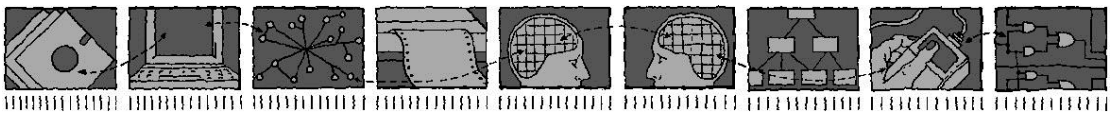
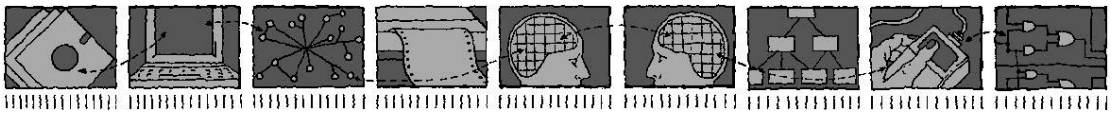


Computing Science and Mathematics
University of Stirling



Predicting Emotional Dysregulation

Kenneth J. Turner, Brian O'Neill,
Gary Cornelius, Evan H. Magill

Technical Report CSM-200

ISSN 1460-9673

September 2017

Computing Science and Mathematics
University of Stirling

Predicting Emotional Dysregulation

Kenneth J. Turner[1], Brian O'Neill[2], Gary Cornelius[3], Evan H. Magill[4]

[1,4] Computing Science, University of Stirling, Stirling, FK9 4LA, UK
Email kjt@cs.stir.ac.uk, ehm@cs.stir.ac.uk

[2] Brain Injury Rehabilitation Trust, Glasgow, G21 1UU, UK
Email brian.oneill@thedtgroup.org

[3] Rapport Network CIC, Dunblane, FK15 0AX, UK
Email gary.cornelius@rapport.net

Technical Report CSM-200

ISSN 1460-9673

September 2017

Abstract

The background is given to mental health problems and their manifestation as emotional dysregulation (inability to control emotion). This motivates the work in the paper to predict emotional episodes using physiological data collected from mass-market devices (smartwatches). The design of a system to achieve this is explained. Data is collected from two commercial smartwatches for analysis and training of classifiers. Using colour-coded indications, feedback is given to staff and patients about the risk of an emotional episode. Medical staff can view detailed patient charts in order to make a considered judgement about the risk and how best to mitigate this. An overview is given of how features are extracted from the physiological data and then fed into machine learning algorithms that predict the likelihood of an episode. Results are presented for a small-scale evaluation with a partner hospital that looks after patients with brain injuries. The approach achieves a typical prediction accuracy up to 82% (aggressive episodes) or up to 68% (normal behaviour), anticipating episodes up to 4 hours ahead (physical data) or up to 4 days ahead (sleep data). Pointers are given to future developments of the work.

Keywords: Assisted Living, Emotional Dysregulation, Machine Learning, Physiological Data, Smartwatch, Wearable Device

1 Introduction

The medical background of the research is explained, along with its motivation and objectives. Related work in the field is discussed.

1.1 Background

Mental health problems place a significant burden on society due to the cost of medical care, the direct loss of working time, and the substantial amount of informal care that is often required. The following figures give an idea of the scale of the problem.

General mental health issues in the UK, for example, are estimated to cost £100 billion per year [14]. Of mental health problems, depression is the major issue. In the United States major depressive disorder affected over 15 million people in 2010, with an associated cost to society of $211 billion [18]. In the UK £9 billion is lost per year due to depression.

Cognitive impairment is widespread and has various causes including the dementias, stroke, traumatic injury, learning difficulties and mental illness. The annual cost of formal and informal care provision for cognitive impairment is over £66 billion in the UK [13] and over $300 billion in the United States [23].

Dementia in various forms is the largest cause of cognitive impairment. The world-wide cost of dementia in 2015 was US $818 billion, affecting 47 million people [3]. An 85% increase in this is predicted by 2030, with the bulk of the cost relating to care provision. In the UK the number of people with dementia is currently around 850,000, and is predicted to rise to 1,740,000 by 2051 [25].

Acquired brain injury is typically divided into traumatic brain injury and non-traumatic brain injury, the latter arising from strokes, tumours, infectious diseases, hypoxia, hypoglycaemia and neurotoxins. The annual cost of traumatic brain injury was estimated in 2010 to be approximately £4.1 billion in Europe [19]. Among non-traumatic brain injuries, stroke is the most prevalent; annual cost estimates in the UK exceed £7 billion [15]. In general, a large proportion of the costs comes from the loss of productive activity, and the need for ongoing care due to physical, cognitive, emotional and behavioural barriers to societal participation.

1.2 Motivation and Objectives

Emotional dysregulation is a reduced capacity to control one's own affective state, and to modulate the intensity of emotional and behavioural responses to environmental stimuli [31]. Difficulty controlling emotional responses is a key feature of many mental health conditions including the anxiety disorders, depression, the dementias, schizophrenia, bipolar disorder, personality disorder, attention deficit hyperactivity disorder and acquired brain injury (e.g. [33], pp. 135–154). Medical management of emotional dysregulation is predominant, but is of limited use in neurological conditions such as acquired brain injury and the dementias. Increased survival rates and longevity mean that these conditions are a growing public health concern.

Treatment of emotional dysregulation needs to be founded on accurate measurement of emotional state. More effective medical care can be enabled by increasing the awareness of patient mental state. Patients can also use feedback on mental state to regulate their emotions more effectively.

The general aim of the work reported here is to help in predicting all forms of emotional dysregulation: anger, anxiety, depression, manic behaviour, panic attacks, etc. However to give the work a concrete application, the initial focus has been on a hospital specialising in brain injury. BIRT (Brain Injury Rehabilitation Trust, *http://www.thedtgroup.org/brain-injury*) operates 14 hospitals across the UK to help patients with a variety of acquired brain injuries, e.g. due to physical trauma, stroke or dementia. BIRT in Glasgow was the clinical partner for the work of this article.

An issue for the BIRT hospitals is that patients often exhibit emotional dysregulation. This typically takes the form of emotional outbursts such as verbal or physical aggression. It can be difficult to predict when these outbursts might occur, so staff and patients have to deal with the situation when it happens. This can be stressful, and can have an adverse effect on other patients and on the patient-carer relationship.

This article reports the approach of a project that aims to predict and thus to help manage emotional dysregulation. TEAMED (Technology Evaluating And Measuring Emotional Dysregulation, *http://www.cs.stir.ac.uk/projects/teamed*) has developed, integrated and assessed a variety of technologies to establish a patient's emotional state. The goal is for patients with emotional problems to receive better care, to achieve a better quality of life, and to self-manage their conditions better.

Physiological changes that are likely to indicate changes in emotional state include heart rate variability [35], galvanic skin resistance [27], and movement [11]. This article is believed to be the first

that reports analysis of physiological data in free-behaving persons with disorders of emotional regulation.

The objectives of the work have been:

- to interface commercial, off-the-shelf smartwatches for collection of physiological data of relevance to emotional dysregulation
- to investigate machine learning techniques as a means of analysing this kind of data
- to determine whether emotional outbursts can be successfully predicted based on physiological measurements
- to provide timely feedback to medical staff and patients about the risk of an emotional outburst
- to carry out a preliminary assessment of the approach in a small evaluation with a limited number of patients.

The long-term aim is to support patients in three contexts:

- Clinical: In-patients have an infrastructure to collect, store and analyse physiological data from the smartwatches. Medical staff are on hand to monitor predictions of emotional dysregulation and to intervene appropriately.
- Domiciliary: Out-patients at home can be supported through an Internet connection (broadband or cellular) that collects and forwards data to a clinic. In this setting, patients would have to self-assess their own behaviour. Local feedback would be important for alerting patients to the risk of emotional dysregulation.
- Mobile: Out-patients away from home, e.g. shopping or visiting, can again be supported through an Internet connection (cellular). Self-assessment and local feedback (via a mobile phone) would be needed.

Of these contexts, the clinical one is the easiest to manage since professional help is at hand to assess behaviour during the training phase and to mitigate the risk of emotional dysregulation. Secure and reliable data storage is also available. However, both domiciliary and mobile contexts are also feasible and will be investigated in future work. The main issue with these would be the need for self-assessment of behaviour. Professional assessment could be employed in a clinical setting initially prior to discharge. However if the patient's condition is likely to change in future then some self-assessment would be desirable.

Medical applications of the work require a condition that manifests itself in episodic behaviour that can be readily assessed, and can be correlated with physical measurements from a wearable device. This holds for the initial application area of aggressive behaviour due to acquired brain injury. However there is the prospect of applying the approach to other conditions. For example, the project approach is now being used to support patients with dementia who are prone to distress. Other conditions such as anxiety attacks, bipolar disorder, depression (often associated with anxiety) and epilepsy are also promising applications. These are associated with observable behaviour such as agitation, hyperactivity, lassitude or seizures, and could hopefully be predicted using physical measurements such as anomalous heart rate, physical movement, skin resistance, skin temperature and sleep patterns. Besides mental health problems, it may also be feasible to handle more general kinds of health issues.

1.3 Relationship to Other Work

The TEAMED project is distinguished from other work in a number of significant ways that have conditioned the approach:

- The aim is to predict episodes of emotional dysregulation (e.g. overt aggression) rather than identifying emotional state *per se*. The focus is on healthcare application rather than on psychological science.
- Physiological data is collected using mass-market devices (smartwatches) rather than specialised devices (e.g. ECGs, EEGs, spirometers). This is a much cheaper and more pragmatic solution, though the data is likely to be less accurate than from certified medical equipment.
- Data is collected in an unsupervised fashion (ambulatory patients wearing watches around a hospital, at home or on the move) rather than in a tightly-controlled environment (e.g. test subjects in a laboratory). Again this may result in less accurate data, e.g. because a watch is worn too loosely for a reliable indication of heart rate or skin resistance.
- Data is collected and analysed long-term on a fairly continuous basis, rather than in the short-term fashion that is typical of lab-based evaluations. As a result, there can be day-to-day or even seasonal changes because of natural variations in physiological response as

well as variations in data measurements (e.g. due to how a smartwatch is worn or to different environmental conditions).

