eighty-four features were extracted from ECG, 21 from SC, 21 from each of the EMG channels, and 67 from the RSP channel. A complete description of these features can be found in (Wagner, 2009).

3.5 Classification methods

Algorithm 1 describes the Winnow updatable ensemble algorithm used in this study. Winnow is an ensemble based algorithm that is similar to a weighted majority voting algorithm because it combines decisions from ensemble members based on their weights (Kuncheva, 2004a). However, it utilizes a different updating approach for member classifiers. This includes promoting ensemble members that make correct predictions and demoting those that make incorrect predictions. This updating strategy ensures that correct decisions made by the ensemble are amplified, and incorrect ones are minimized. Updating of the weights is done automatically based on incoming data, which makes this approach suitable for online applications. In order to construct the ensemble, we used a fixed ensemble size, which is equal to the number of day-datasets per participant. In our approach, each day-dataset was used to train a separate classifier in batch mode, which was then added to the ensemble.

Algorithm 1 The Winnow ensemble algorithm

```
1 Initialization: Construct a classifier ensemble
  $D = (D_1, ..., D_n)$, each classifier is trained in batch
  mode on a given dataset, Initialize all classifiers
  weights; $W_i = 1$. $i = 1 : n$.
2 Classification: For a new example $X$, calculate the
  support for each class as the sum of the weights of all
  member classifiers $D_i$ that suggest class label $C_k$ for
  $X$. Set $X$ to the class with largest support.
  $K = 1$:number of classes.
3 Updating: if $X$ is classified correctly by classifier $D_i$
  then its weight is increased (promotion):
  $W_i = alpha * W_i$, where $alpha > 1$. If classifier $D_i$
  incorrectly classifies $X$, then its weight is decreased
  (demotion): $W_i = W_i/alpha$
```

There are other adaptive ensemble classification algorithms (e.g., Dynamic Weighted Majority, and Hedge β algorithms) described in literature (Kolter and Maloof, 2003; Kuncheva, 2004a). Some of these algorithms may have different strategies for building and updating the ensemble compared to the one described above. This simple adaptive ensemble algorithm was adopted in order to demonstrate the efficacy of adaptive ensembles for handling non-stationarities of physiological data in comparison to batch static classification. This decision was also motivated by the nature of the classification problem at hand, with data generated from multiple session recordings.

We used a fixed ensemble size of four base classifiers; each classifier trained on data from a single session recording. The reason behind this decision is twofold. Firstly, the nature of the data; we found in our previous work that data that comes from each session showed high clustering cohesion compared to data from other sessions (AlZoubi et al, 2011). Each ensemble member can then be viewed as a specialist or an expert classifier. Secondly, the small data sample size; we only have data from five recording sessions. However, we believe that a dynamic ensemble size is mandatory when more data becomes available. In this case the lowest performing classifier/s can be removed from the ensemble and newer members are added. This is a common approach for online ensembles with large data throughput (Bifet et al, 2009).

The WEKA machine learning software (Witten and Frank, 2005) and PRTools 4.0 (Heijden et al, 2004) were used for preprocessing, features selection and classification. PRTools offers a variety of data preprocessing and classification methods that allow for the design of custom-specific classification programs in MATLAB. Chi-square features selection was used to reduce the dimensionality of the features space in order to avoid various problems associated with large features spaces. A preliminary analysis showed that using top-five ranked features were sufficient to produce consistent classification results without sacrificing performance. Therefore, using Chi-square features selection, the top five features were selected from each dataset and used in all subsequent analysis.

4 Results

We first tested the effectiveness and reliability of the IAPS images for inducing both valence and arousal (using inter-rater Cohen's kappa as an evaluation metric). We then tested how diagnostic features of affect change over time by applying feature ranking (chi-square) to separate day-datasets. A number of training strategies were designed to help mitigate the issue of features changes over time. A number of experiments were carried out to test the performance of the Winnow ensemble algorithm under different conditions of features engineering and selection. We tested four training strategies, which are: 1) day cross-validation; 2) Winnow with pooled features; 3) Winnow with day-specific features; and 4) Winnow with decision fusion from individual channels data . The details of these training strategies are explained below:

- Static classification (SCL): A single baseline classification model was constructed from pooled data of four days, and testing was done on data from the remaining day. This process was repeated five times in order to test on all available data. This training strategy represents a static classification approach without an update mechanism.
- Winnow with pooled features (WPF): The Winnow ensemble algorithm was run with an ensemble of four base classifiers each trained on a pooled features set. Pooled features are features selected from four days data combined. Testing was done on the remaining day-data. The procedure was repeated five times to test on all available data.
- Winnow with day-specific features (WSP): The Winnow ensemble algorithm was run with an ensemble of four base classifiers each trained on a separate day-specific features dataset. Testing was done on the remaining day-data. The procedure was repeated five times to test on all available data.
- Winnow with decision fusion from individual channels data (WDF): This method used base classifiers that were trained on individual channels data using pooled features. Using datasets from four days, a new member classifier was constructed from the 5 physiological measures resulting in an ensemble with 20 classifiers. Testing was done on the remaining day-data. The procedure was repeated five times to test on all available data. In order to classify a new example, the decisions of these base classifiers are combined using the Winnow decision fusion strategy.

