Affect Detection in Daily Life Using Machine Learning and Wearable Devices

Jonne Antti Kristian Lohilahti
Master's Thesis
Cognitive Science
Faculty of Arts
University of Helsinki

May 2022

Supervisor: Ilmari Määttänen

Research Project: Sisu at Work

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Faculty of Arts	Koulutusohjelma – Utbildningsprogram – Degree Programme Cognitive science

Opintosuunta – Studieinriktning – Study Track
Cognitive science

Tekijä – Författare – Author
Jonne Antti Kristian Lohilahti

Työn nimi – Arbetets titel – Title
Affect Detection in Daily Life Using Machine Learning and Wearable Devices

Työn laji – Arbetets art – Level Master’s thesis	Aika – Datum – Month and year May 2022	Sivumäärä– Sidoantal – Number of pages 42+13

Tiivistelmä – Referat – Abstract

Objectives. This study aims to evaluate feasibility of affect detection in daily life using wearable devices and machine learning models. Affective states play an important role in decision making, perception and behaviour, making objective detection of affective states a desirable goal both for potential applications and as a way to gain insight into affective phenomena. Affective states have been found to have measurable physiological and behavioral changes, which allows training of machine learning models for detecting the underlying affects. Majority of affect detection studies have been conducted in laboratory conditions using affect elicitation stimuli or tasks, raising the question whether results from these studies will generalize to daily life. Although development of wearable devices and mobile surveys have facilitated evaluation in the context of daily life, research here remains sparse. In this study, self-reported affective states are predicted using machine learning models to identify which affective states can be detected in daily life. Additionally, model interpretation methods will be used to identify which relationships the models found important for their predictions.

Methods. Data for this thesis came from a study conducted as a part of Sisu at Work project between University of Helsinki and VTT, where 82 knowledge workers from four Finnish organizations were studied for a period of three weeks. During this period, the participants were queried by mobile surveys about their affective states thrice a day, while they also used wearable devices to record photoplethysmography (PPG), electrodermal activity (EDA) and accelerometry (ACC) signals. A signal processing pipeline was implemented to deal with movement artefacts and other issues with the data. Features describing heart rate (HR) and heart rate variation (HRV) were extraced from PPG, physiological activation from EDA and movement from ACC signals. Models were then fitted to predict the reported affective states using the extracted features. Model performance was compared against a baseline to identify which affects could be reliably detected, while permutation importance and Shapley additive explanations (SHAP) values were used to identify important relationships established by the models.

Results and conclusions. Models for affective state vigor showed improvements over baseline with statistical significance, while improvements were also noted for affects focused and enthusiastic. Permutation importance highlighted the significance of movement and HRV features, while examination of SHAP values indicated that low movement, low EDA and high HRV impacted model predictions the most. These results indicate potential for detecting high activation affective states in daily life and propose potential relationships for future research.

Avainsanat – Nyckelord – Keywords
affect detection, ecological validity, machine learning, signal processing, wearable devices

Säilytyspaikka – Förvaringställe – Where deposited
Helsinki University Library / Helda / E-thesis (opinnäytteet) ethesis.helsinki.fi

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Humanistinen tiedekunta	Koulutusohjelma – Utbildningsprogram – Degree Programme Kognitiotiede	
Opintosuunta – Studieinriktning – Study Track Kognitiotiede		
Tekijä – Författare – Author Jonne Antti Kristian Lohilahti		
Työn nimi – Arbetets titel – Title Tunteiden Havaitseminen Arkielämässä Koneoppimisen ja Puettavien Laitteiden Avulla		
Työn laji – Arbetets art – Level Pro Gradu	Aika – Datum – Month and year Toukokuu 2022	Sivumäärä– Sidoantal – Number of pages 42+13

Tiivistelmä – Referat – Abstract

Tavoitteet. Tämän tutkimuksen tavoitteena on arvioida tunteiden havaitsemisen mahdollisuutta arkielämässä puettavien laitteiden ja koneoppimismallien avulla. Tunnetiloilla on tärkeä rooli päätöksenteossa, havaitsemisessa ja käyttäytymisessä, mikä tekee objektiivisesta tunnetilojen havaitsemisesta arvokkaan tavoitteen, sekä mahdollisten sovellusten että tunnetiloja koskevan ymmärryksen syventämisen kannalta. Tunnetiloihin usein liittyy mitattavissa olevia fysiologisia ja käyttäymisen muutoksia, mikä mahdollistaa koneoppimismallien kouluttamisen muutoksia aiheuttaneen tunnetilan havaitsemiseksi. Suurin osa tunteiden havaitsemiseen liittyvästä tutkimuksesta on toteutettu laboratorio-olosuhteissa käyttämällä tunteita herättäviä ärsykkeitä tai tehtäviä, mikä herättää kysymyksen siitä että yleistyvätkö näissä olosuhteissa saadut tulokset arkielämään. Vaikka puettavien laitteiden ja kännykkäkyselyiden kehittyminen on helpottanut aiheen tutkimista arkielämässä, tutkimusta tässä ympäristössä on vielä niukasti. Tässä tutkimuksessa itseraportoituja tunnetiloja ennustetaan koneoppimismallien avulla arkielämässä havaittavissa olevien tunnetilojen selvittämiseksi. Lisäksi tutkimuksessa käytetään mallintulkintamenetelmiä mallien hyödyntämien yhteyksien tunnistamiseksi.

Metodit. Aineisto tätä tutkielmaa varten on peräisin tutkimuksesta joka suoritettiin osana Helsingin Yliopiston ja VTT:n Sisu at Work projektia, missä 82:ta tietotyöläistä neljästä suomalaisesta organisaatiosta tutkittiin kolmen viikon ajan. Osallistujilla oli jakson aikana käytettävissään mittalaitteet jotka mittasivat fotoplethysmografiaa (PPG), ihon sähkönjohtavuutta (EDA) ja kiihtyvyysanturi (ACC) signaaleita, lisäksi heille esitettiin kysymyksiä koetuista tunnetiloista kolmesti päivässä puhelinsovelluksen avulla. Signaalinkäsittelymenetelmiä sovellettiin signaaleissa esiintyvien liikeartefaktien ja muiden ongelmien korjaamiseksi. Sykettä (HR) ja sykevälinvaihtelua (HRV) kuvaavia piirteitä irroitettiin PPG signaalista, fysiologista aktivaatiota kuvaavia piirteitä EDA signaalista, sekä liikettä kuvaavia piirteitä ACC signaalista. Seuraavaksi koneoppimismalleja koulutettiin ennustamaan raportoituja tunnetiloja irroitetujen piirteiden avulla. Mallien suoriutumista vertailtiin suhteessa odotusarvoihin havaittavissa olevien tunnetilojen määrittämiseksi. Lisäksi permutaatiotärkeyttä sekä Shapley additive explanations (SHAP) arvoja hyödynnettiin malleille tärkeiden yhteyksien selvittämiseksi.

Tulokset ja johtopäätökset. Mallit tunnetiloille virkeä, keskittynyt ja innostunut paransivat suoriutumistaan yli odotusarvon, joista mallit tunnetilalle virkeä paransivat suoriutumista tilastollisesti merkitsevästi. Permutaatiotärkeys korosti liike- ja HRV-piirteiden merkitystä, kun SHAP arvojen tarkastelu nosti esiin matalan liikkeen, matalan EDA:n, sekä korkean HRV:n merkityksen mallien ennusteille. Nämä tulokset ovat lupaavia korkean aktivaation positiivisten tunnetilojen havaitsemiselle arkielämässä, sekä nostavat esiin mahdollisia yhteyksiä jatkotutkimusta varten.