- The approach has to deal with frequent gaps in the data ranging from tens of minutes to days (Section 4.1 explains why this may happen). This presents a significant challenge for the analysis since it must be robust in the face of data gaps.

The link between emotional state and physiological response has long been recognised. For example William James investigated this in the 1880s, while in the 1900s Karl Jung analysed GSR (Galvanic Skin Response) in relation to negative aspects of word associations.

There has been considerable research into emotional state which has highlighted the need to distinguish various affective phenomena such as feeling, mood, attitude, emotion, etc. [29]. A common approach is to represent emotion as points in a two-dimensional plane such as the Russell Circumplex Model [28] or the Geneva Emotion Wheel [32]. Typically, valence (emotional label) is shown on a horizontal axis and arousal level on a vertical axis. Although this is useful background, the aim of TEAMED is not to identify emotions. Rather the goal is to predict observable behaviour such as agitation, aggression, hyperactivity or lassitude.

In psychology an affect is behaviour resulting from emotional state. Affective computing deals with computer-based systems that can recognise, interpret and even simulate human emotions [27]. This approach has been used to recognise emotions, to enable self-awareness of emotional state, and to improve human-computer communication.

A number of techniques have been used to measure physiological parameters associated with affect. These include electrodermal activity (skin data, [27]), electrocardiograms (heart data, [35]), electroencephalograms (brain data, [6]), electromyograms (muscle data, [36]), environmental temperature [5], facial expression [16] and respiration [9].

The approach of combining sensors to identify specific emotions as felt by the wearer has had partial success [20]. The limited accuracy is perhaps due to the complexity of relating multiple data streams, various emotion labels and different intensity levels. TEAMED has focused on consequences of emotions rather than their specific recognition. The goal was to interpret physiological measurements as predictors of a binary outcome: the occurrence of an emotional event or its absence. More pragmatically, the project has aimed to discover whether physiological data from mass-market devices, rather than specialised ones, can be used for predictions in healthcare applications.

The approach of [20] resembles TEAMED in that it aimed to assess emotions in a real-world situations rather than in a lab. Discrete devices were used for heart rate, skin resistance and movement (unlike the single smartwatches used on TEAMED). Participants were asked to record their emotions using a Mood Map similar to the Circumplex model. It was found that self-reporting of emotional state could be problematic for a variety of reasons such as the need for subjective self-assessment. Another issue was that mental state is usually a mixture of several emotions. For the TEAMED work so far, classification has been undertaken on the basis of a professional assessment that identifies the nature and severity of emotional outbursts.

[34] aimed to identify mental stress in mobile patients. A challenge in that work was separating physiological responses due to stress from changes arising due to physical activity. Like [20], discrete sensors were used. Baseline physiological measurements were obtained for sitting, standing and walking, allowing changes due to stress to be identified. So far, TEAMED has focused on ambulatory but relatively static in-patients. When the approach is extended to more mobile out-patients, it is expected that solutions may similarly be needed due to the effect on measurements of physical activity.

As pointed out in [4], an issue in classifying physiological data can be the variability of measurements over time. This can arise due to varying activities, natural changes in physiology, or different environmental conditions (e.g. a hot, humid or cold day). The use of pooled day data has been advocated as a way of compensating for this. When training classifiers on the TEAMED project, relatively lengthy data sets are used (e.g. 4 weeks) so daily variations are effectively taken into account.

TEAMED is an example of approaches that use feature-based models [22]. The general idea is to extract features from physiological signals and use these to recognise affective phenomena. It has been argued that feature-based models can usefully be combined with sequential models [21]. For example, Hidden Markov Models can be used to predict the evolution of an affect (e.g. its gradual onset and subsequent decay). For the initial application of TEAMED (predicting aggressive outbursts) a feature-driven approach was judged to be most appropriate, but sequential aspects will be investigated in future.

The measurements collected by TEAMED can be viewed as multivariate time series data. This is common in clinical applications where a series of measurements is made over time (e.g. blood pressure, body temperature or heart rate). Most classifiers assume that data samples are independent and identically distributed, which may not be the case for a time series. The data may be collected at

precise intervals or more irregularly. If data collection is periodic, signal processing techniques such as Hidden Markov Models or recurrent neural networks can be used.

However, if data collection is irregular then such techniques cannot be used. Instead temporal abstractions can be used to replace time series data with interval-based abstractions over values [2]. Various temporal relationships can be established such as *before*, *overlaps* or *during*. The aim of such an approach is to address the issue of correlated data samples, allowing conventional classifiers to be used.

This is the technique used by the RÉSUMÉ system [30] that has been applied to medical applications such as monitoring children's growth, diabetes treatment and AIDS therapy. The system is able to identify significant classes of value (e.g. low, medium high) as well as temporal relationships (e.g. increasing, steady, decreasing).

A similar approach is taken by the Minimal Predictive Temporal Patterns framework [7]. This is able to automatically identify value abstractions and trend abstractions much as RÉSUMÉ does. The work was evaluated on data from patients who had undergone cardiac surgery. The use of an anticoagulant with these patients may lead to an anomalous platelet count and hence thrombosis. Based on historical data about when platelet tests are ordered, the approach is able to predict the need for tests more successfully when temporal patterns are used.

Temporal abstractions have also been used to infer medical knowledge about patterns. The KLS framework [26] aims to efficiently discover Time Interval Related Patterns in the style of [2]. The value of the approach has been demonstrated through applications to datasets for diabetes, hepatitis and intensive care.

The work on temporal abstractions is broadly relevant to the TEAMED project. As explained in Section 4.1, there are fairly frequent gaps in the data such that signal processing techniques would not be appropriate. However, the time series are sufficiently continuous that it is reasonable to abstract the data in fixed periods (hourly for physical data, daily for sleep data). Section 3.2 describes the abstractions that are used, though these are different from those of, say, [7].

For TEAMED purposes, the abstractions were motivated by insights from the clinical partner and from an early study of data patterns relevant to predicting aggressive behaviour. In general, absolute data values are not used for the reasons discussed in Section 3.2; rather, changes over time are abstracted. Although minute-by-minute data is similar due to common underlying physiological mechanisms, as Figure 3 shows the hour-by-hour data varies significantly. It is therefore believed that the abstracted features are sufficiently decoupled over time for meaningful classification.

The current focus of TEAMED is on prediction of aggressive outbursts, so formal assessment of aggression is needed. The Overt Aggression Scale (OAS [37]) is a classification commonly employed in psychiatry to describe aggressive behaviour. The Overt Aggression Scale Modified for Neurorehabilitation (OAS-MNR [1]) is an adaptation of this for patients who are undergoing rehabilitation following an acquired brain injury. For this scale, aggressive behaviour is placed into four categories: verbal aggression, physical aggression against objects, physical aggression against self, or physical aggression against other people (carers or patients). For each category the behaviour may be at one of four levels: mild, minor, moderate, or severe.

1.4 Structure of The Paper

Section 2 presents the overall design of the system for collecting and analysing physiological data. Section 3 discusses how the data was analysed in order to predict emotional dysregulation. Section 4 describes a small-scale evaluation of the approach and the results from this. Section 5 summarises the key points and indicates future work.

2 System Design

The TEAMED system collects physiological data for a number of patients and then analyses it in order to predict patient behaviour. The overall structure of the system is shown in Figure 1 and is explained in more detail below.

2.1 Data Measurement

Smartwatches offer a variety of sensors, wireless communication with a mobile phone, and additional capabilities such as message display and phone call functions. Smartwatches are a common example of a wearable device. Many kinds of smartwatch are available, falling roughly into three categories: fitness watches (for assessing exercise), health watches (for measuring health-related factors), and utility watches (to complement a mobile phone).

For the purposes of the TEAMED project, smartwatches are ideal as they look like watches (and are therefore acceptable to patients), are mass-market (and are therefore affordable), and can support measurement of relevant physiological data. Smartwatches are convenient for patients as they are familiar, robust and comfortable. For out-patients, they can be used and kept charged with minimal training.

In principle other kinds of devices could be used. For example a chest-band could monitor heart rate, finger electrodes could measure skin resistance, or a digital sphygmanometer could assess blood pressure. In a clinical setting such devices are appropriate and achieve greater accuracy than is possible with a smartwatch. However, they are much less practical for a number of reasons. They can be awkward to use and to wear on a continual basis. They are much more expensive than a smartwatch and would require multiple data sources to be coordinated. They would also require multiple devices to be charged on a regular basis. In comparison a smartwatch offers a single, acceptable and cost-effective solution that can be used by patients with minimal training and maintenance.