The fundamental assumption behind utilizing classification techniques is that pre-trained classification models can be used to predict future unseen input. Thus, a day cross-validation procedure was adopted so that training data included data from four days, and the fifth day-data was used for testing. This procedure was repeated five times to test on all day-datasets. The objective of this analysis is to assess the accuracy of classifiers that are trained on different day-data to predict exemplars from other days. We also test the effect of individual physiological channels on affect detection accuracy. In addition, we evaluated the effect of the two factors that can affect the performance of the Winnow ensemble. These are: *a*) The baseline classifier; and *b*) The *alpha* factor used to update the weights of the ensemble.

4.1 The effectiveness of IAPS at inducing affect

We used inter-rater Cohen's Kappa to test the effectiveness of the IAPS stimuli in inducing both valence and arousal. We test the level of agreement between participants' self-reports and IAPS normative ratings. Participants' self-reported valence showed higher agreement (kappa = 0.89) with IAPS normative ratings, whereas arousal self-reported arousal did not show that level of agreement (kappa = 0.41). It is evident that the IAPS stimuli were quite successful in eliciting valence, but was much less effective in influencing arousal. However, both ratings dimensions will be used to assess affect detection accuracy.

4.2 Day-specific features

As an example of how diagnostic features change across days, Table 2 presents the results of chi-square features selection applied to participant S_1 (applied to each day-data separately). It can be seen from the list of features that the diagnostic features are different for each day. The chi-square value represents the degree of relevance of a feature to class category. Table 2 presents the features selected from one participant using IAPS ratings only, however data from other participants showed similar patterns. An interesting observation is that there are some features which reoccur on different days (e.g., ZYG-EMG-1Diff-maxRatio, ZYG-EMG-1Diff-minRatio). This is promising as it allows for easier calibration of affect detection classification models. However this leaves us the question of whether classification models that are built from these day-specific features are more accurate than those built using pooled features.

4.3 Winnow results with day-specific and pooled features

The SCL, WPF, and WSF training strategies were used to classify data from the four categories; Valence-IAPS, Arousal-IAPS, Valence-Self, and Arousal-Self. The results in Table 3 were obtained using features fusion from all five physiological channels; top five features were selected from all channels. A SVM classifier with a linear kernel was used as a base classifier for the training. We also used using the same training strategies to classify individual channels data. Table 4 shows the

Table 2 The top five selected features using chi-square features selection performed on day-data separately for participant S_1 with valence(IAPS) as class label

Chi Square/Feature Name				
Day 1 Features	*Day 2 Features*	*Day 3 Features*	*Day 4 Features*	*Day 5 Features*
43 SC-2Diff-minRatio	23 ZYG-EMG-1Diff-minRatio	10 SC-1Diff-minRatio	16 ZYG-EMG-max	28 ZYG-EMG-2Diff-minRatio
43 SC-2Diff-maxRatio	23 ZYG-EMG-1DiffmaxRatio	10 RSP-Ampl-1Diff-max	10 ECG-QS-min	26 SC-2Diff-maxRatio
23 ZYG-EMG-1Diff-maxRatio	14 RSP-2Diff-range	10 SC-1Diff-maxRatio	9 ECG-QS-range 2	4 SC-2Diff-minRatio
23 ZYG-EMG-1Diff-minRatio	13 RSP-2Diff-min	76 RSP-Ampl2Diff-maxRatio	8 ECG-HrvDistr-mean	23 ZYG-EMG-2Diff-maxRatio
23 RSP-Pulse-max	12 ZYG-EMG-2Diff-mean	5 ECG-HrvDistr-mean	6 RSP-Pulse1Diff-maxRatio	12 RSP-Ampl-mean

*ZYG: Zygomatic facial muscle, Amp: Amplitude, min: Minimum, max, Maximum, HRV: Heart rate variability, 1 Diff: First Difference, 2 Diff: Second Difference

classification accuracies for the individual physiological channels for the four emotional categories using pooled features only. Results for day cross-validation and day-specific features are not shown here but will be outlined in the Analysis of Variance (ANOVA) analysis described next.