Avainsanat – Nyckelord – Keywords
ekologinen validiteetti, koneoppiminen, puettavat laitteet, signaalinkäsittely, tunteiden havaitseminen

Säilytyspaikka – Förvaringställe – Where deposited
Helsingin Yliopiston Kirjasto / Helda / E-thesis (opinnäytteet) ethesis.helsinki.fi

Muita tietoja – Övriga uppgifter – Additional information

Contents

Introduction 1

Emotions, mood and core affect . . . 1

Psychophysiological signals . . . 3

Cardiovascular activity . . . 4

Electrodermal activity . . . 5

Accelerometry . . . 6

Affect detection as a machine learning problem . . . 7

Affect detection in daily life . . . 9

Methods 12

Participants . . . 12

Procedure . . . 12

Materials . . . 13

Ecological momentary assessments . . . 13

Wearable devices . . . 14

Preprocessing . . . 15

Accelerometry autocalibration procedure . . . 15

Segmentation . . . 16

Inter beat interval masking procedure . . . 16

Feature extraction . . . 18

Final dataset . . . 20

Model fitting . . . 21

Statistical testing and model interpretation . . . 22

Results 24

Model performance . . . 24

Feature importance . . . 24

Discussion **27**

Model performance 29

Model interpretation 30

Limitations and future research 31

Conclusions 33

Acronyms

ACC Accelerometry. 4, 6, 14–17, 19, 20, 24, 27, 33, 34, 43

ANS Autonomous Nervous System. 3, 4

ECG Electrocardiography. 4, 5, 10, 14, 20

EDA Electrodermal activity. 4–7, 10, 14, 15, 19, 24, 30, 32, 34, 43

EMA Ecological Momentary Assessment. 9–14, 21, 27, 30, 32, 33

EN Euclidian Norm. 15–17, 20

ENMONZ Euclidian Norm Minus One with Negatives set to Zero. 20, 21

HR Heart Rate. 4, 5, 14, 18, 19, 21, 24, 28, 31, 43

HRV Heart Rate Variation. 4, 5, 17–19, 21, 24, 25, 27–31, 33, 34, 43

IBI Inter Beat Interval. 5, 11, 14–18, 20

PNS Parasympathetic Nervous System. 4, 14, 30

PPG Photoplethysmography. 5, 7, 10, 14, 15, 19, 20, 32, 34

RF Random Forests. 21–24, 26, 44

RMSSD Root Mean Square of Successive Differences. 17, 18, 21

SCL Skin Conductance Level. 6, 14, 20, 32

SCR Skin Conductance Responses. 6, 14, 20, 25, 32

SHAP SHapley Additive exPlanations. 23–25, 27, 29–31, 34, 47

SNS Sympathetic Nervous System. 3–5, 14, 29, 32

SQE Signal Quality Estimation. 14–16, 18

VMU Vector Magnitude Units. 20, 24, 25

XGBoost eXtreme Gradient Boosting. 21–25, 27, 30, 44, 47

Introduction

Affective computing, pioneered by Picard (2000), is a multidisciplinary field of research spanning computer science, psychology and cognitive science. It aims to study and develop computation that relates to affective phenomena, promising applications in human-computer interaction, healthcare, marketing and more. A crucial subfield within affective computing is affect detection – the detection of affective states via measurement of physiology and other observables. Usually this mapping between affective states and observables is established using machine learning, while the affective states under study are induced by chosen stimuli or tasks in a laboratory setting. In this context, classification accuracies over 90% are often observed (Bota et al., 2019). It is however an open question how these results generalize to unconstrained daily life – the domain of application for many potential affective computing applications. Fortunately, developments in wearable sensor and mobile technology have facilitated the investigation of affect detection in daily life, although research here remains sparse.

To evaluate feasibility of affect detection in daily life, physiological and inertial signals from wearable devices will be used to train machine learning models for predicting self-reported affective states from mobile questionnaires. A signal processing pipeline will be presented to deal with challenges in the data and model interpretation methods will be used to examine relationships that the models found important for affect detection. Results of this study indicated that affect detection is possible for high activation positive affective states with modest accuracy and that low movement, low electrodermal activity level and high heart rate variation are predictive for these affects.

Emotions, mood and core affect

Detection of emotions, moods, affects and other affective phenomena is the central goal of the field of affect detection. Decades of proliferation of research into these concepts has however yielded a variety of definitions and conceptualizations for these terms. This has lead to imprecise usage, with affects, emotions and moods used partially or completely interchangeably while referring to substantially different constructs (Ekkekakis, 2013). Furthermore, no single measurement technique or theoretical framework has emerged to dominate others. As such, these concepts can be viewed either as discrete entities, categories or as phenomena defined by various formulations of two or more underlying dimensions.

To bring conceptual clarity to the definitions of emotions, moods and affects, the terminological system outlined by Ekkekakis (2013) is adopted. Here, core affect is defined as the most elementary consciously accessible feelings that is constantly present (Russell, 2003; Russell & Barrett, 1999). Core affect does not necessarily have to be about any specific thing or be related to any antecedent appraisal, although it can be experienced embedded with other affective phenomena like emotions or moods (Russell, 2005). Emotions are seen as comprising of several interconnected and coordinated components like core affect, cognitive appraisal, bodily changes, expressive behaviour and action tendencies (Russell & Barrett, 1999). Emotions occur rarely, are short (duration seconds to minutes) but have comparatively high intensity and are about some immediate and identifiable stimulus (Ekman, 1992; Scherer, 2005). Compared to emotions, moods are longer lasting (hours to days) and present more often but with lower intensity (Ekman, 1992). Moods include synchronized activation of the same components as emotions, but this activation is less pronounced and distinct. In general moods are described as "having a certain diffuseness" compared to emotions, with not necessarily having an immediate or clearly identifiable object of appraisal or cause (Fridja, 2009, p. 258). Based on these definitions the term *affective state* will be used as an umbrella term to refer to a state that necessarily contains core affect, but can also meet the criteria for emotions or moods, while the term affect is used to refer to affective states in general.

Affective states have been mainly modeled as distinct entities or by a set of underlying dimensions. In the distinct-states approach, each state is examined as being unique and distinct from all others (Roseman et al., 1994), in order to highlight the unique features of different emotions and moods for a deep and detailed analysis. Examples of this approach include Ekman (1992)s efforts to identify "basic" emotions or the more broad approach of describing categories of states by similarity to each other and a prototypical example (Russell, 1991). In the dimensional view, the goal is to determine the underlying dimensions that explain similarities and differences among affective states. While formulations of these types of models vary widely in the number of dimensions used and in the dimensions uni- or bipolarity, approaches from multiple perspectives have converged on the two dimensions of valence and arousal (Ekkekakis, 2013). For example, the circumplex model proposed by Russell (1980) is one instance of these two-dimensional models, where affective states are characterized by two underlying orthogonal and bipolar dimensions of valence (pleasure-displeasure) and activation (arousal). While distinct-state and dimensional models can seem incompatible, a hierarchical synthesis of the two approaches has also been suggested (Russell, 2003). In this view, dimensional models can be used to capture significant amounts of variation in affective states, but are unable to differentiate between discrete states, where distinct-state or categorical approaches would be more applicable. Thus, the proper choice of model depends on the target of study. In this thesis, affective states will be characterized by their placement on the axes of a two-dimensional valence (negative to positive) - arousal (low activation to high activation) model.