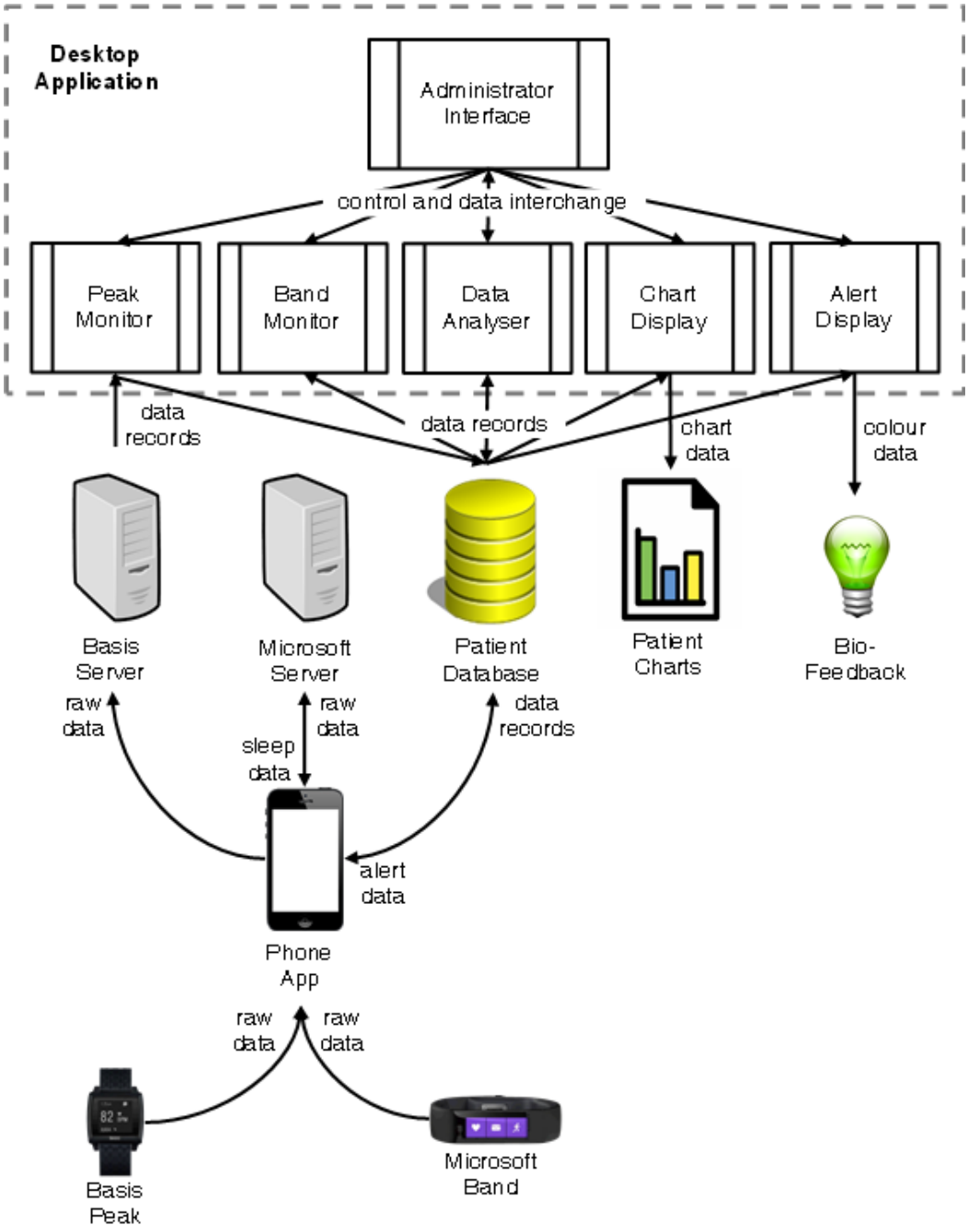


Figure 1. System Structure

Ideally a smartwatch should be able to communicate with a central hub in the hospital or home. However, the present communication range of smartwatches (10m at best, but often less than this) means that this may be infeasible. Most smartwatches are designed on the assumption that a mobile phone will be nearby, typically carried by the user. Fortunately this is now common practice for many people. A smartwatch usually sends data to the phone using BLE (Bluetooth Low Energy), and the phone forwards this to the manufacturer's server using a secure Internet connection (WiFi or cellular).

Requirements were agreed with the clinical partner for the devices to be used by TEAMED. Some smartwatches were not considered because they look more like tracking bands than watches and so are viewed with some suspicion by patients. Other smartwatches were excluded due to their price or lack of a suitable SDK (Software Development Kit). The main selection criterion was that the smartwatches should provide the kinds of physiological data that were likely to be useful in identifying emotional dysregulation. Based on previous reports in the literature, the following measurements were expected to be relevant for this: heart rate (and heart rate variability), physical movement, skin resistance, skin temperature, sleep quality and sleep total.

For the initial work of TEAMED, two makes of smartwatch were selected as the best fit to requirements: the Basis Peak and the Microsoft Band 2. However, the approach is agnostic regarding the kind of watch and would allow the use of other kinds of smartwatch or digital device in future. For example, a digital sphygmomanometer could be useful to measure blood pressure (a difficult task on a smartwatch). The two kinds of watch used on the project are broadly similar in functionality but differ a lot in detail.

Both the Basis Peak and the Microsoft Band are able to measure key factors that were expected to be relevant for predicting emotional episodes. Due to differing timescales, these physiological measurements are classified as physical data (heart rate, skin resistance, skin temperature, step count) or sleep data (sleep quality, sleep total). The Microsoft Band is also able to report heart rate variability. Although smartwatches may have additional sensors such as for orientation, atmospheric pressure, air temperature or UV, the clinical partner did not expect this kind of data to be useful for predicting emotional outbursts so it is not analysed.

Battery life for the Basis Peak is up to 4 days, which is reasonably convenient in practice as recharges need to be scheduled only every 3 days. Battery life for the Microsoft Band is up to 2 days, which in practice means that it needs a daily recharge of an hour or two. For the Band a daily schedule is needed for recharging, choosing a time when loss of data is unlikely to be significant. Overnight charging is not appropriate as the watches need to be worn in order to gather sleep data.

The selected watches differ in their communication range. In practice it was found that the Basis Peak has an effective range of 2 to 4 metres while the Microsoft Band has an effective range of 5 to 10 metres. The kinds of patients supported by TEAMED are mobile, so this means the patient needs to carry the associated mobile phone (or leave it by the bedside overnight).

Both kinds of watch will lose communication with the phone if these become too separated. In theory the watch and the phone should resynchronise when they come together again. In practice the Basis Peak often fails to do this and requires a manual resynchronisation. This is viable in a clinical setting, where the medical staff will be alerted to the lack of synchronisation and can deal with this. If the Basis Peak were used by an out-patient, this could be more problematic.

Data collection approaches vary according to the smartwatch. Measurements are made on the watch and transmitted wirelessly to a smartphone. Most watches such as the Basis Peak then forward the raw data to the manufacturer's server where it can be downloaded by an authorised application. Some watches such as the Microsoft Band make the raw data available only to authorised applications on the associated smartphone. This means that data is collected quite differently for the two watches used on the project. If there is a break in communication, the Basis Peak will store data for later transfer. However raw data is obtained live on the phone for the Microsoft Band, so loss of communication means loss of data.

2.2 Data Storage

Physiological data is collected and stored from various sensors. Because of differing timescales, physiological data is split in two categories: physical data (minute-by-minute measurements of heart rate, skin resistance, etc.) and sleep data (day-by-day measurements of sleep quality and total).

Although physical measurements are made in near-real-time, they are averaged per minute before storage. For some smartwatches such as the Basis Peak, this is imposed by the manufacturer to reduce transmissions and therefore to save battery life. For other smartwatches such as the Microsoft Band, which supports a real-time data feed, several samples are averaged per minute by the TEAMED software for the same reason. Per-minute averaging is desirable anyway as the data may be noisy (particularly heart rate). For the purposes of TEAMED it is not necessary to have a real-time feed of, say, heart beats.

The data is preprocessed to validate its integrity and to remove outliers. For example, skin resistance measurements often go out of range: very low if the user showers or sweats after exertion, or very high if the user wears the watch loosely or removes it temporarily. After preprocessing, the data is stored in a secure central database.

Ideally the patient database would be physically within the clinic or hospital that collects the patient data. However in the experimental work of TEAMED, the database was located with the academic partner. For confidentiality and security, each patient's data is stored anonymously in a separate logical database with different access credentials, and data transfer uses a secure network connection.

Although a regular relational database could have been used, a time-series database was preferred as this is optimised for large volumes of real-time data and for efficient time-based queries. InfluxDB (*https://influxdata.com*) was selected to store the TEAMED data. Each patient database contains the following tables (series in InfluxDB terminology):

- technical data
- physical data
- sleep data
- report data
- prediction data.

Each data record has an associated timestamp (always UTC/GMT, but presented as local time in the user interface). To handle data collected over a long period, a retention policy can be set so that large amounts of out-of-date information do not clutter the database.

The technical data includes the kind of smartwatch and the preferred means of patient feedback. The report data contains observations entered by medical staff about previous emotional episodes; this is used to train the prediction of future emotional outbursts. The prediction data contains hour-by-hour predictions of how likely an emotional episode is to occur; this is used retrospectively to assess how well the system has performed.

The BIRT hospitals associated with the project already keep records of aggressive behaviour on the OAS-MNR aggression scale mentioned in Section 1.3, so the same information is entered into the TEAMED database. Aggressive behaviour often occurs several times in a short period, and is recorded by the hospital as one episode with several behaviours. As an example, one 15-minute episode during the evaluation had 15 instances of mild verbal aggression, 10 instances of minor verbal aggression, 10 instances of moderate verbal aggression, and 1 instance of minor physical aggression against other people.

Figure 2 illustrates how an aggressive outburst is reported for a patient using the TEAMED desktop application: here, two instances of moderate physical aggression against other people. Currently only aggression reports are supported, but the approach can be readily extended for other kinds of reports (e.g. for manic behaviour or anxiety attacks). Reports for in-patients are made by medical staff. Reports for out-patients could be submitted by a modified version of the TEAMED mobile phone application.

2.3 Prediction and Feedback

As described in Section 3, the physiological data is analysed in order to predict episodes of emotional dysregulation. Feedback on this is provided to medical staff and also to patients.

Using the desktop application, medical staff are able to review physiological data. Charts of this data are useful when monitoring a patient at risk of emotional dysregulation. Figure 3 shows an example set of charts produced by the desktop application (using JFreeChart, *http://www.jfree.org/jfreechart*). Most charts show measurements on a linear scale, but skin resistance is shown on a log scale because it varies so widely. The charts are interactive in various ways. For example, hovering over a chart will display time and value for each data point. It is also possible to zoom into a chart for more detail.