In order to examine the effect of training strategy on affect detection accuracy, an ANOVA was conducted on all accuracy scores obtained from the above described procedures. This is a one-way ANOVA with accuracy as the dependent variable and training strategy (SCL, WPF, WSF) as the independent variable. The analysis showed significant main effect for training strategy ($F_{(2,285)} = 57.67, p < 0.05$). Bonferroni posthoc tests revealed that accuracy scores for WPF ($M = 65.14$) were higher than those for WSF ($M = 57.81$) and SCL ($M = 55.55$). The accuracy scores using WPF and WSP were higher than SCL baseline accuracy. This indicates that we were able to leverage the dynamic learning ability of Winnow algorithm to enhance classification accuracy. We also found that WPF outperformed WSF, although we were expecting that day-specific features would provide higher performance. The explanation for the lower performance of day-specific features might be that day-data tends to have higher clustering cohesion compared to data for the same emotion category across multiple days. This suggests that using pooled features are more suitable for building predictive models of affect than using day-specific features.

Table 4 Average classification accuracy using WPF from individual channels data (%)

Subject Id	*ECG*	EMG_{cur}	EMG_{zyg}	*SC*	*RSP*
S_1	58	76	73	71	60
S_2	55	65	58	62	61
S_3	63	86	69	60	59
S_4	58	88	74	54	59
Average	58.50	78.75	68.50	61.75	59.75

(a) Valence (IAPS)

Subject Id	*ECG*	EMG_{cur}	EMG_{zyg}	*SC*	*RSP*
S_1	62	68	59	54	61
S_2	68	79	60	60	57
S_3	61	60	63	61	70
S_4	58	74	57	58	57
Average	62.25	70.25	59.75	58.25	61.25

(b) Arousal (IAPS)

Subject Id	*ECG*	EMG_{cur}	EMG_{zyg}	*SC*	*RSP*
S_1	61	71	67	71	63
S_2	64	64	61	63	63
S_3	65	67	66	60	59
S_4	57	88	74	52	63
Average	61.75	72.50	67.00	61.50	62.00

(c) Valence (self)

Subject Id	*ECG*	EMG_{cur}	EMG_{zyg}	*SC*	*RSP*
S_1	59	67	60	56	58
S_2	58	68	62	66	59
S_3	70	70	50	67	67
S_4	63	73	57	71	64
Average	62.50	69.50	57.25	65.00	62.00

(d) Arousal (self)

4.4 Physiological channels effect on affect detection accuracy

The effect of both channel and emotion on affect detection accuracy was tested using a two-way ANOVA. The ANOVA was conducted on accuracy scores obtained from the WPF training strategy. We found significant main effect of channel ($F_{(5,72)} = 12.60, p < 0.05$). Posthoc tests revealed that accuracy scores for EMG-cur ($M = 69.87$) and features fusion ($M = 72.25$) were significantly higher than for other channels ECG ($M = 62.37$), EMG-zyg ($M = 58.5$), RSP ($M = 61.62$), and SC ($M = 61.62$). We did not find significant effect for emotion category ($F_{(3,72)} = 1.14, p = 0.34$). But when the levels of emotional categories were decreased to two (valence, arousal), the effect of emotion was only marginally insignificant ($F_{(1,84)} = 3.5, p = 0.065$). However, a significant effect ($F_{(5,84)} = 2.36, p < 0.05$) for the interaction between channel and emotion was found.

The interaction effect was further explored by conducting simple effects test. The test revealed that EMG-cur ($F_{(1,84)} = 4, p < 0.05$), and EMG-zyg ($F(1, 84) = 10.44, p < 0.05$) were more useful for detecting valence than arousal. Other channels, ECG ($F_{(1,84)} = 43, p =$

Table 3 Average classification accuracies using static classification (SCL), Winnow with pooled features (WPF), and Winnow with specific features (WSP) for four participants (S_1–S_4), using features fusion from all channels (%)

	Valence (IAPS)			Arousal (IAPS)			Valence (self)			Arousal (self)		
Subject Id	*SCL*	*WPF*	*WSF*	*SCL*	*WPF*	*WSF*	*SCL*	*WPF*	*WSF*	*SCL*	*WPF*	*WSF*
S_1	59	74	64	52	72	56	52	73	68	50	63	62
S_2	54	63	61	51	69	53	53	62	59	49	65	48
S_3	50	74	61	52	75	49	51	79	64	51	76	60
S_4	50	76	54	48	73	59	50	70	54	52	69	69
Average	53.25	71.75	60.00	50.75	72.25	54.25	51.50	71.00	61.25	50.50	68.25	59.75