Psychophysiological signals

As reviewed by Ekkekakis (2013), many theorists have paid attention to the physiological changes, behavioral expressions and action tendencies associated with emotions, and moods where these effects are less distinct and more diffuse. The existence of these overt behaviours and changes, which are proposed to yield adaptive advantages (Ekman, 1992; Picard, 2000), means they can be registered and used to predict the presence of an underlying affective state. These psychophysiological signals arise from the motor division of the peripheral nervous system, which can be divided to the somatic nervous system responsible for voluntary motor activity and Autonomous Nervous System (ANS) mainly responsible for involuntary internal homeostasis maintenance. The ANS can be further subdivided to the Sympathetic Nervous System

(SNS) which is often characterized as relating to "fight-or-flight" responses and Parasympathetic Nervous System (PNS) corresponding to "rest-and-digest" types of behaviour. The involuntary nature of ANS activity has made its measurement an objective way of detecting affective states (Bota et al., 2019). Although there exists a great variety of psychophysiological signals (Cowley et al., 2016, for review), the three signal types used in this thesis will be highlighted here: cardiovascular activity, Electrodermal activity (EDA) and Accelerometry (ACC).

Cardiovascular activity

Heart function is controlled by both the sympathetic and parasympathetic branches of the ANS. Sympathetic activation via the adrenergically mediated sympathetic fibers generally increases Heart Rate (HR), while parasympathetic – also known as vagal – effects via cholinergically mediated parasympathetic fibers lower it (Ernst, 2014, pp. 35, 40–41). This antagonistic relationship between the effects of the SNS and PNS activation on heart function, is known as sympathovagal balance. Effects of SNS and PNS activity can also be seen in Heart Rate Variation (HRV) which describes how the rhythm of the heart varies. Respiration is also known to influence HRV due to respiratory sinus arrythmia – the shortening of heart beat intervals in inspiration and lengthening in expiration (Yasuma & Hayano, 2004). This effect is mediated via PNS activation, which can apply effects on heart beat intervals on almost a beat-to-beat basis, due to the quick acting cholinergic fibers (Franchini & Cowley Jr, 2004). Vasoconstrictive effects, i.e. the expansion and constriction of blood vessels, can also be studied to gain indices of SNS and PNS activity (Vinik, 2012).

One way of measuring heart activity is using Electrocardiography (ECG). In this method the potential differences arising from the contraction and relaxation of the cardiac muscle are registered by electrodes attached to the skin. In laboratory and medical contexts multiple electrodes are typically attached to the chest area, although single lead recordings can also be used to allow recordings in ambulatory settings. Timings of R-peaks representing the contractions of ventricles of the heart can be then detected from the recorded waveforms. To analyze the signal, the RR-intervals representing the time differences of subsequent R-peaks are calculated and can be used to extract HR and HRV features (Bota et al., 2019).

Photoplethysmography (PPG) is an alternative way to measure heart activity via the optical measurement of blood volume variations in the body. The PPG signal is registered by a photodiode measuring the amount of backreflected light when a light source is shined on the skin. The amount of reflected light depends on the blood volume in capillaries of the skin and deeper tissue vasculature, which in turn varies due to the pumping action of the heart. Thus, by detecting local peak timings from this signal and calculating their successive differences an Inter Beat Interval (IBI) – also known as pulse to pulse interval – series is formed. This IBI-series can then be used for analysis similarly to RR-interval series (Tamura et al., 2014). In addition, the raw PPG signal can be used to measure vasoconstriction by examining the blood volume pulse waveform amplitude changes (Bota et al., 2019).

Compared to ECG, the PPG signal is especially artefact prone. Motion artefacts caused by gravity or subject movements can easily influence the light propagation path through the tissue and corrupt the PPG signal. Furthermore, external light artefacts, skin tone, structure, and temperature, as well as blood oxygen saturation level and flow rate can cause problems and degrade the signal quality (Delgado-Gonzalo et al., 2015; Schmidt et al., 2019). Despite these downsides, PPG-sensor devices have generally good usability in daily life by allowing long-term cardiovascular recordings without the need for an ECG chest strap (Heikkilä et al., 2018).

Electrodermal activity

EDA is primarily driven by eccrine sweat gland activity, which is in turn regulated by the SNS. In more detail, sudomotor nerve fibers originating from the sympathetic chain terminate to sudomotor cells in eccrine sweat glands, which in turn activate the release of sweat through sweat ducts on to the skin. As *stratum corneum* of the skin gets saturated by sweat, the electrical resistance is decreased and conductance – the reciprocal of resistance – is increased. This conductance change can then be measured, for example by passing a small fixed direct current or voltage through the body and measuring voltage or current between two electrode leads, from which the conductance level can be calculated (Boucsein, 2012; Cacioppo et al., 2007). To gain the best signal, electrodes for measuring EDA are usually placed on areas of high eccrine sweat gland densities, such as fingers, palms or soles.

The EDA signal is characterised by a tonic baseline called Skin Conductance Level (SCL) from which phasic variations referred as Skin Conductance Responses (SCR) arise. SCRs can result either from orienting responses to environmental stimuli or be non-specific in nature, corresponding for example to respiratory activity or body movements (Cacioppo et al., 2007, p. 164). The primary signal processing associated with EDA analyses involves separating the tonic and phasic components and detecting the SCRs. Filtering or deconvolution techniques can be used to extract these signals for further feature extraction (Benedek & Kaernbach, 2010).

Accelerometry

Although often not considered to be a physiological signal per se, the measurement of body movement can be used as an informative signal in ambulatory or in-field studies (Schmidt et al., 2019). Measurement of movement can be seen as an index of skeletal muscle activity, which is under voluntary control by somatic nervous system. Since many affective theories include action tendencies, expressive behaviours and general preparation for adaptive behaviour as defining properties of emotions (Ekman, 1992; Levenson, 2014), it can be reasoned that by measuring occurred movement inferences can be made of an antecedent affective state. Different aspects of movement can be measured by a variety of inertial sensors such as accelerometers, gyroscopes and magnetometers. Three axes-accelerometer records the acceleration forces acting on the sensor in orthogonal directions, most often in units of g denoting the standard gravitational acceleration on Earth. The ACC signal can be then used to quantify movement as is, or to conduct activity recognition for additional contextual information extraction (Schmidt et al., 2019). ACC sensors have the advantage of being small, having low power consumption, low cost and being easily integrable to wearable devices (Godfrey et al., 2008). Additionally, the ACC signal can be used for movement artefact correction when integrated to a sensor platform collecting movement sensitive signals such as PPG or EDA.

Affect detection as a machine learning problem

Affect detection is typically handled as a supervised machine learning problem. Here, affective states measured using a chosen method (e.g. dimensional or distinct-states model) are predicted using features extracted from physiological signals or other observables by a chosen machine learning model, in order to generate predictions that generalize to novel data. The predominant approach in affect detection literature is to handle the task as a classification problem (Bota et al., 2019, for review). Here, the models are predicting categories of states – e.g. whether an affective state is present or not – referred as labels. These labels can be formed by discretizing a numeric or ordinal scale by some cutoff value or by defining labels based on some external factor, like the used stimuli. The task can also be handled as a regression problem if numeric affect scores are collected, which allows the models to use a more fine grained error metric compared to classification. However suitability of the chosen approach depends on goals of the study and whether classes can be formed in a meaningful way.