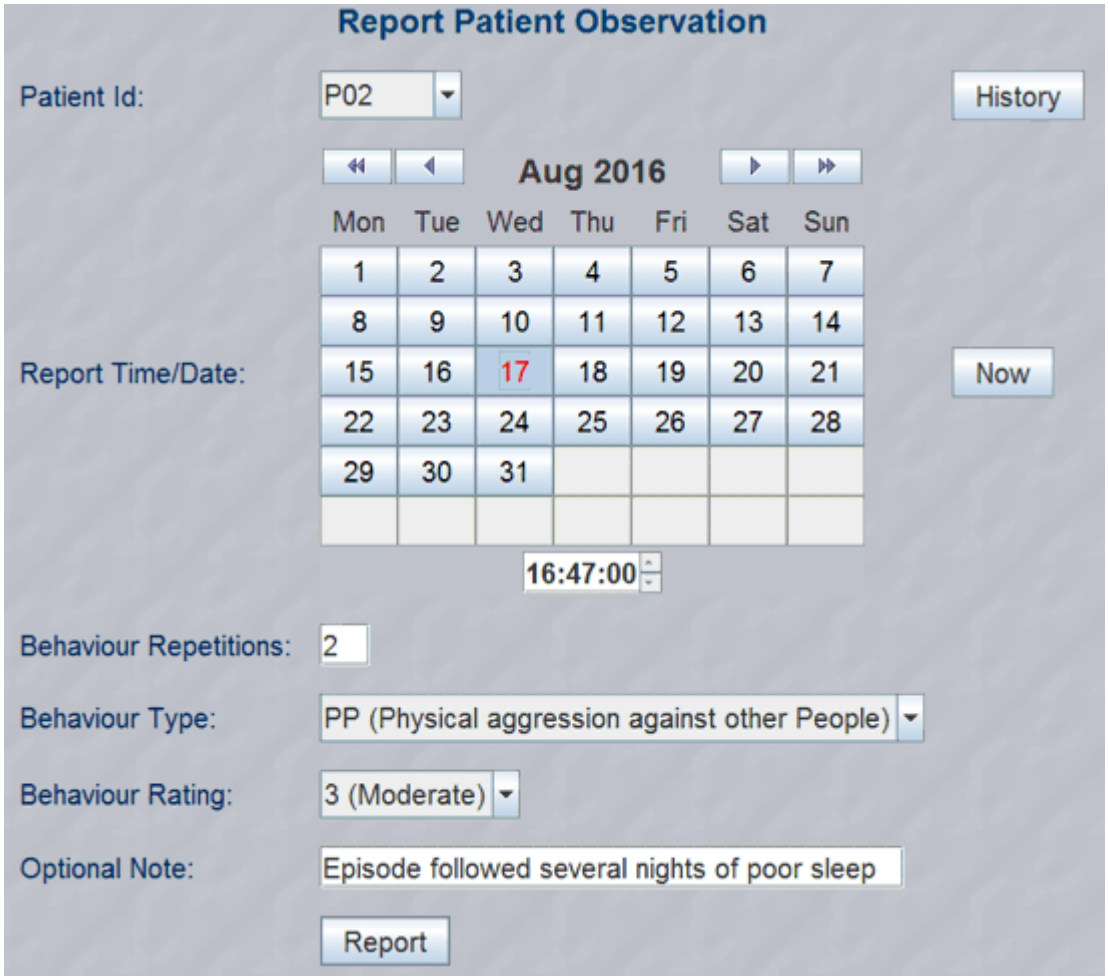


Figure 2. Example of Reporting Aggression

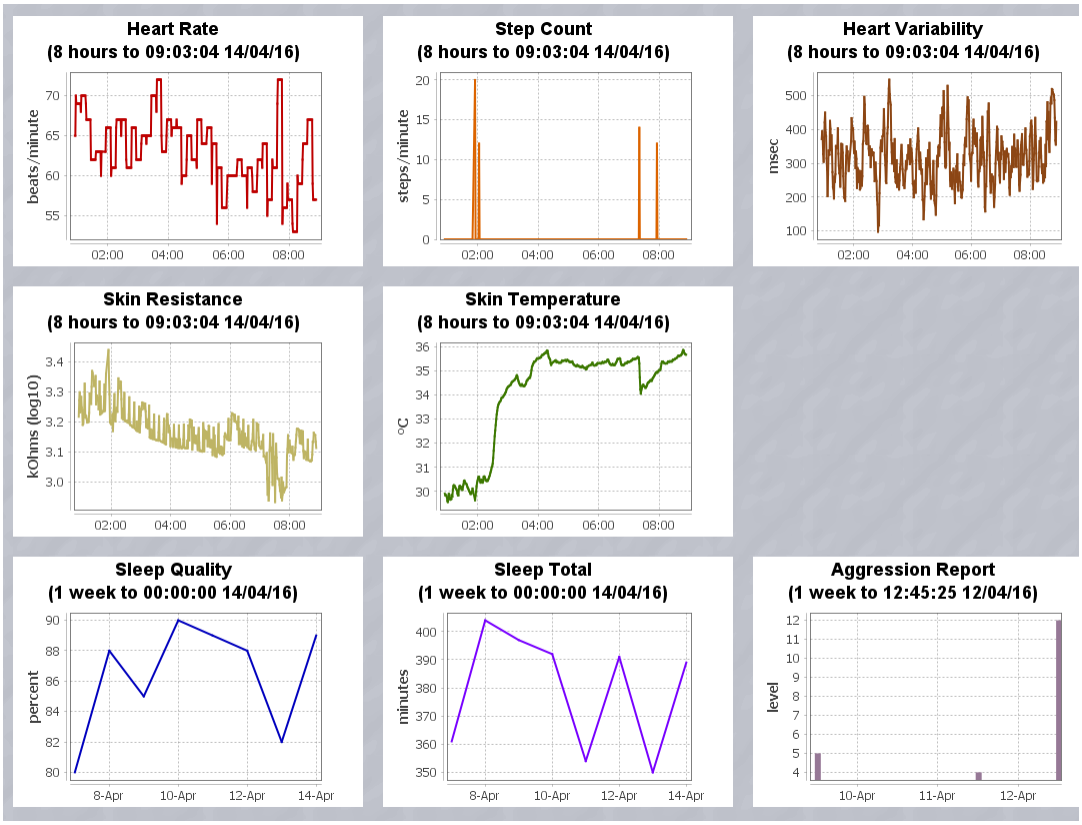


Figure 3. Sample Patient Charts

The prediction algorithm estimates the likelihood of an emotional episode. In the desktop application this is displayed as a coloured disc for each patient, accompanied by the percentage risk. A wide range of colours is used from blue (low risk) to green (medium risk) to red (high risk). Although not the 'traffic lights' that might have been expected, the full spectrum is used for two reasons: to exploit a wider range of colour, and to deal with a common form of colour blindness that causes green and red to be indistinguishable. With green as low risk and red as high risk, it could be difficult for some people to notice a significant episode risk from colour alone. With the full spectrum, blue as low risk should always be distinguishable from higher risk values (green or red).

Feedback for out-patients would be particularly important as they are not directly cared for by medical staff. Patients can be alerted in two ways to the risk of an emotional episode. The TEAMED mobile phone application shows a coloured disc following the same scheme as the desktop application. This is accompanied by brief textual advice, for example to undertake a relaxing activity if the risk is assessed as high. Since the user may not always be watching the phone, a home user can also receive biofeedback using a coloured light that follows the same colour scheme. Several coloured LED lights are available commercially; currently TEAMED is using the EasyBulb (http://www.easybulb.com), also marketed under other brand names. One or more lights can be positioned discreetly in the home to provide the user with feedback as to their emotional state. Although primarily intended for the home user, the lights can also be used in a clinical setting (e.g. in an individual's room or in a staff room).

Other forms of feedback could be used, e.g. a discrete chime or vibration if the risk of an episode risk is predicted to be high. Since the user needs to be in range of a mobile phone for communication with the watch, the phone would be an obvious feedback device. However the use of colours allows a range of risk values to be shown, rather than the binary possibilities of chimes or vibration. It also avoids disturbing the users (e.g. while sleeping).

2.4 Software Components

Apart from proprietary software on the phone and on the manufacturers' servers, the main software components developed by TEAMED are a desktop application and a mobile phone application.

The TEAMED ADMIN desktop application is used by medical staff in the clinic responsible for the patients. The application is written in Java and so is portable across platforms. As will be seen from Figure 1, the application has a number of distinct modules. It includes functions for managing patient data: creating, amending and deleting patient records, as well as starting and stopping monitoring of individual patients. Since TEAMED is at the experimental stage, patient data is currently stored separately from other medical records, but integration of these is an obvious future extension. The desktop application also supports more specialised functions:

- data collection and storage from smartwatches (or other devices)
- entry of emotional episodes by medical staff (or out-patients in future)
- data analysis and prediction of emotional episodes
- providing charts of physiological data to medical staff
- informing medical staff and patients about the risk of an emotional episode.

The TEAMED PATIENT mobile phone application runs on the phone associated with the watch. (This is separate from the manufacturer-supplied application to interface the watch to the phone.) So far only an Android implementation has been created as such phones are inexpensive for widespread deployment. However, the application could also be readily ported to, say, an Apple or Microsoft smartphone. Besides a log of key events, the mobile application displays the likelihood of an emotional episode as explained in Section 2.3.

For a smartwatch that requires data to be collected on the phone (the Microsoft Band here), the TEAMED PATIENT application also receives data from the watch and sends it to the project database after preprocessing. To reduce the effect of communication on watch battery life, sensor data is sampled 4 times per minute, averaged per minute, and sent to the database every 10 minutes. For the Microsoft Band this reduces the watch battery life from 2 days (when the watch is idle, i.e. not being monitored) to 1.5 days (when watch data is actively being collected). Frequent sampling requires more frequent transmissions by the Band. This can reduce the watch battery life to under a day, which would then be operationally inconvenient.

3 Analysis Strategy

The reasons for predicting episodes using machine learning are discussed. The types of features that can be defined are described. The features are then used to analyse physiological data, building classifiers that predict episodes based on physical or sleep data.

3.1 Prediction Technique

Having collected physiological data, it is then necessary to analyse this in order to predict episodes of emotional dysregulation. In fact there are multiple time series (6 or 7 physiological data streams) that need to be correlated against the report time series. It was not definitively known *a priori* which kinds of data would be most effective in predicting episodes, nor what features of the data would be most relevant. The quality of the data was also not known at the start of the work.

Broadly speaking, episode prediction using the physiological data could employ two kinds of technique: purely statistical approaches or machine learning. There are certainly statistical techniques for time series analysis. Approaches like ARIMA (Autoregressive Integrated Moving Average [10]) tend to concentrate on forecasting, i.e. predicting future values of a time series based on past data. Non-linear regression analysis was also considered as a way of relating episodes to other factors. However, the authors suspected that it would be too limiting to focus on a single technique ahead of actually collecting and analysing data.