0.62), features fusion ($F_{(1,84)} = 70, p = 0.15$), RSP ($F_{(1,84)} = 0.07, p = 0.79$), and SC ($F_{(1,84)} = 0, p = 0.99$), were equally likely to detect both valence and arousal with the same accuracy. Figure 4 shows the interaction effect between channel and emotion. Previous research has shown that SC for example is more useful for detecting arousal than valence (Lang, 1995; Levenson, 1992). Our findings were in accordance with the literature with regards to both EMG channels. The corrugator and zygomatic EMG have always shown consistent changes with the valence component of emotion (Hamm et al, 2003). On the other hand, previous research has always considered SC as an index of arousal (Levenson, 1992), but this was not observed here.

Our results show that detecting arousal with acceptable accuracy required more physiological markers in comparison to valence. It can be seen from Fig. 4 that features fusion has the highest mean compared to other channels for detecting arousal component. The literature is somehow inconsistent in this regard, with studies reporting higher detection rates for arousal than valence (Haag et al, 2004; Lichtenstein et al, 2008) and the contrary (Kim and André, 2008). However, an interesting study conducted by Gomez et al (2009) found that induced physiological changes of participants' valence lasted longer than those of arousal which dissipates quickly. This might explain the higher detection rates of valence compared to arousal. However, it should also be noted that some researchers state that valence detection can be more difficult to detect compared to arousal as valence information is conveyed more subtly (Picard, 1997).

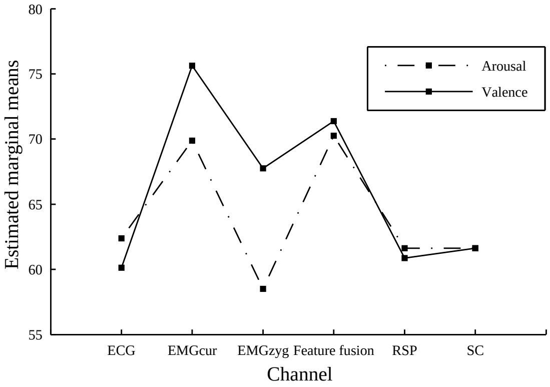


Fig. 4 Interaction effect between channel and emotion (valence and arousal).

4.5 Classifier and alpha factor effect on the performance of the Winnow ensemble

In this experiment we tested the two main factors that may have an effect on the performance of the Winnow ensemble. These are the base classifier, and the alpha factor. In order to test the classifier effect, we used a WPF strategy with different base classifiers. These are listed in Table 5. SVMs and linear perceptron are attractive methods due to their high generalization capability, while k-nearest neighbor classifier is a good choice, especially with small datasets (Alzoubi, 2012). NaiveBayes was included as a standard bench mark method. We used data from all participants, but we used one category (Valence-IAPS) only as a demonstration. In this experiment, the *alpha* factor was set to 2. We conducted one-way ANOVA on accuracy scores obtained from training the ensemble with different base classifiers. We found significant differences in the performance of the Winnow ensemble with different base classifiers ($F_{(4,75)} = 9.06, p < 0.05$). Bonferroni posthoc tests revealed that accuracy scores for SVM with a linear kernel ($M = 70.81$) and Linear perceptron ($M = 68.94$) were higher than those for K-nearest neighbor ($M = 62.25$), Naive Bayes ($M = 61.75$), and SVM with sigmoid kernel ($M = 66.31$). Figure 5 shows an error bar chart of the classifier accuracy data. The error bars show the 95% confidence interval around the mean. These results indicate that the choice of the base classifier has an effect on the performance of the ensemble. It is known that there is no classification al-

Table 5 Base classifiers description

Classifier	Description
SVM-linear	Support vector machine classifier that combines a maximal margin strategy with a kernel method to find an optimal boundary in the features space, this process is called a kernel machine. The machine is trained according to the structural risk minimization criterion. svm-linear uses a linear kernel
KNNC	K-nearest neighbor, classical instance-based algorithm; uses normalized Euclidean distance, k is optimized using the leave-one error. It assigns the class label by majority voting among nearest neighbors.
SVM-sigmoid	support vector machine classifier with sigmoid kernel.
perlc	Linear perceptron classifier, learning rate set to 0.1
naivebc	Naive Bayes, standard probabilistic classifier, the classifier assigns an example to the class that has the maximum estimated posterior probability.