It is important to gather sufficient amount of observations with a preferably balanced distribution of target variable labels or values, in order to give models enough examples to identify relationships between the target variable and features, and to make sure that these predictions generalize over all target values (Domingos, 2012). To generate informative features for the model to use, statistical transformations can be applied on recorded signals in a process known as feature engineering. Although having more features is often useful, high correlations and large number of redundant features can be problematic for some model types. To correct this, a feature selection or regularization procedures can be applied to select only the most informative features.

To train models and evaluate their performance, training and test sets need to be specified. A popular method to do this is by using k-fold cross validation, where the data is randomly assigned to a k-number of non-overlapping sets, referred as folds.

Then given model will be trained on data from all folds but one and the trained models performance is evaluated on the remaining fold using chosen error metric. When this procedure is repeated over all folds, error measures can be averaged over the whole dataset. Leave-one-subject-out cross-validation is an especially relevant variation of k-fold cross-validation recommended by Schmidt et al. (2019) for the use of affect detection. In this method, the used folds are defined based on the subject from which the data was gathered from. This allows evaluation of subject-independent performance measures, i.e. how well trained model would perform on a new subjects data (Bota et al., 2019; Tohka & van Gils, 2021). Finally, some models require the selection of hyperparameters that control the behaviour of the model. This can be done manually, but it can also be automatized by running a hyperparameter selection procedure nested within each training set using a method called nested cross-validation (Tohka & van Gils, 2021).

As the primary goal of affect detection studies is studying the generalization of the trained model to novel data, results are succinctly summarised by evaluation of performance measures, often concluding the study. Model interpretation methods can however also be applied to gain insight into the function of the model, instead of treating it as a black-box. Model interpretation is desirable, since it allows identifying which features and underlying phenomena were important for model performance, guiding feature engineering for future studies and giving insight into the mechanisms that generated the informative signals. In the simplest case, interpreting a model can be done by studying the structure of the trained model itself. This however only possible with the simpler models, such as linear regression, where the number of coefficients is manageable. One way of organizing different interpretation methods is by considering whether the importance measure is global or local (Lundberg et al., 2019). Global methods assign a single importance value to every feature included in a model, summarising overall model behaviour over entire dataset. Local methods on the other hand assign features importance values for a single, individualized prediction, allowing aggregation of multiple individualized importances to estimate patterns of change

(Molnar, 2022).

Affect detection in daily life

One of the most important considerations in affect detection is establishing a reliable ground-truth for the studied affective states. The primary method to achieve this is by employing questionnaires in order to subjects self-report their current or recalled affective states using a chosen measure. Self-reports are however limited by variability in the ways different people experience and report their affective states (Schmidt et al., 2019). Training personalized models for each subject has been found to be a good way to combat this problem (Taylor et al., 2020; Tervonen et al., 2020), but applying this approach requires a sufficient sample size collected from each subject. Another way of ensuring that the measure values correspond to the affective state of interest, is the usage of a pre-validated set of stimuli that have been found to reliably elicit the desired affects in different subjects (Bota et al., 2019). These emotion elicitation procedures are most often used in a controlled laboratory setting to maximize the self-reported affective state ratings for analysis. There also exists studies where the ground-truth is defined directly by the employed conditions or stimuli, but in these cases it is important to verify that the desired affective states were successfully evoked using questionnaires (Bota et al., 2019).

Affective states can also be studied nonspecifically or without a preceding elicitation procedure in subjects daily lives by using the Ecological Momentary Assessment (EMA) method, also known as the experience sampling method (Shiffman et al., 2008). Using this method, questionnaires for affective states are presented to subjects throughout their daily life, either at specific time intervals, pseudorandomly or when associated with some event of interest. Modern EMA-based studies tend to employ mobile platforms to do this, which also allows the collection of contextual information. While powerful at generating large amounts of labeled data for affect detection, care should be paid to the duration, frequency and reward structure associated with the EMAs to keep response rates high in longer term studies (Schmidt et al., 2019).

Laboratory studies using affect elicitation procedures and EMA-based studies in daily life represent a trade-off between certainty in the self-reported elicited affective state versus generalization of the results to daily life (Bota et al., 2019). While affect elicitation in laboratory conditions is the predominant approach in affect detection, evaluation of the results in an ecologically valid context is of special interest for affective computing applications, many of which are practically applied in naturalistic environments. Differences in these two contexts of study can be also seen in signal quality and ground-truth distributions: Devices recording physiological data in in-field studies are practically limited by power consumption, memory constraints and general usability in daily life. Some signal types employed in laboratory settings like electroencephalography or eye movements are not practical for long term ambulatory recordings, while EDA, ECG, PPG and other signal types that can be recorded in both contexts often have their sample rates reduced in order to allow constant recording. Also, movement artefacts and other environmental effects which can be controlled in a laboratory can easily corrupt the signal in the wild. Even with good quality data, upload for offline analysis is reliant on device connectivity and implementation. Laboratory-based affect inducement procedures are selected due to their ability to elicit desired affective states in the limited time allotted for the study. This approach is likely to produce a much more intense and balanced distributions of affective states compared to daily life, where positive affect levels of higher intensity and frequency are observed over negative affects (Komulainen et al., 2014; Zelenski & Larsen, 2000). Laboratory-based studies are also practically restricted to the study of emotions or core affect due to their limited durations, causing the interdependencies between emotions, moods (Ekkekakis, 2013), personality (Komulainen et al., 2014) and environmental factors go unappreciated. These reasons indicate that the excellent performances achieved in induced laboratory-based affect detection studies (Bota et al., 2019) are not likely to generalize to daily life and that affective states should be examined in daily life to capture the variability present in daily affective phenomena.

There exists few affect detection studies using EMAs to detect affective states as they occur in daily life as reviewed by Bota et al. (2019). Although the majority of these studies focus on the healthcare context predicting stress or panic attacks, there are some that focus on affective states more broadly. For example Jaques et al. (2016) and Taylor et al. (2020) predicted self reported sad-happy classes using EMA-based surveys in a student population, reaching 14% and 15% accuracy improvements over expected baseline performance using the best model for single task learning. However, in these studies the original mood scorings were collected in a 0-100 range, which were later discretized to classes by discarding the middle 40% and 20% of affect ratings in the respective studies to make the task easier. Zenonos et al. (2016) introduced a system to predict a variety of office workers moods during their workday using physiological and inertial signals. The system was able to improve accuracy by 12.4% on average in generalized models over the baseline while predicting mood classes discretized to 5 steps of intensity. The study however included only a small sample of 4 subjects. Zhu et al. (2016) presented the novel approach of predicting moods as angles on Russels circumplex complex (Russell, 1980) using inertial and contextual signals. A mean absolute error rate of 0.24π radians (out of a maximum error π) was reported, which outperformed statistically significantly the used benchmarks. While all these studies showed modest performance in daily mood prediction, there was significant variation in the ground-truth definitions used, making comparison difficult. The trained models were also treated as black boxes, with up to hundreds of inputted features, making interpretation of the relationships found by the models very difficult.