It was therefore decided to use a machine learning package because of its flexibility. For this the authors were attracted to the data mining tool called WEKA (Waikato Environment for Knowledge Analysis, *http://www.cs.waikato.ac.nz/ml/weka*). This offers a wide range of classification algorithms and easy manual experimentation with different classification approaches. Several statistical techniques are supported by WEKA, including regression analysis. Being Java-based, WEKA capabilities can also readily be integrated with the TEAMED desktop application (which is also written in Java).

3.2 Classification Features

The physiological data is quite voluminous: multiple series, with physical data recorded every minute: up to 7200 measurements per day. The sleep data is less bulky, with only daily values: 2 measurements per day. For classification purposes a number of features are extracted from the data. These are computed per hour for physical data, though still making use of the minute-by-minute measurements. For sleep data the features are computed per day.

A common issue in processing physiological data is that it needs to be normalised across individuals or even for the same individual across time [24]. On TEAMED this has been achieved by defining features that avoid the use of absolute values for data. This is because absolute values vary between patients, and even for the same patient on different days and at different points of the day. As an example, the resting heart rate for one patient might be 60 beats/minute and for another 75 beats/minute. If heart rate is measured at 80 beats/minute, this could be significant for the first patient but not for the second. Skin resistance for a patient might be 500 kOhms today compared to 1000 kOhms yesterday due to more humid weather or the patient being better hydrated. How sensor readings vary is thus more important than their absolute values.

It was not known in advance how long a period of data would be relevant in predicting episodes. The general approach illustrated in Figure 4 was therefore adopted. A reference point in time is defined for computing a feature. Nearly all features have a sample period in the past defined relative to this: data in the sample period is used to calculate the feature. Report features normally have a sample period in the future when it is ascertained if an episode occurred. However, it can also be useful to check for episodes in the past to see if they correlate with future episodes.

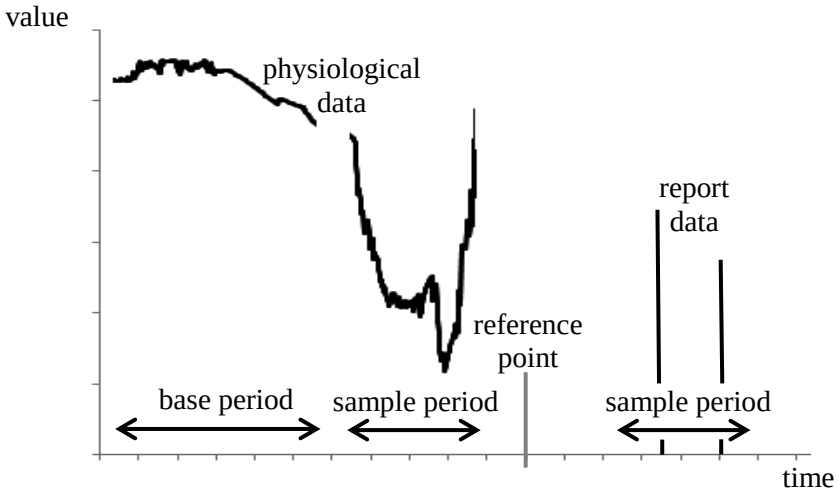


Figure 4. Feature Periods

Besides the sample period, a few features (notably for calculating standard deviation) also have a base period: values in the sample period are compared to this. The reason for a base period is so that values in the sample period are not compared with themselves. For example, it might be relevant that heart rate in the past hour has become more irregular compared to the previous hour.

It was also not known in advance what features of the data would be useful in predicting episodes. A range of plausible features was therefore defined initially. Once actual data had been collected, this was extended to add peak and trough detection since these features looked as though they would be significant. Table 1 summarises the types of features that can be used. For those that are not immediately obvious, an explanation is given below. The features used in practice are temporal abstractions in that they identify patterns that are not correlated with actual time. That is, they focus on the pattern of measurements during a relatively lengthy time periods (hourly or daily) where temporal correlation can be expected to be limited.

Most features apply to physical or sleep data, but `REPORT` features are used to analyse only report data. For the OAS-MNR aggression scale mentioned in Section 1.3, aggression reports have four categories and four levels. For ease in analysis, these reports are converted to a 16-point numerical scale measured by their `LEVEL`. As an example, PS 3 (moderate Physical Aggression against Self) is converted to a value of 11 (8 for PS plus 3 for moderate). Some classifiers cannot deal with a numeric predicted class, so the alternative `DEGREE` feature is provided to compute a discrete ('nominal') class: PS 3 corresponds to 'medium'.

Calculating the slope of a graph through line fitting is a computationally expensive operation. The `TREND` feature in Table 1 is therefore defined as an efficient way of estimating whether a graph is generally rising or falling, similar to the idea of local angles [21]. Suppose that heart rate in some period is mostly rising, but has some downward dips. `TREND MAX` would compute the largest number of consecutive rising steps in the period as an indication of whether there is a general trend upwards.

`VALUE PEAKS` and `TROUGHS` are detected using an adaptation of the algorithm by Billauer (http://www.billauer.co.il/peakdet.html). To allow for noisy data, points in peaks or troughs must be distinguished from their neighbours by a certain threshold.

Type	Variant	Meaning
REPORT	DEGREE	severity of an episode report on a discrete/continuous scale
	LEVEL	
STDEV	MAX	number of base standard deviations that the maximum/mean/minimum sample value is away from the base mean
	MEAN	
	MIN	
TREND	MAX	largest number of consecutive rising/falling steps in the sample period
	MIN	
VALUE	COUNT	number of data points
	EXISTS	whether any data points exist
	MAX	maximum/minimum value in the sample period (normally not used as the values are absolute)
	MIN	
	PEAKS	number of sustained peaks/troughs in the sample period
	TROUGHS	
	STDEV	standard deviation in the sample period

Table 1. Feature Types

3.3 Analysing Data

To keep the approach flexible, the selection of features for analysis is not hard-coded. Instead the data analyser is parameterised by feature definitions that specify the feature characteristics, base period and (if used) sample period. The choice of features was determined in consultation with the clinical partner, based on expectations of physiological influences and also a prior study of how physiological measurements seemed to correlate with aggressive episodes.

Since it was not known *a priori* which features would be the best indicators of episode risk, a wide range of plausible features was therefore included. As an example of a clinically *implausible* feature, it would not be likely that peaks in skin temperature would correlate with an episode a day later.

A nice characteristic of machine learning is that building a classifier can discover which features are good indicators of an episode. There is only a small performance penalty in defining more features than actually prove to be useful.

Table 2 shows the physical features (periods in hours) used in the evaluation, while Table 3 shows the sleep features (periods in days). As noted in Section 3.2, it is undesirable to use absolute data values so some of the feature types in Table 1 were not used (e.g. `VALUE MAX/MIN`) even though they were, in principle, available.

Each feature set must end with `REPORT DEGREE` or `REPORT VALUE EXISTS` since this will be the class to be predicted. The measurements are, of course, those that can be collected from the watches. Feature types and variants are described in Table 1, while base and sample periods are described in Figure 4. As experience with the approach and medical insight mature, it will be easy to change the features that are calculated. The following examples illustrate how features can be defined.

HEART RATE	STDEV	MAX	-1 … 0	-2 … -1

This determines how atypical the recent maximum heart rate is compared to the previous values. It first computes the mean and standard deviation of heart rate during the base period two hours ago. The maximum heart rate is then determined during the sample period during the last hour. The value of the feature is how many base standard deviations the maximum is away from the base mean during the sample period.

SKIN TEMPERATURE	VALUE	TROUGHS	-2.5 … -1.5

This determines whether skin temperature has regularly dipped. It counts how many sustained troughs there were in skin temperature during the period from 2.5 hours ago to 1.5 hours ago.

REPORT	VALUE	EXISTS	3 … 4

This determines whether an episode report was made relatively soon after the reference point, specifically whether an episode (of any type or level) was reported during the fourth hour from now.

REPORT	REPORT	DEGREE	1 … 2

This determines the degree of episodes that were reported shortly after the reference point. It computes the report degree (typically none, low, medium or high) during the next but one hour.

SLEEP QUALITY	TREND	MIN	-7 … 0

This determines whether sleep quality has been tending to deteriorate recently. It counts the largest number of consecutive falling steps in sleep quality during the past seven days.

The data analyser is given a period of past data to analyse: 1, 2 or 4 weeks. It extracts physical/sleep features such as those in Table 2 and Table 3 for each hour/day in the data. WEKA ARFF files (Attribute-Relation File Format) are created for physical and sleep data (one pair of files per patient).

The feature files are then supplied as training data to classifiers for physical and sleep data. These build and store models that are able to predict episodes based on the measurement of features at some future time. Suppose that a high variation in maximum heart rate and a large number of troughs in skin resistance are found to be good predictors of an episode. If such values of these features are later measured, this might predict that an episode was likely. Predictions have a probability (e.g. 0.85) that is reported to the user via a coloured indicator (e.g. orange) as discussed in Section 2.3. For medical staff the probability is also given as a percentage risk, while for patients the probability is rendered as appropriate advice about relaxing.