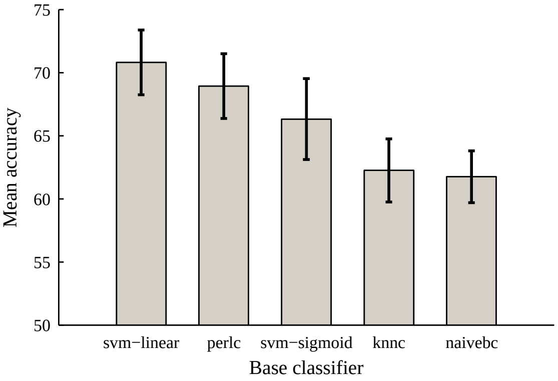


Fig. 5 Winnow ensemble accuracy data with different base classifiers (means with 95% CI) across all participants' data and categories.

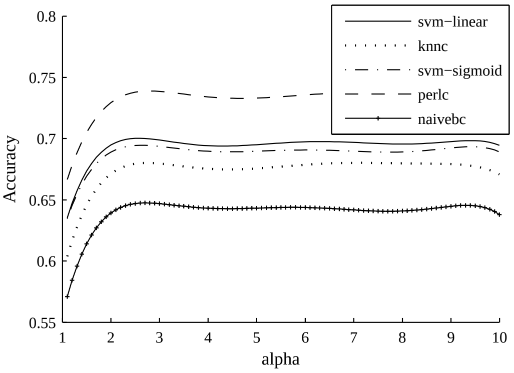


Fig. 6 Alpha factor effect on the performance of the ensemble.

gorithm that can outperform all other methods in all contexts (Duda et al, 2001).

We tested the performance of the ensemble for a range of *alpha* values (1.1–10) with the same base classifiers mentioned earlier. Figure 6 shows a plot of accuracy scores, averaged across all participants' data, against *alpha* values. It can be seen that accuracy scores peeks around ($alpha = 2$), and remains relatively stable afterward. Previous research has shown that setting *alpha* to two is a proper choice (Kuncheva, 2004a).

4.6 Individual channels decision fusion

There are a number of fusion techniques that are used to fuse affective information from multiple channels. These were discussed in Section 2.3. Table 6 shows classification accuracy scores obtained using WDF training strategy. We compare the performance of decision fusion training strategy (WDF) to that of features level fusion represented by (WPF) training strategy. The WPF selects features from all physiological channels, so it represents a features level fusion strategy. We conducted one-way ANOVA on accuracy scores obtained from both training strategies and using different base classifiers. We found significant differences in the performance of the Winnow ensemble by using the two training strategies ($F_{(1,158)} = 74.27, p < 0.05$). The mean for WDF was ($M = 75.67$), and for WPF ($M = 66.01$). This indicates that decision fusion from individual independent channels' data outperformed the features fusion strategy. This could be due to the nature of asynchronous physiological markers of affect occurring at different times within a specific time frame in response to a stimulus. Similar findings were observed in similar setups by fusing decisions obtained from EEG and peripheral physiological signals (Chanel et al, 2006).

5 Discussions and conclusions

We have shown that diagnostic physiological features of affect exhibit daily variations. This is a challenging issue for building effective physiological-based affect detection systems. In order to be able to detect affect from future physiological data, there is a need for a classification system that can handle these day/session variations. We have shown that a classifier ensemble

Table 6 Average classification accuracies obtained using WDP training strategy(%)

Subject Id	svm_{linear}	*knnc*	$svm_{sigmoid}$	*perlc*	*naivebc*
S_1	67	83	65	74	70
S_2	82	71	82	77	83
S_3	84	82	78	84	79
S_4	86	88	84	84	84
Average	79.75	81.00	77.25	79.75	79.00

(a) Valence (IAPS)

Subject Id	svm_{linear}	*knnc*	$svm_{sigmoid}$	*perlc*	*naivebc*
S_1	77	77	77	83	81
S_2	76	80	70	81	73
S_3	74	81	76	82	82
S_4	84	80	82	82	84
Average	77.75	79.50	76.25	82.00	80.00

(b) Arousal (IAPS)

Subject Id	svm_{linear}	*knnc*	$svm_{sigmoid}$	*perlc*	*naivebc*
S_1	66	67	63	75	65
S_2	82	76	81	86	87
S_3	58	71	53	76	72
S_4	85	83	83	82	83
Average	72.75	74.25	70.00	79.75	76.75

(c) Valence (self)

Subject Id	svm_{linear}	*knnc*	$svm_{sigmoid}$	*perlc*	*naivebc*
S_1	69	67	66	75	70
S_2	70	66	67	62	59
S_3	64	63	76	72	70
S_4	74	76	71	78	76
Average	69.25	68.00	70.00	71.75	68.75