In this thesis, the feasibility of an affect detection system for predicting a variety of self-reported affective state scorings from EMAs using physiological and inertial signals from wearable sensors is evaluated. A novel masking algorithm for IBI-series data will be presented, along with a signal processing pipeline description. Model performances will be evaluated using leave-one-subject-out cross-validation against a baseline and modern model interpretation methods will be used to identify important features and relationships governing model performance. Due to the exploratory nature of the study, no specific hypotheses are presented.

Methods

Participants

The data used in this thesis came from a study conducted as a part of *Sisu at Work*-project[1], by University of Helsinki and VTT. The participants (N = 82, 30 male) were aged between 24 and 58 (M = 41.03, SD = 8.51). They were knowledge workers recruited from four different Finnish organizations mailing lists and occupied various positions from expert roles to managers. The study was approved by the VTT ethical committee (27.5.2019) and University of Helsinki ethics review board. The participants also signed a form of consent before participating.

The participants were informed that the study aimed to further the development of digital tools for individual well-being in the workplace. As compensation for participating, every participant received a summary report about their mental resources and sleep quality during the experimental period. Additionally, the participants who actively took part in the study received two movie tickets. Participants were required to have Android phones for mobile applications used in the study. All participants passed health criteria evaluated based on subjective health and medications and no participants dropped out during the experimental period.

Procedure

The study analyzed in this thesis consisted of an experimental period of three weeks, run in stages between 28.5.2019-12.12.2019. During this period, participants answered EMA questionnaires using a mobile app developed by VTT. The EMA questionnaires were queried thrice a day: In the morning at 10 am (9 am for the first organization), in the afternoon at 4 pm and in the evening at 9 pm. Every questionnaire contained questions about experienced affective states, social company and activities, as

[1] Principal investigator Ilmari Määttänen, Academy of Finland decision number 313399

well as more specific questions about sleep, sisu states and stress once a day. Each questionnaire took around a minute to fill. All questions were given in Finnish.

The participants were also given a PulseOn (PulseOn Technologies Ltd., Espoo) activity wristband and a Moodmetric (Vigofere Ltd., Finland) smart ring to wear for the duration of the study period. Participants were instructed to wear the PulseOn wristband on their non-dominant hand and to try to use the devices continuously during the three week period. Instructions were given to charge the wristband every day and the smart ring every week. The same application that was used for EMA questionnaires also uploaded data from the PulseOn wristband to cloud, while data upload from the smart ring was done using Moodmetric-application.

Materials

Ecological momentary assessments

The affective state questions queried accuracy of different affective state claims of the form "In the last 30 minutes, I've felt like:" on a 7-step Likert-scale slider, with answers ranging from 1 (*not at all*) to 7 (*completely*). Eight Finnish translations of affective state claims were queried in this manner: angry [*vihainen*], anxious [*ahdistunut*], enthusiastic [*innostunut*], focused [*keskittynyt*], happy [*iloinen*], sad [*surullinen*], satisfied [*tyytyväinen*] and vigor [*virkeä*]. The affect claims were designed in adherence to Positive and Negative Affect Schedule by Watson et al. (1988). Missing values in the affective state answers were marked with 0. Questionnaires that were answered within three hours of each query and that had at least one non-missing affective state answer were included in analysis. In cases where there was multiple answers within the three hour period, only the first one was considered. Altogether, this yielded 3537 EMAs with at least one affective state rating.

Significant amounts of missing values was noted in answers to negative affective state questions (angry = 79.1%, anxious = 62.0%, sad = 80.9%), compared to positive affect states (range = 2.5 - 4.2%). The number of affective state scores with value 1 seemed to be underrepresented in negative affect scores, where the distributions where otherwise highly positively skewed. Examination of correlations of missing values also indicated that the proportion of missing values in one affect was inversely related to scores in affects of opposite valence. This indicated the possibility that participants had used omitting an answer as a way to report the lowest value on the scale 1: *not at all*. There was however no clear way of determining when omitting an answer was used in this manner and when it was genuinely missing. For this reason, missing values in EMA affect claim answers were excluded from further analysis.

Wearable devices

PulseOn activity wristband reported IBI-series values measured using using green and infra-red PPG signals, as well as 12.5 Hz 3-axes ACC in a dynamic range of ±8 *g*. While it was also possible to record 25 Hz ACC and PPG signals using the PulseOn wristband, this was not done in order to save recording space and to limit device power consumption. Heart beats were detected using a proprietary algorithm by PulseOn, with previous research indicating heart beat detection accuracy up to 99.57% in night time recordings compared against an ECG-baseline (Parak et al., 2015). A Signal Quality Estimation (SQE) rating calculated from PPG wave morphology and ACC signals ('Accuracy of Beat-to-Beat Heart Rate Estimation Using the PulseOn Optical Heart Rate Monitor', 2018) for masking out artefacts was also provided by the wristband. The wristband also reported HR, predicted activity class and step counts, which were not included in the analysis. PulseOn recordings yielded 348.0 hours of recordings on average per participant (SD = 143.6 hr, range = 14.9-564.5 hr).

The Moodmetric smart ring recorded per-minute SCR counts, SCL, step count and Moodmetric-index score. The Moodmetric index was calculated using a proprietary algorithm from EDA signal. It ranged between 0 and 100, with high values indicating high SNS activity and low values PNS activity according to Jussila et al. (2018). In more detail, value 1 indicated deep sleep, below 30 relaxation, 50 mental activity and 100 extreme stress, excitement, anxiety or fright (Jussila et al., 2018). The Moodmetric-index is calibrated to user specific physiological activity level, requiring a minimum calibration period of 12 hours. While it was also possible to receive 3 hz

EDA-signals from the Moodmetric devices, it was omitted over the simplified per-minute recordings. Comparison of Moodmetric prototype model against a laboratory-grade EDA sensor found a cosine similarity of .83 between extracted EDA-features, which was deemed adequate for in field recordings (Torniainen et al., 2015). Moodmetric recordings yielded 380.5 hours of recordings on average per participant (SD = 94.5 hr, range = 86.4-506.4 hr).

Preprocessing

Even though Parak et al. (2015) found the beat-to-beat detection accuracy of PulseOn to be excellent, these analyses were limited to nighttime recordings without movement. Movement artefacts are however ubiquitous in daytime PPG recordings, since even small movements of the sensor can affect the LED light propagation path, and thus the resulting signal (Delgado-Gonzalo et al., 2015). Movement artefacts can be removed from PPG signals by a variety of methods, being a popular subject of signal processing competitions (Pankaj et al., 2021, for review), but all of these approaches require the raw PPG signal, which was not available here. Alternatively, a masking algorithm can be used to determine the beat to beat quality of given IBI-series. Applying a masking procedure can however lead to significant amounts of data loss in daytime recordings, with 'Accuracy of Beat-to-Beat Heart Rate Estimation Using the PulseOn Optical Heart Rate Monitor' (2018) reporting 67.6% IBI observations being excluded using the PulseOn SQE rating as a mask during daytime activities. When these issues are combined with non-wear time, ectopic beats, challenges in data upload and calibration issues in a sizeable, long-term in-field study, then the required signal processing can be extensive.