Measurement	Type	Variant	Sample Period	Base Period
HEART RATE	STDEV	MAX	-1 … 0	-2 … -1
HEART RATE	STDEV	MAX	-2 … -1	-3 … -2
HEART RATE	VALUE	PEAKS	-1 … 0	
HEART RATE	VALUE	PEAKS	-2 … -1	
HEART VARIABILITY	STDEV	MAX	-1 … 0	-2 … -1
HEART VARIABILITY	STDEV	MAX	-2 … -1	-3 … -2
HEART VARIABILITY	VALUE	PEAKS	-1 … 0	
HEART VARIABILITY	VALUE	PEAKS	-2 … -1	
SKIN RESISTANCE	STDEV	MIN	-1 … 0	-2 … -1
SKIN RESISTANCE	STDEV	MIN	-2 … -1	-3 … -2
SKIN RESISTANCE	VALUE	TROUGHS	-1 … 0	
SKIN RESISTANCE	VALUE	TROUGHS	-2 … -1	
SKIN TEMPERATURE	STDEV	MIN	-1 … 0	-2 … -1
SKIN TEMPERATURE	STDEV	MIN	-2 … -1	-3 … -2
SKIN TEMPERATURE	VALUE	TROUGHS	-1 … 0	
SKIN TEMPERATURE	VALUE	TROUGHS	-2 … 0	
STEP COUNT	STDEV	MAX	-1 … 0	-2 … -1
STEP COUNT	STDEV	MAX	-2 … -1	-3 … -2
STEP COUNT	VALUE	PEAKS	-1 … 0	
STEP COUNT	VALUE	PEAKS	-2 … -1	
REPORT	VALUE	EXISTS	0 … 1	

Table 2. Physical Features used during Evaluation Phase

Measurement	Type	Variant	Sample Period	Base Period
SLEEP QUALITY	STDEV	MIN	-1 … 0	-6 … -1
SLEEP QUALITY	STDEV	MIN	-2 … -1	-7 … -2
SLEEP QUALITY	VALUE	TROUGHS	-7 … 0	
SLEEP QUALITY	TREND	MIN	-7 … 0	
SLEEP TOTAL	STDEV	MIN	-1 … 0	-6 … -1
SLEEP TOTAL	STDEV	MIN	-2 … -1	-7 … -2
SLEEP TOTAL	TREND	MIN	-7 … 0	
SLEEP TOTAL	VALUE	TROUGHS	-7 … 0	
REPORT	VALUE	COUNT	-1 … 0	
REPORT	VALUE	COUNT	-2 … -1	
REPORT	VALUE	EXISTS	0 … 1	

Table 3. Sleep Features used during Evaluation Phase

4 Evaluation

The way the approach has been evaluated in practice so far is presented. The approach to predicting episodes, its theoretical and practical accuracy are then discussed.

4.1 Conduct of The Evaluation

A small-scale evaluation was carried out in order to gain confidence in the approach and to establish whether emotional episodes could in practice be predicted. The evaluation was conducted with patients at BIRT (Brain Injury Rehabilitation Trust) in Glasgow. This is a 29-bed, specialist, multidisciplinary hospital supporting neurobehavioural rehabilitation of in-patients. For the evaluation, all interactions with patients were through therapy staff at the hospital; technical staff from the project did not meet or work with the patients.

Ethical approval for the evaluation was granted by the Research Ethics Committee of the Disabilities Trust UK for monitoring and data collection; approval for active intervention is planned for a later evaluation. Initially, staff at BIRT were briefed as to why the evaluation was being carried out, what it would involve, and how to use the hardware and software.

A request was then made of patients at BIRT to take part in the evaluation. These patients have a variety of acquired brain injuries, but only some exhibited frequent enough aggressive behaviour for classifier training to be feasible. Of these, some patients were excluded as they were insufficiently well to give informed consent to the evaluation. Yet others were not considered reliable enough to use a watch and phone without breaking them.

This reduced potential participants to 3 patients, all of whom agreed to take part in the evaluation. The results that follow are aggregate figures across all participants. The patients were in receipt of psychoactive medication that remained constant throughout the evaluation. Their medical background was as follows:

- Patient 01 was a 41-year old male with a 22-year history of severe traumatic brain injury due to an assault with blunt weapons and alcohol-induced psychotic disorder. The patient presented with hyperacusis – a heightened sensitivity to noise that preceded many behavioural episodes. He displayed periods of dysregulated emotional responses, resulting in highly escalated periods of agitation and concomitant assaults on other patients and staff.
- Patient 02 was a 39-year old male with a 19-year history of organic personality disorder, diabetes mellitus hypoglycaemic coma, and tonic-clonic (grand mal) seizures. Diurnal variations in attentiveness level were apparent and he experienced daytime sleepiness. He has a very low memory index, impaired executive function, and severe cognitive impairment across domains.
- Patient 03 was a 26-year old male with a 15-year history of localisation-related epilepsy with simple partial seizures, mild cognitive disorder, and acute disseminated encephalitis. He presented with intermittent periods of severely challenging behaviour since his acquired brain injury, including self-injurious behaviour and physical aggression to others. He has impaired attention level, verbal memory and visual memory.

The participants wore a Basis Peak or Microsoft Band watch according to their preference. They carried a mobile phone in a belt pouch. Data was collected over a period of 24 weeks. The total amount of data collected was 178 potential days of physical data (211,844 records) and 163 potential days of sleep data (105 records). In addition to this, medical staff entered a total of 121 reports of aggressive episodes (423 behaviour records).

There were fairly frequent gaps in the data such that only 71% of potential physical data and 64% of potential sleep data was collected. For physical data the gaps ranged from tens of minutes to a few hours to a few days. Although this posed a challenge for the analysis it is probably typical of what can realistically be achieved in a clinical setting, where patient care must take priority over data collection. The desktop application used by the medical staff will warn if no data is received for an hour. However, the staff cannot always respond quickly to this. There were a number of reasons for gaps in data collection:

- The watch or phone would sometimes run out of charge. Although the hospital staff planned a regular recharging schedule, patient priorities sometimes meant that this could not be done in time (especially over the weekend).
- The watch might lose synchronisation with the phone due to their becoming separated, e.g. because the phone was left by the bedside. This was mainly a problem with the Basis Peak and often required a manual resynchronisation.
- The patient might remove the watch for a time, e.g. to shower or because it was uncomfortable. The reason that less sleep data was collected is that patients sometimes preferred to take the watch off at night.
- Wearing a watch loosely sometimes meant that out-of-range values were received so that the data could not be used.

When the approach is used in future with out-patients, data might be lost for similar reasons but also if the Internet connection to the phone is interrupted (e.g. due to moving into an area with weak cellular reception).

In general, overfitting can be a problem when training classifiers. As will be seen in Section 4.2, once initial data was collected some experiments were conducted with cross-validation to evaluate whether the approach had promise. At that stage there was a risk of overfitting leading to deceptive results, so limited credence was attached to these outcomes. However, this was only a stepping stone to the work in Section 4.3 that validated actual performance on new data against classifiers based on training data.

Data about each patient is collected regularly. Retraining is at the discretion of the clinician, but is recommended at intervals of a week or two – particularly if there are reasons to believe that the patient's condition is improving or worsening. As a result, the classifiers are always being updated with

fresh data. When training is triggered, the results of cross-validation at that point are reported. This is only a very broad indication of how well the classifiers can perform. The system (and the medical staff) subsequently monitor how accurate the predictions are. If later performance is disappointing, this is a good reason to initiate training again.

The classification process intentionally uses relatively simple and well-known classifiers. Classifier parameters *per se* are not tweaked, though misclassification cost can be optimised during training. However, with regular refreshes of training data and monitoring performance it is felt that there is sufficient safeguard against the risk overfitting.

4.2 Classification with Training Data

Once 4 weeks of data had been collected for each patient, this was used to train classifiers based on physical and sleep data. Classifiers vary widely in their characteristics such as whether they accept discrete or continuous values, how well they cope with large feature sets, or whether they can be coupled with other techniques such as feature selection or cost-sensitive classification.

It was not known in advance which kind of classifier would be most appropriate, so around a dozen were evaluated through cross-validation on training data. The classifiers that were tried included well-known algorithms such as C4.5, Decision Tables, Decision Trees, Naïve Bayes, Multilayer Perceptrons and Support Vector Machines. Their performance was informally compared on the basis of 'accuracy' as discussed below. It is perhaps worth saying that the goal of the work reported here was to assess the overall approach rather than to come up with optimal classifiers and features: that would be the subject of future research once further data has been collected.

A single type of classifier is used for all patients, though of course it is trained on individual patient data. In fact the classifier type is a parameter to the data analyser so it is easy change (as happened during early experiments).

At this point in the work, only training data was available since there were insufficient episode reports to make separate test data feasible. The performance of the generated classifiers therefore had to be evaluated at this stage using cross-validation.

A problem with the classification was immediately apparent. Only a few percent of the hourly records (typically 2%) are associated with an episode, the vast majority being associated with normal behaviour. For a typical analysis period of 4 weeks, 12 hours with episodes were recorded on average for 489 hours of physical data. Across all patients, this ranged from 9 episodic hours in 464 hours to 14 episodic hours in 514 hours. (Recall that one episode is often associated with multiple outbursts, giving rise to multiple reports in a short period.)

The issue of data imbalance is well-known in machine learning. The situation arises when one class (normal behaviour here) heavily outnumbers other classes (episodic behaviour here). If an attempt is made to train with the raw data, the result will be a classifier that strongly prefers to identify only the majority class. The accuracy in identifying minority classes is then likely to be very poor.

There are a number of techniques for dealing with the situation of data imbalance [8]. Two common approaches were tried: oversampling the minority class(es) and/or undersampling the majority class, and using a cost matrix to bias the classifier away from incorrect decisions.

For the first approach, WEKA supports filtering using SMOTE (Synthetic Minority Oversampling Technique [12]). It was, however, necessary to modify this slightly. WEKA has just one kind of numeric attribute, but a number of feature types have integer rather than real values. If SMOTE is used to oversample such data, it creates synthetic instances with real values in nearly all cases. As a result the classifier tries to distinguish cases that cannot arise in practice (e.g. a peak count in the range 1.2 to 1.8, when only the values 1 and 2 are possible). The effect is that the classifier does not perform as well as it should. The SMOTE filter was therefore modified to synthesise only integer values where the raw instances had integers (real values continuing to be synthesised in the case of reals).

With this change, a variety of classifiers was tried. When cross-validation was carried out (using 10 folds and leave-one-out), the most accurate classifier proved to be a Decision Table. Table 4 illustrates typical accuracy when predicting episodes and normal behaviour based on physical and sleep data. The period for training data is shown as well as how far in advance episodes were predicted.