(d) Arousal (self)

approach offer such a capability by combining multiple classifier's decisions and updating their weights according to their performance. This enhances the generalization capability of the system on future data. Our analysis showed that affect detection using day-specific features did not yield improved performance over using pooled features set. Our analysis also showed that a decision fusion strategy which combines decisions from classifiers trained on individual channels data of each day (WDF) outperformed a features fusion strategy (WPF). We have tested the two factors that may have an effect on the performance of the ensemble. We found that the performance of the ensemble is affected by the choice of the base classifier - SVM with a linear kernel provided robust performance. We also found that the *alpha* factor with values close to 2 provided the best performance. The facial EMG of corrugator and zygomatic were more predictive of valence than arousal compared to ECG, RSP and SC. This should have implications if designers of affect detection systems were more interested in detecting valence than arousal. This also might suggest that facial EMG is more reliable than other measures when considering affect detection over multiple sessions. Additionally, EMG-cur and a fusion of features from all channels yielded the highest recognition rates for both valence and arousal.

Adaptive classification enhanced the detection rate of affect on this type of changing data compared to static classification. However, there are a number of limitations with this approach. First, adaptation comes with the cost of additional complexity, and computation time. This might not be favorable if time for making decisions is critical. Second, there is no single adaption system that fits all uses, since it is application dependent. Third, the updating mechanism of the system might require (at some point) the existence of true class labels (ground truth). The absence of ground truth when needed might affect the performance of an adaptive classification system. Although ground truth labels could be potentially estimated using unsupervised techniques, this comes with additional complexity and time cost. Alternatively, ground truth labels can be obtained from users when the confidence about the predicted class label drops below a certain level and a user intervention is needed in order to maintain the consistency of the system. This might not seem practical, however obtaining periodic and sparse self-reports of emotions from users will help maintain the effective operation and consistency of the system.

When more data becomes available, the ensemble size may grow unbound. In order to control the size of the ensemble, making structural changes for the ensemble is inevitable. This can be done by removing some of lowest performing member classifiers. This is necessary when fast decisions in real time is required. One primary advantage of the use of the ensemble is that there is no need to retrain existing member classifiers; only newly added member classifiers are needed to be trained on the newly available data, thereby saving time when these kinds of structural changes are required.

This study has provided evidence on the time-varying nature of affective physiological data as indicated from data that was acquired on multiple recording sessions. This characteristic of the data affected the performance of static classifiers which assume stationarity of the data considerably, with performance near baseline. This has major implication on building automatic physiology-based affect detectors. As an alternative to static classification approach, an updatable ensemble-based classification approach proved to offer significant performance enhancement over static classifiers. Although the results of the ensemble were moderate, they warrant improvement via further research. Classifiers ensemble with an update mechanism could possibly offer solutions to many problems that result from the time-varying nature of physiological data. For example, the classifiers ensemble approach might be able to address changes in the classification environment which include 1) data distribution changes (features space), as is the case when data is obtained from different days or sessions, 2) changes in class distribution, which are quite prevalent during naturalistic interactions, 3) changes in diagnostic features, where features for discriminating particular affective states may change over time, 4) the introduction of new users over time, i.e., building user-independent models. Therefore, an updatable ensemble-based modeling technique might be a more practical option for building real-world affect detection systems than static classifiers which are trained on initial data and are never updated to reflect new data.

There are two primary limitations with the present study. One limitation of our work is the relatively small sample size, so replication with a larger sample is warranted. The second limitation is that emotions were artificially induced rather than spontaneously experienced. This approach was adopted because strict laboratory control was desired for assessing the day-data phenomenon. These types of methodological issues can only be solved in larger studies. Thus, replicating this research in more naturalistic contexts is an important step for future work. Naturalistic affective interactions cover a wide range of applications; one such example that stirred interest from researchers is recording emotional responses of call center workers going under various degrees of stress.

Acknowledgements Omar AlZoubi and Davide Fossati are supported by award NPRP 5-939-1-115 from the Qatar National Research Fund. Sidney DMello is supported by NSF grants (HCC 0834847, and DRL 1235958).