Accelerometry autocalibration procedure

In the absence of movement, an correctly calibrated ACC signal should represent the gravitational pull of the earth and residual noise (van Hees et al., 2013). A visual examination of the recorded ACC signals however indicated that the Euclidian Norm (EN) $\sqrt{a_1^2 + a_2^2 + a_3^2}$ of the triaxial signals consistently did not sum to the expected value of 1 g during periods of low movement. This was problematic since it indicated that EN-based ACC features would include a sensor-specific calibration bias. To correct this, an autocalibration method for ACC signals developed by van Hees et al. (2014) and implemented in GGIR-package (Hees et al., 2021) was used. The method first segments the ACC signal to 10 second segments, detects ones with low movement and then calculates the best fitting model intercepts and coefficients for the triaxial signals using an iterative closest-point fitting process to correct the error. Before calibration the sensor specific low-movement segments differed on average by 28.5 mg (range = 19.4-37.9 mg) from expected and after calibration by 2.7 mg on average (range = 1.4-8.4 mg).

Segmentation

Signals recorded by the PulseOn wristband and Moodmetric ring were segmented to 30 minute windows before mobile questionnaire response times. This yielded 2,918 PulseOn and 2,970 Moodmetric signal segments with any data present. Missing data was found in 240 PulseOn data segments, while 521 segments of non-wear time was noted by visual inspection. The non-wear time periods, which were characterized by a static ACC-signal and IBI-values ranging around 400-600 ms over at least a 1-minute period, were manually labeled in the data and removed. The resulting IBI segments contained only 17.7 % completely reliable IBI-observations according to the PulseOn SQE rating. For the Moodmetric-observations, 753 segments with missing data was found. Also 112 Moodmetric signal segments that were found to fall within the first 12 hours of recordings were excluded to allow user specific calibration of the Moodmetric index time to function.

Inter beat interval masking procedure

While the PulseOn SQE rating was deemed quite reliable based on visual examination of passed values, it was found to be overly conservative as a masking algorithm. Therefore, a new masking algorithm was developed using amount of movement and IBI-value dispersion (Algorithm 1). In this algorithm, only IBI-values within 250 ms of a 40 observation rolling centered median and IBI values with less than 50 *mg* of movement during the previous 4 seconds measured by Root Mean Square of Successive Differences (RMSSD) of EN of calibrated ACC signal, were passed. As an additional condition, every passed IBI observation had to belong to a continuous sequence of at least 3 other passed IBI-observations. This was done in order to ensure that every passed IBI-observation belonged to a continuous sequence of values that could be used to calculate HRV-features, many of which require lagged or differenced values. After applying the mask 44.7 % of the IBI observations passed, while visual examination of the results indicated that the accuracy of the mask was good.

Algorithm 1: IBI masking algorithm

Input : List D_i, $i = 1, 2, \cdots, n$, with absolute differences of IBI-values to 40 observation rolling centered median and list M_i, $i = 1, 2, \cdots, n$, with amount of movement measured by RMSSD of EN from previous 4 seconds of each IBI observation. Parameters *Dlim* and *Mlim* for cutoff values of IBI dispersion and movement.

Output: IBI mask.

```
$Mask = [\ ]$
$prevMask = 0$
for $i \leftarrow 4$ to $n$ do
    if $prevMask \neq 0$ then
        if $D_k < Dlim$ AND $M_k < Mlim$ then
            $Mask_k \leftarrow 1$
        else
            $Mask_k \leftarrow 0$
            $prevMask \leftarrow 0$
        end if
    else
        if ALL $D_{k-3...k} < Dlim$ AND ALL $M_{k-3...k} < Mlim$ then
            $Mask_{k-3...k} \leftarrow 1$
            $prevMask \leftarrow 1$
        else
            $Mask_k \leftarrow 0$
        end if
    end if
end for
return $Mask$
```

Feature extraction

A wide variety of features was chosen to be extracted from the preprocessed signal segments to allow models to make use of the most informative ones. Feature abbreviations and explanations are presented in table 1. Two primary signal types were extracted from the IBI-series segments: HR and HRV. HR features were calculated as segments IBI-timings mean, median and percentile values. HRV features are often divided to time-domain, frequency-domain and non-linear features (Shaffer & Ginsberg, 2017). Time-domain features included SDNN, CVNNI, IQRRR and HRVTI calculated from IBI-values, and pNN20, pNN50, SDSD, RMSSD, CVSD and MADRR calculated from differenced IBI-values. Non-linear features, which quantified unpredictability and complexity of the signal included CVI, CSI, modified CSI, SD1, SD2 and SD2/SD1-ratio calculated from lagged IBI-values (Jeppesen et al., 2014). Frequency-domain feature calculation was done by first linearly interpolating given IBI-value sequence to time domain at 4 Hz, centering the signal and then applying fast Fourier transformation for frequency decomposition, yielding the VLF, LF, HF, LFNU, HFNU and TotPow features.

After extraction, features from the IBI-signals were normalised to participant specific nighttime values. Within-subject normalization is a method that has been previously found to increase model performances in affect detection studies (Pettersson et al., 2020; Tervonen et al., 2020). Although subject-specific normalization is usually performed using the same data that is being normalized, here the abundant relatively good quality night time recordings were able to be used as the baseline. To calculate the normalization statistics, night time recordings of IBI-series between 1:00 AM and 6:00 AM were segmented to 30 minute, non-overlapping segments. From these segments, the ones with over 85 % completely reliable observations according to PulseOn SQE rating were selected and had unreliable observations excluded. This yielded 8,112 segments in total, with 100 observations on average for each participant (SD = 53.4, Range = 7-195). The same features as outlined previously were extracted from these segments, from which participant specific means and standard deviations

Table 1

Extracted features and their descriptions. Features included in the final models are marked with bold.

Signal	Feature	Description
PPG	**HRmean**, HRmed	Mean and median HR
	HRp05, HRp25, HRp75, **HRp95**	5th, 25th, 75th & 95th HR percentiles
	HRmin, **HRmax**	Minimum and maximum HR
	SDNN, SDSD	Std of IBIs, std of IBI differences
	pNN20, **pNN50**	Proportion of IBI differences differing by more than 20 & 50 ms
	RMSSD	Root mean square of IBI differences
	CVNNI, **CVSD**	Ratio of SDNN and mean IBI, and RMSSD and mean IBI
	VLF, **LF**, **HF**, **TotPow**	Power in (0.0033-0.04 Hz), (0.04-0.15 Hz), (0.15-0.4 Hz) & (0-0.4 Hz) frequency bands
	LF/HF	Ratio of LF and HF
	LFNU, **HFNU**	LF and HF with normalized units
	LFNU, **HFNU**	LF/(TotPow-VLF) and HF/(TotPow-VLF)
	HRVTI	HRV triangular index using a binwidth of 100 ms
	IQRRR	Interquartile range of IBI values
	MADRR	Median absolute IBI difference
	CVI, CSI, **modified CSI**	Cardial vagal index, (modified) cardial sympathetic index
	SD1, SD2, **SD2/SD1**	Poincaré plot minor and major axis length, ratio of major and minor axes
ACC	VMUmean, **VMUmed**	Mean and median VMU
	VMUskew, **VMUmax**	Skewness and maximum of VMU
	VMUp75, **VMUp95**	75th and 95th percentiles of VMU
	ENMONZmean, ENMONZmed	Mean adn median ENMONZ
	ENMONZskew, ENMONZmax	Skewness and maximum of ENMONZ
	ENMONZp875, **ENMONZp975**	87.5th and 97.5th percentiles of ENMONZ
	STEPmean	Mean number of steps
EDA	**MMmean**, MMmed	Mean and median MM-index
	MMsd, **MMtrend**	Std and trend of MM-index
	SCRNmean, **SCRNmed**, **SCRNmax**	Mean, median, and maximum skin conductance response count

Note. Despite the multitude of features, they can be thought to summarise only four primary signal types: HR, HRV, movement and EDA.

were calculated and used to normalize the used features for each participant.