Data Type	Episode Behaviour	Normal Behaviour	Training Period	Prediction Period
Physical	91%	39%	2 weeks	3–4 hours
Sleep	68%	46%	4 weeks	2–3 days

Table 4. Prediction Accuracy for Cross-Validation of Physical and Sleep Data

'Accuracy' in this article has its everyday meaning of what proportion of some behaviour is correctly identified. In machine learning terms, this corresponds to 'true positive rate' or 'sensitivity' when used of episodes, or 'true negative rate' or 'specificity' when used of normal behaviour. 'Accuracy' was chosen as it is immediately meaningful to the medical staff using the system. It was important to be able to monitor and communicate prediction performance to people such as medical staff who are not knowledgeable about machine learning. For them, the only visible aspect of performance is how well predictions pan out. However it should be noted that in machine learning 'accuracy' means the proportion of correct identifications (true positives plus true negatives) compared to the total number of cases.

Various theoretical measures of classifier performance are used by researchers such as F-Measure, Kappa Statistic, MCC (Matthews Correlation Coefficient), ROC (Receiver Operating Characteristic) and Youden's Index. The development phase experimented with all of these when assessing classifiers. However, it was felt that they were too technical for easy interpretation by medical staff. Instead, only 'accuracy' as above was employed during the validation phase. The small scale of the evaluation means that more extended measures of performance would not be very meaningful.

Every time an episode prediction is made (i.e. every hour of new physiological data), the prediction type and episode probability are logged in the database. Accuracy can then be assessed by comparing predictions to what actually happened (i.e. whether an episode was subsequently recorded by medical staff). If the predicted episode risk exceeds 70% and an episode is recorded during the prediction period, this is treated as an accurate prediction; otherwise the prediction is treated as inaccurate.

The accuracy shown in Table 4 refers to whether any kind of episode was predicted. In fact the OAS-MNR aggression scale mentioned in Section 1.3 records aggression at multiple levels of severity. It was therefore investigated whether better accuracy could be achieved by trying to predicting multiple degrees of episode. (SMOTE works only with nominal, i.e. discrete, rather than numerical classes.) Potentially this could allow for a finer-grained classification and hence greater accuracy.

In place of the 2-degree classification in Table 4 (episode vs. normal behaviour), 3 or 4 report degrees were investigated (ranging from none to high). This sometimes showed better accuracy, but often resulted in too few instances of the episode class. Since the number of episodes is relatively small, subdividing them into different degrees meant that there were often too few for sensible classification. Later work therefore considered only a binary (2-degree) classification: episode or normal behaviour.

4.3 Classification with Actual Data

After a further 4 weeks of data collection, it was then possible to assess how well the classifiers performed against actual data. It had been wondered whether performance when cross-validating training data would be typical of performance with oversampling and actual data. However, this proved not to be the case, probably because SMOTE creates artificial data that may not be typical of real data. When actual performance was compared to cross-validation performance, it was found that the classifiers had low accuracy in practice when identifying episodes (27%) though they were reasonably accurate at identifying normal behaviour (85%).

In practice there is a penalty associated with misclassification. If an episode is predicted but normal behaviour ensues, the medical staff will have responded to a false alarm. This is not actually serious as it just means the staff keep a closer eye on the patient in the following period. However if normal behaviour is predicted but an episode occurs, the medical staff will have to deal with this as an unanticipated and possibly disruptive event. In fact, the goal of TEAMED was to avoid this situation. The opinion of medical staff at the partner hospital is that they would rather deal with more false alarms about episodes than failures to predict an episode.

A second technique (a cost matrix) for dealing with class imbalance was therefore tried since the first technique (oversampling) had proved disappointing on actual data. When a range of classifiers was tried with a cost matrix, a Partial Decision Tree [17] proved to be most accurate.

Table 5 shows the classifier cost matrix for predicting normal or episodic behaviour; the confusion matrix has the same form. As usual, the cost of correct classifications is set to 0 since the prediction is accurate. In principle, separate costs could be set for incorrect classifications. Instead a nominal cost of 1 is used for incorrectly predicting an episode (a false alarm about an episode), and some higher cost C is set for incorrectly predicting normal behaviour (a missed episode). As noted above, false alarms are not serious and have little impact on medical staff. However, failing to predict an episode should be penalised more heavily.

		Predicted	
		Normal	**Episode**
Actual	**Normal**	0	1
	Episode	C	0

Table 5. Classifier Cost Matrix

The most appropriate choice for cost *C* is determined automatically. It was found in practice that the best value for this cost depended on whether physical or sleep data was being analysed, and also varied across patients. When a new batch of training data is analysed, cross-validation is performed with a range of typical costs. This allows the best cost to be determined as the one that leads to the highest accuracy. Although a heuristic search method could have been used, this was not felt to be worthwhile since only cross-validation can be performed at this point. It would also not be sensible to over-train the classifier on purely training data (as opposed to the actual data that follows training). If necessary the computed cost can be fine-tuned by an expert through a definition in the program configuration file. The misclassification cost is subsequently used when classifying new data.

With an optimal choice of cost matrix for each patient, Table 6 shows prediction performance averaged across all participants when using physical data. This gives the prediction accuracy for episodes and normal behaviour. The variation is shown for different training and actual periods: 1, 2 or 4 weeks for training and then actual performance in the subsequent 1, 2 or 4 weeks. The variation is also shown with different prediction periods: 1, 2, 3 or 4 hours ahead.

	Training/Actual (weeks)					
	1		**2**		**4**	
Prediction (hours)	**Episode**	**Normal**	**Episode**	**Normal**	**Episode**	**Normal**
0–1	75%	16%	**81%**	**42%**	35%	64%
1–2	75%	16%	91%	9%	41%	53%
2–3	100%	17%	64%	45%	60%	45%
3–4	75%	16%	41%	59%	**68%**	**42%**

Table 6. Prediction Accuracy with Cost Matrix and Full Validation of Physical Data

Table 6 shows that useful results were achieved with 4 weeks of training data and predictions 3 to 4 hours ahead. An average of 68% of episodes and 42% of normal behaviours are correctly predicted; the range across different patients was 63% to 71% of episodes and 32% to 54% of normal behaviour. Making predictions up to 4 hours ahead gives medical staff adequate time to monitor a patient at risk and to intervene as appropriate. It suggests that the mechanisms that cause aggressive behaviour have a measurable physiological effect a few hours ahead of the actual aggression. Although performance is actually better with 2 weeks of training data and predictions up to 1 hour ahead, this would not give much warning of a potential problem and so was not selected.

With an optimal choice of cost matrix for each patient, Table 7 shows prediction performance when using sleep data; unlike Table 6 the prediction periods here are in days. The best performance was for 2 weeks of training data and predictions 3 to 4 days ahead. An average of 82% of episodes and 68% of normal behaviours are correctly predicted; the range across different patients was 50% to 100% of episodes and 54% to 100% of normal behaviour. It is perhaps surprising that reasonably accurate predictions can be made so far ahead. The reason is partly that several nights of poor sleep can trigger an episode, and partly because episodes a few days ago are a reasonable predictor of future episodes.

	Training/Actual (weeks)					
	1		**2**		**4**	
Prediction (days)	**Episode**	**Normal**	**Episode**	**Normal**	**Episode**	**Normal**
0–1	25%	84%	82%	42%	47%	59%
1–2	60%	9%	82%	63%	90%	21%
2–3	25%	75%	80%	60%	72%	38%
3–4	60%	91%	**82%**	**68%**	69%	49%

Table 7. Prediction Accuracy with Cost Matrix and Full Validation of Sleep Data

Comparing Table 4 (cross-validation) with Table 6 and Table 7 (actual performance), the predictions are not quite as good in practice as the theoretical performance might suggest. However, this is to be expected as cross-validation is not a true assessment. Since the risk of an episode is reported as the greater of the physically-based and sleep-based probabilities, this is more accurate than using either form of data alone. Actual accuracy is therefore better than suggested by the results in Table 6 or Table 7 alone.

After reviewing prediction results from the evaluation, staff of the clinical partner (BIRT) judged that the accuracy was acceptable and that the approach would be useful. Indeed it is a big improvement on the current situation where episodes are unpredictable and cannot easily be anticipated by the staff. Furthermore, the hospital was pleased that they would receive adequate warning of possible aggressive episodes. The hospital reports that when their patients become aroused and aggressive they respond well to being 'talked down' and encouraged to relax. There is therefore a real opportunity for intervention by medical staff to head off an episode.

When the approach is in future used by out-patients (not under the immediate supervision of medical staff), it is hoped that the feedback to patients about the risk of an episode will similarly allow them to self-manage their arousal.

5 Conclusion

5.1 Summary

It has been seen that mental health issues have a significant impact on society. Emotional dysregulation, common to a variety of mental health conditions, greatly affects someone's ability to participate in normal daily life. The general aim of the work reported here has been to predict several forms of emotional dysregulation, though the initial evaluation has been of predicting aggressive outbursts.

The system for collecting, storing and analysing physiological data has been explained. The analysis and prediction strategy has been discussed. The performance of the approach has been seen from the results of a small-scale evaluation.

In terms of the objectives set out in Section 1.2:

- It has been possible to interface commercial, off-the-shelf smartwatches for collection of physiological data of relevance to emotional dysregulation. A small-scale evaluation collected 178 days of physical data and 137 days of sleep data, amounting to nearly 212,000 records.
- Machine learning techniques have proven useful as a means of analysing this kind of data. A partial decision tree with a cost matrix for dealing with data imbalance proved to be the best classifier.
- It has been shown that emotional outbursts can be successfully predicted based on physiological measurements. The evaluation achieved a typical accuracy of up to 82% in predicting episodes and up to 68% in predicting normal behaviour.
- It has been possible to give timely feedback to medical staff and patients about the risk of an emotional outburst, typically 3 to 4 hours ahead based on physical data and 3 to 4 days ahead based on sleep data.
- A preliminary assessment of the approach has been carried out with a limited number of patients.

The clinical partner in the project has expressed satisfaction with the results, and is planning on using the approach in future with more patients.

5.2 Future Work

A number of future directions were identified during the work reported here. An extended evaluation has been already planned for the near future. This will involve more patients across a number of BIRT hospitals in order to collect substantial evidence of efficacy. Initially the approach has been used on an advisory basis. BIRT are now intending to base actual interventions on the predictions made by the system.