References

Allanson J, Fairclough SH (2004) A research agenda for physiological computing. Interacting with Computers 16(5):857–878

Alzoubi O (2012) Automatic affect detection from physiological signals: Practical issues. PhD thesis, University of Sydney

AlZoubi O, Calvo RA, Stevens RH (2009) Classification of eeg for affect recognition: an adaptive approach. In: AI 2009: Advances in Artificial Intelligence, Springer, pp 52–61

AlZoubi O, Hussain MS, D'Mello S, Calvo RA (2011) Affective modeling from multichannel physiology: analysis of day differences. In: Proceedings of the 4th international conference on Affective computing and intelligent interaction-Volume Part I, Springer-Verlag, pp 4–13

AlZoubi O, D'Mello SK, Calvo RA (2012) Detecting naturalistic expressions of nonbasic affect using physiological signals. IEEE Transactions on Affective Computing 3(3):298–310

Andreassi JL (2007) Psychophysiology: Human behavior and physiological response, 5th edn. Lawrence Erlbaum Associates, Publishers, New Jersey

Angelov P, Filev DP, Kasabov N (2010) Evolving intelligent systems: methodology and applications, vol 12. John Wiley & Sons, New York

Bifet A, Holmes G, Pfahringer B, Kirkby R, Gavaldà R (2009) New ensemble methods for evolving data streams. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '09, pp 139–148

Bradley M, Lang PJ (2007) The international affective picture system (iaps) in the study of emotion and attention, Oxford University Press, New York:, pp 29–46

van den Broek EL, Schut MH, Westerink JHDM, Tuinenbreijer K (2009) Unobtrusive sensing of emotions (use). Journal of Ambient Intelligence and Smart Environments 1(3):287–299

Chanel G, Kronegg J, Grandjean D, Pun T (2006) Emotion assessment: Arousal evaluation using eegs and peripheral physiological signals. In: Multimedia content representation, classification and security, Springer, pp 530–537

Cieslak D, Chawla N (2009) A framework for monitoring classifiers performance: When and why failure occurs? Knowledge and Information Systems 18(1):83–109

Duda R, Hart P, Stork D (2001) Pattern Classification. John Wiley & Sons, New York

Ekman P (1992) An argument for basic emotions. Cognition and Emotion 6(3):169–200

Ekman P (1994) Moods, emotions and traits, Oxford University Press, New York, pp 56–58

Gomez P, Zimmermann PG, Schär SG, Danuser B (2009) Valence lasts longer than arousal. Journal of Psychophysiology 23(1):7–17

Haag A, Goronzy S, Schaich P, Williams J (2004) Emotion recognition using bio-sensors: First steps towards an automatic system. In: Affective dialogue systems, Springer, pp 36–48

Hamm AO, Schupp HT, Weike AI (2003) Motivational organization of emotions: Autonomic changes, cortical responses, and reflex modulation. Handbook of affective sciences pp 187–211

Heijden F, Duin R, Ridder D, Tax D (2004) Classification, parameter estimation and state estimation - an engineering approach using Matlab. John Wiley and Sons, Chichester

Jain AK, Duin RPW, Jianchang M (2000) Statistical pattern recognition: a review. Pattern Analysis and Machine Intelligence, IEEE Transactions on 22(1):4–37

Kim J, André E (2006) Emotion recognition using physiological and speech signal in short-term observation. In: Perception and Interactive Technologies, Springer Berlin Heidelberg, pp 53–64

Kim J, André E (2008) Emotion recognition based on physiological changes in music listening. IEEE Trans Pattern Anal Mach Intell 30(12):2067–2083

Kim K, Bang S, Kim S (2004) Emotion recognition system using short-term monitoring of physiological signals. Medical and Biological Engineering and Computing 42(3):419–427

Kolter JZ, Maloof M (2003) Dynamic weighted majority: A new ensemble method for tracking concept drift. In: Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, IEEE, pp 123–130

Kreibig SD (2010) Autonomic nervous system activity in emotion: A review. Biological Psychology 84(3):394–421

Kuncheva L (2004a) Classifier ensembles for changing environments. In: Roli F, Kittler J, Windeatt T (eds) Multiple

Classifier Systems, Lecture Notes in Computer Science, vol 3077, Springer Berlin Heidelberg, pp 1–15

Kuncheva L (2004b) Combining pattern classifiers : methods and algorithms. J. Wiley, Hoboken, NJ

Kuncheva L, Christy T, Pierce I, Mansoor S (2011) Multimodal biometric emotion recognition using classifier ensembles. In: Mehrotra K, Mohan C, Oh J, Varshney P, Ali M (eds) Modern Approaches in Applied Intelligence, Lecture Notes in Computer Science, vol 6703, Springer Berlin / Heidelberg, pp 317–326

Lang PJ (1995) The emotion probe. studies of motivation and attention. American Psychologist 50(5):372–85

Lang PJ, Bradley MM, Cuthbert BN (1995) International affective picture system (iaps): Technical manual and affective ratings. Gainesville, FL: The Center for Research in Psychophysiology, University of Florida