The features extracted from ACC signals for affective state prediction are not as well established than in the context of ECG and PPG research. Therefore, two different approaches in terms of gravitational component removal were chosen for ACC feature extraction. First one was using the Euclidian Norm Minus One with Negatives set to Zero (ENMONZ) from the calibrated ACC signal. This feature introduced by van Hees et al. (2013) weighs vertical movements more heavily than others and has been found to be a good daily physical energy expenditure estimation metric. Another approach was measuring Vector Magnitude Units (VMU), also known as high frequency filtered EN. This measure was chosen since it has been previously found to be a good predictor in a highly similar study by Määttänen et al. (2021). It was extracted from the uncalibrated triaxial ACC signal using a 4th order high-pass Butterworth filter with a cut-off frequency of 0.5 Hz to keep the formulation of the feature consistent to Määttänen et al. (2021). After these measures were transformed from the ACC signals, they were epoched at 1 second intervals, from which statistical features were extracted.

Statistical features were extracted from Moodmetric-index and SCR counts, while SCL measures were found to contain too many missing values to be used for feature extraction. The average number of steps reported by Moodmetric ring was also used as a feature.

Final dataset

Two criteria were used select the observations included in the final dataset; each feature value had to be formed from at least 5 minutes of data, and each participant needed to have at least 5 unique observations. These were chosen to ensure a minimum sensibility criteria for the extracted feature reliability and to limit the variance in model performance estimates due to participant specific observation counts. In practice, the formation of features from differenced or lagged IBI-values limited the number of observations the most due to the requirement of passing the mask with both values. After applying these criteria and joining the features together, a dataset with 1642 observations was formed, containing 24 observations for each subject on average (SD =

11.6, Range = 5-50). Missing values in EMA questionnaire affective state answers further limited the observation counts for affect-specific models (Angry = 214, Anxious = 552, Enthusiastic = 1,591, Focused = 1,585, Happy = 1,612, Sad = 189, Satisfied = 1,611, Vigor = 1,610).

A final sensibility check was performed based on the extracted feature correlations. First, as noted by Ciccone et al. (2017), SD1 and SDSD were to be identical to RMSSD, as well as CSI to SD2/SD1-ratio and were consequently removed. A further limit of no feature having a Pearson correlation over .9 with another feature in the data was also applied. This was done in order to reduce model training times, increase interpretability and to control the effects of multicollinearity on feature importance estimates (Strobl et al., 2008). To determine the features to retain, usage in similar studies and performance in preliminary analyses was considered. As a result, many of the ENMONZ, HR and HRV-features were excluded (see Table 1). Final dataset feature correlations are reported in appendix A1.

Model fitting

Random Forests (RF) and eXtreme Gradient Boosting (XGBoost) models were fitted to predict affective state scores using the extracted features. Random forests, by Breiman (2001), is a method for training ensembles of decision trees with added randomness to prevent overfitting. In this method, predictions are generated by aggregating predictions over a group of trained decision trees – corresponding to averaging in a regression context. Randomness is injected to each tree by means of bootstrapping, or selecting a random subsample of observations with replacement, referred as in-bag observations, for training each tree. Further randomness is then introduced to by only considering a random subset of features at each split and selecting the feature that provides the best split according to the used objective function. The tree training and split selection is then continued until the desired number of trees is reached. XGBoost models are also a class of tree ensembles, where the results from multiple decision trees are combined using boosting instead of aggregation. Boosting means training each decision tree serially, with each tree trying to minimize errors of the previous tree. This allows sequentially combining together simple decision trees to form an overall strong learner (Chen & Guestrin, 2016; Friedman, 2001).

These two algorithms were chosen due to their ability to handle the large number of possibly uninformative or highly redundant features included in the analysis. While a feature selection procedure like wrapper search could have also been used, it was omitted due to prohibitive computational requirements. The chosen models also had the convenience factor of not requiring feature normalization and having readily implemented feature importance methods.

Model performance was evaluated using leave-one-subject-out cross-validation. This method was chosen to evaluate model generalization performance on a previously unseen users data as recommended by Schmidt et al. (2019). Nested within these subject-specific folds, a 3-fold cross-validation was run while retaining subject-specific groupings to select model hyperparameters (Tohka & van Gils, 2021). The best performing hyperparameter set in the inner folds was used to train the model for outer fold predictions. Model fitting, cross-validation and hyperparameter selection was done using the *mlr3*-framework (Lang et al., 2019), with packages *ranger* (Wright & Ziegler, 2017) and *xgboost* (Chen et al., 2021) being used as model implementations. RF models were trained using 1,000 trees, with maximum tree depth and number of features to consider at each split selected via tuning. XGBoost models were trained with maximum of 1,000 trees with a early stopping criterion of 5 rounds, and had maximum tree depth, learning rate and number of features to consider for each tree tuned.

Statistical testing and model interpretation

Statistical tests were performed to see whether the models were able to improve predictions over a baseline model predicting the training set outcome variable mean for each fold. Paired one sided t-tests comparing fold-specific root mean square error decrease from baseline models to RF and XGBoost models were utilized to do this. Benjamini-Hochberg method (Benjamini & Hochberg, 1995) was used for adjusting p-values for multiple comparisons within each model type.

Trained model behaviour was analysed using global and individualized feature importance methods to identify the most important features and to see how changes in different features impacted model predictions. Permutation importance was utilized to quantify feature importance in RF models. In this method, feature importance is measured as the amount of performance degradation when evaluated feature values are permuted in a given test set using the previously trained model. If the permuted feature is important, then randomizing the feature values should cause a significant increase in the model error (Breiman, 2001). Out of bag observations were used to calculate the permutation importance values by the *ranger* package (Wright & Ziegler, 2017), with the final importance values calculated by averaging across all trained models for a given task/affect. Although permutation importance is often regarded as a robust feature importance method, it should be noted that RF permutation importance measures can be overestimated for mutually correlated features (Nicodemus et al., 2010).

SHapley Additive exPlanations (SHAP) values were used to examine XGBoost feature importances and individualized feature attributions. SHAP values are based on a combination of approaches from cooperative game theory and local explanations to calculate additive feature impacts on model outputs while ensuring several desirable properties, which are not guaranteed with other popular tree ensemble feature importance methods (Lundberg et al., 2019; Lundberg & Lee, 2017). SHAP values essentially describe model output changes from expected value by averaging the impacts on a prediction when adding an evaluated feature to a model when some subset of other features are missing, over all possible orderings of adding the evaluated feature to the model. Feature being missing means being replaced by its expected value in this context. Tree SHAP algorithm was used to calculate the SHAP-values using the *xgboost*-package (Chen et al., 2021). SHAP summary plots were used to visualize all SHAP-values for given task test sets, for identifying most important features for the XGBoost models and overall patterns of changes. Since examination of SHAP summary plots can be challenging due to the amount of visualized information, SHAP dependence plots were used to highlight selected patterns of SHAP value impacts on model predictions. All analyses were done in R 4.1.3 (R Core Team, 2022).