So far the approach has been applied to predicting aggressive outbursts among brain-injured patients in a clinical setting. The extension to out-patients is an obvious next step. Training the classifiers will then require a degree of patient self-assessment, e.g. using questionnaires on a mobile phone.

A new project has been started with a Health Board in Scotland, using physiological data from smartwatches to support patients with dementia who are prone to distress. A number of other useful

applications of the work have also been identified. The possibilities include helping people who suffer from anxiety attacks, depression (which is often accompanied by anxiety) or bipolar disorder. Another development envisaged is with prisoners who are clinically fit but tend to have violent behaviour.

Some technical developments would also be desirable. The smartwatch market continues to be very volatile. For example, the Basis Peak has now been discontinued and the future of the Microsoft Band is uncertain. Fortunately the approach can readily be extended to support other kinds of smartwatch. New smartwatches are appearing all the time, e.g. the Alphabet Health Watch and the Empatica Embrace could be used with the project approach.

In the work up to this point, oversampling and cost matrixes have been used to tackle the issue of data imbalance. In future, ensembles will be investigated to see if they handle this problem any better.

So far, temporal abstractions inspired by the application domain have been used. However, the more general approach of temporal abstraction (e.g. [7]) will be an interesting new development to investigate.

At present no attempt is made to build a sequential model of emotional outbursts. This could be an interesting future development if the patient's condition evolves over time such that aggression becomes more or less likely.

It could also be useful to study other methods of predicting outbursts. The present approach is statistical in nature and so does not identify a reason for outbursts occurring. A model might be developed that identifies behavioural patterns and what leads to outbursts.

Other physiological measurements might be investigated as factors in predicting outbursts. For example physical agitation, blood pressure or voice stress indications could be relevant. However, it would be important to maintain the project emphasis on data that is easily obtained from a wearable device or a non-intrusive sensor.

The risk of an episode is currently fed back to patients and staff using a coloured indicator. The possibilities of audio and haptic feedback will be investigated in future as alternative mechanisms.

Physiological data and predictions based on this are currently collected in an isolated database. In time it would be worth considering integrating this information with other medical data. However, this would be feasible only if the approach were adopted as part of wider health service provision since there are implications for database security and compatibility with Electronic Health Records.

Acknowledgements

The TEAMED project was supported by a grant from the Digital Health and Care Institute, Scotland. The authors are grateful to Erin Dalziel, Colin Farrell and Naomi Bowers (Brain Injury Rehabilitation Trust) and to their patients for making the evaluation possible. Thanks are due to Kevin Swingler (University of Stirling) for providing key advice on machine learning, and to Roman Poppat (Data Lab) for discussion of appropriate analytic techniques.

References

[1] N. Alderman, C. Knight, and C. Morgan. Use of a modified version of the Overt Aggression Scale in the measurement and assessment of aggressive behaviours following brain injury. *Brain Injury*, 11(7):503–523, Nov. 1997.

[2] J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, Nov. 1983.

[3] Alzheimer's Disease International. The global impact of dementia: An analysis of prevalence, incidence, cost and trends. Alzheimer's Disease International, London, UK, Aug. 2015.

[4] O. AlZoubi, D. Fossati, S. D'Mello, and R. A. Calvo. Affect detection and classification from non-stationary physiological data. In M. A. Wani, G. Tecuci, M. Boicu, M. Kubat, T. M. Khoshgoftaa, and N. Seliya, editors, *Proc. 12th Int. Conf. on Machine Learning and Applications*, pages 240–245. IEEE Computer Society, Los Alamitos, California, USA, Dec. 2013.

[5] C. A. Anderson, W. E. Deuser, and K. M. DeNeve. Hostile cognition, and arousal: Tests of a general model of affective aggression. *Personality and Social Psychology Bulletin*, 21(5):434–448, May 1995.

[6] E. S. Barratt, M. S. Stanford, T. A. Kent, and A. Felthous. Neuropsychological and cognitive psychophysiological substrates of impulsive aggression. *Biological Psychiatry*, 41(10):1045–1061, May 1997.

[7] I. Batal, H. Valizadegan, G. F. Cooper, and M. Hauskrecht. A temporal pattern mining approach for classifying electronic health record data. *Intelligent Systems Technology*, 4(4):20–29, Sept. 2013.

[8] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *Knowledge Discovery and Data Mining*, 6(1):20–29, June 2004.
[9] F. A. Boiten, N. H. Frijda, and C. J. E. Wientjes. Emotions and respiratory patterns: Review and critical analysis. *Int. J. Psychophysiology*, 17(2):103–128, July 1994.
[10] G. Box and G. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, USA, 1970.
[11] G. Castellano, S. D. Villalba, and A. Camurri. Recognising human emotions from body movement and gesture dynamics. In A. C. R. Paiva, R. Prada, and R. W. Picard, editors, *Proc. 8th Int. Conf. on Affective Computing and Intelligent Interaction*, number 4738 in Lecture Notes in Computer Science, pages 71–82. Springer, Berlin, Germany, Sept. 2007.
[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Artificial Intelligence Research*, 16(1):321–357, June 2002.
[13] Counsel and Care. A charter for change. Counsel and Care, London, UK, Jan. 2008.
[14] S. Davies. The Chief Medical Officer annual report: Public mental health. Department of Health, London, UK, Sept. 2014.
[15] Department of Health Stroke Team. National stroke strategy. Department of Health, London, UK, Dec. 2007.
[16] R. El-Kaliouby and P. Robinson. The emotional hearing aid: An assistive tool for children with Asperger syndrome. *Universal Access in the Information Society*, 4(2):121—134, Dec. 2005.
[17] E. Frank and I. H. Witten. Generating accurate rule sets without global optimization. In J. W. Shavlik, editor, *Proc. 15th Int. Conf. on Machine Learning*, pages 144–151, San Francisco, USA, 1998. Morgan Kaufmann.
[18] P. E. Greenberg, A.-A. Fournier, T. Sitisky, C. T. Pike, and R. C. Kessler. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *Clinical Psychiatry*, 72(2):155–173, Feb. 2015.
[19] A. Gustavsson, M. Svensson, F. Jacobi, et al. Cost of disorders of the brain in Europe 2010. *European Neuropsychopharmacology*, 21(10):718–779, Oct. 2011.
[20] J. Healey, L. Nachman, S. Subramanian, J. Shahabdeen, and M. Morris. Out of the lab and into the fray: Towards modeling emotion in everyday life. In P. Floréen, A. Krüger, and M. Spasojevic, editors, *Proc. 8th Int. Conf. on Pervasive Computing*, number 6030 in Lecture Notes in Computer Science, pages 156–173. Springer, Berlin, Germany, May 2010.
[21] R. Henriques and A. Paiva. Learning effective models of emotions from physiological signals: The seven principles. In H. P. da Silva, A. Holzinger, S. Fairclough, and D. Majoe, editors, *Proc. 1st Int. Conf. on Physiological Computing Systems*, number 8908 in Lecture Notes in Computer Science, pages 137–155. Springer, Berlin, Germany, Jan. 2014.
[22] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wann. Physiological signals based human emotion recognition: A review. In M. N. Taib, R. Adnan, A. M. Samad, N. M. Tahir, Z. Hussain, and M. H. F. Rahiman, editors, *Proc. 7th Int. Coll. on Signal Processing and Its Applications*, pages 410–415. Institution of Electrical and Electronic Engineers Press, New York, USA, Mar. 2011.
[23] M. P. LaPlante, C. Harrington, and T. Kang. Estimating paid and unpaid hours of personal assistance services in activities of daily living provided to adults living at home. *Health Services Research*, 37(2):397–415, Apr. 2002.
[24] J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In M. J. Zaki and C. C. Aggarwal, editors, *Proc. 8th Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 2–11. ACM Press, New York, USA, Mar. 2003.
[25] R. Luengo-Fernández, J. Leal, and A. Gray. The economic burden of dementia and associated research funding in the United Kingdom. Alzheimer's Society, London, UK, Mar. 2010.
[26] R. Moskovitch and Y. Shahar. Classification of multivariate time series via temporal abstraction and time-intervals mining. *Knowledge and Information Systems*, 45(1):35–74, Oct. 2015.
[27] R. W. Picard. *Affective Computing*. MIT Press, Boston, USA, Sept. 1997.
[28] J. A. Russell. A circumplex model of affect. *Personality and Social Psychology*, 39(6):1161–1178, Dec. 1980.
[29] K. R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, Dec. 2005.
[30] Y. Shahar. A framework for knowledge-based temporal abstraction. *Artificial Intelligence*, 90(1-2):79–133, Feb. 1997.
[31] P. Shaw, A. Stringaris, J. Nigg, and E. Leibenluft. Emotion dysregulation in attention deficit hyperactivity disorder. *American Journal of Psychiatry*, 171(3):276–293, Mar. 2014.

[32] V. Shuman, K. Schlegel, and K. R. Scherer. Geneva Emotion Wheel rating study. Technical report, Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland, Aug. 2015.
[33] Y. Suchy. *Clinical Neuropsychology of Emotion*. Guildford Press, New York, USA, Mar. 2011.
[34] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss. Activity-aware mental stress detection using physiological sensors. In M. Gris and G. Yang, editors, *Proc. 2nd Conference on Mobile Computing, Applications and Services*, number 76 in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pages 211–230. Springer, Berlin, Germany, Oct. 2010.
[35] J. F. Thayer, F. Åhs, M. Fredrikson, J. J. Sollers III, and T. D. Wager. A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience and Biobehavioral Reviews*, 36(2):747—756, Feb. 2012.
[36] W. Wang, G. Sun, X. Ye, M. Shen, R. Zhu, and Y. Xu. Exteroceptive suppression of temporalis muscle activity in subjects with high and low aggression traits. *Clinical Neurophysiology*, 36(2):63–69, May 2006.
[37] S. C. Yudofsky, J. M. Silver, W. Jackson, J. Endicott, and D. Williams. The Overt Aggression Scale for the objective rating of verbal and physical aggression. *American Journal of Psychiatry*, 143(1):35–39, Jan. 1986.