Lang PJ, Bradley MM, Cuthbert BN, et al (2005) International affective picture system (IAPS): Affective ratings of pictures and instruction manual. NIMH, Center for the Study of Emotion & Attention

Last M (2002) Online classification of nonstationary data streams. Intell Data Anal 6(2):129–147

Lazarus R (1991) Emotion and adaptation. Oxford University Press, New York

Lee H, Shackman A, Jackson D, Davidson R (2009) Test-retest reliability of voluntary emotion regulation. Psychophysiology 46(4):874–879

Levenson RW (1992) Autonomic nervous system differences among emotions. Psychological science 3(1):23–27

Lichtenstein A, Oehme A, Kupschick S, Jrgensohn T (2008) Comparing two emotion models for deriving affective states from physiological data. In: Peter C, Beale R (eds) Affect and Emotion in Human-Computer Interaction, Lecture Notes in Computer Science, vol 4868, Springer Berlin Heidelberg, pp 35–50

Lowne DR, Roberts SJ, Garnett R (2010) Sequential non-stationary dynamic classification with sparse feedback. Pattern Recognition 43(3):897–905

Maier-Hein L, Metze F, Schultz T, Waibel A (2005) Session independent non-audible speech recognition using surface electromyography. In: Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on, pp 331–336

Muhlbaier M, Polikar R (2007) An ensemble approach for incremental learning in nonstationary environments. In: Haindl M, Kittler J, Roli F (eds) Multiple Classifier Systems, vol 4472, Springer Berlin Heidelberg, pp 490–500

Nishida K, Yamauchi K, Omori T (2005) Ace: Adaptive classifiers-ensemble system for concept-drifting environments. In: Oza N, Polikar R, Kittler J, Roli F (eds) Multiple Classifier Systems, Springer Berlin / Heidelberg, Lecture Notes in Computer Science, vol 3541, pp 176–185

Oza NC, Russell S (2001) Online bagging and boosting. In: Artificial Intelligence and Statistics, pp 105–112

Oza NC, Tumer K (2008) Classifier ensembles: Select real-world applications. Information Fusion 9(1):4–20

Picard RW (1997) Affective Computing, 2nd edn. The MIT Press, Cambridge, Massachusetts

Picard RW, Vyzas E, Healey J (2001) Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Trans Pattern Anal Mach Intell 23(10):1175–1191

Plarre K, Raij A, Hossain M, Ali A, Nakajima M, al Absi M, Ertin E, Kamarck T, Kumar S, Scott M, Siewiorek D, Smailagic A, Wittmers L (2011) Continuous inference of psychological stress from sensory measurements collected in the natural environment. In: Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on, pp 97–108

Polikar R (2006) Ensemble based systems in decision making. Circuits and Systems Magazine, IEEE 6(3):21–45

Polikar R, Upda L, Upda SS, Honavar V (2001) Learn++: An incremental learning algorithm for supervised neural networks. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 31(4):497–508

Russell JA (1980) A circumplex model of affect. Journal of Personality and Social Psychology 39:1178, 1161–1178, 1161

Russell JA, Weiss A, Mendelsohn GA (1989) Affect grid: a single-item scale of pleasure and arousal. Journal of Personality and Social psychology 57(3):493–502

Sayed-Mouchaweh M, Lughofer E (2012) Learning in non-stationary environments. Springer, New York

Vyzas E, Picard RW (1998) Affective pattern classification. In: AAAI Fall Symposium Series: Emotional and Intelligent: The Tangled Knot of Cognition, pp 176–182

Wagner J (2009) Augsburg biosignal toolbox (aubt). URL `http://hcm-lab.de/files/project_content/33/219_AuBTGuide.pdf`, online; accessed: 2014-4-25

Wagner J, Kim J, André E (2005) From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In: Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, pp 940–943

Webb AR (2002) Statistical pattern recognition. Wiley, West Sussex, England; New Jersey

Whang M, Lim J (2008) A physiological approach to affective computing. In: Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition, I-Tech Education and Publishing, pp 310–318

Widmer G, Kubat M (1996) Learning in the presence of concept drift and hidden contexts. Mach Learn 23(1):69–101

Witten I, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques, second edition edn. Series in Data Management Systems, Morgan Kaufmann

Yue S, Guojun M, Xu L, Chunnian L (2007) Mining concept drifts from data streams based on multi-classifiers. In: Advanced Information Networking and Applications Workshops, 2007, AINAW '07. 21st International Conference on, vol 2, pp 257–263

Zeng ZH, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: Audio, visual, and spontaneous expressions. Ieee Transactions on Pattern Analysis and Machine Intelligence 31(1):39–58