Results

Model performance

The paired t-tests comparing fold specific performances of trained models against baseline models found improvements in RF models for affects vigor (p = .005) and focused (p = .013), while non-significant improvements were noted in affects enthusiastic (p = .155) and angry (p = .621) (Table 2). XGBoost models found similar results with a statistically significant improvement in affect vigor (p = .001) and non-significant improvements for affects focused (p = .068) and enthusiastic (p = .431). The magnitude of improvement was modest in general, with a 0.045 - 0.051 average root mean square error decrease noted for vigor and 0.025 and 0.013 for affects focused and enthusiastic respectively. Inspection of the percentage of folds/subjects that showed improvements over baseline, indicated percentages ranging from 68.1 to 71.0 % for affect vigor, 65.2 to 68.1 % for focused and 59.4 to 63.8 % for enthusiastic.

Inspection of predicted value distributions reported in appendices B1 and B2 indicated that model performance for affects vigor, focused and enthusiastic relied mainly on the models ability to discern low affect scores from average and high values. Only models for affect focused seemed to be able to discern high affect scores from low and middle ones.

Feature importance

ACC and HRV features were found to be most significant for RF models in terms of permutation importance (Figure 1). Especially permutation of the median and 75th percentile of VMU decreased performance heavily, followed by HRV features pNN50, CVI, MADRR and CVSD. Minimum HR was the most important HR-derived feature while EDA-activity described by MMmean and MMsd were found to be important for affect vigor.

Table 2

Model performance statistics and paired one-sided t-test results comparing performance of trained model against baseline. imp % is the percentage of folds where performance improved over baseline model.

		Random Forests				XGBoost			
Task	N	$\bar{d}$	t	adj. p	imp. %	$\bar{d}$	t	adj p	imp. %
Angry	22	-0.009	-0.502	0.621	50.0	0.016	0.332	0.999	50.0
Anxious	42	0.004	0.453	0.898	52.4	0.030	1.450	0.999	40.5
Enthusiastic	69	-0.013	-1.592	0.155	63.8	-0.013	-0.995	0.431	59.4
Focused	69	-0.025	-2.796	0.013	68.1	-0.025	-2.163	0.068	65.2
Happy	69	0.001	0.319	0.898	63.8	0.032	2.262	0.999	47.8
Sad	17	0.015	3.729	0.999	23.5	0.078	1.369	0.999	29.4
Satisfied	69	0.003	1.176	0.999	56.5	0.034	3.238	0.999	31.9
Vigor	69	-0.045	-3.352	0.005	68.1	-0.051	-3.754	0.001	71.0

Note. p-values are adjusted for multiple testing using Benjamini-Hochberg method within each model type.

Examination of SHAP dependence plots for XGBoost models predicting positive affects indicated a pattern of decreased scores when movement measured by VMUmed and VMUp75 was under 8-16 *mg*, while low maximum VMU values were found to increase affect focused scores (Figure 2). High values of minimum heart rate increased positive affect scores, as did high short term HRV indexed by high CVSD, MADRR and low a SD2/SD1 ratio (Figure 3). Low values of average of Moodmetric index were found to decrease positive affect predictions, while high dispersions and trend of Moodmetric index also decreased the scores. High number of SCRs was also found to slightly decrease affect focused scores (Figure 4). Interestingly, many of the effects detailed in the dependence plots could be seen in all of the examined positive affects, although the magnitude of these relationships depended on the achieved model performance. To inspect the SHAP value impacts of all features for each affective state, SHAP summary plots are reported in appendices C1-C8 due to space constraints.

Figure 1

Permutation importance results from RF models. The error bars describe 95 % confidence intervals of the means. Note the legend at lower right corner.

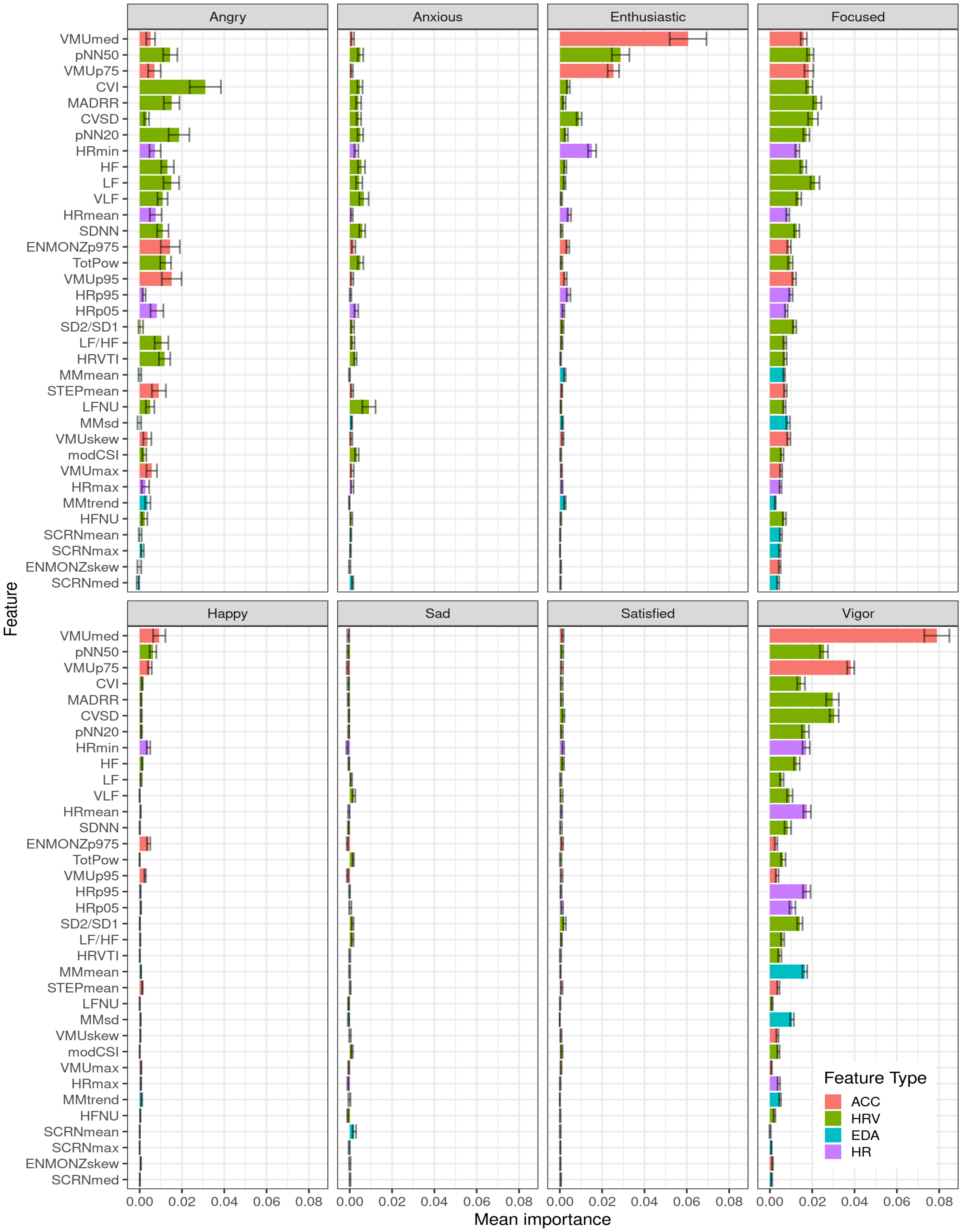
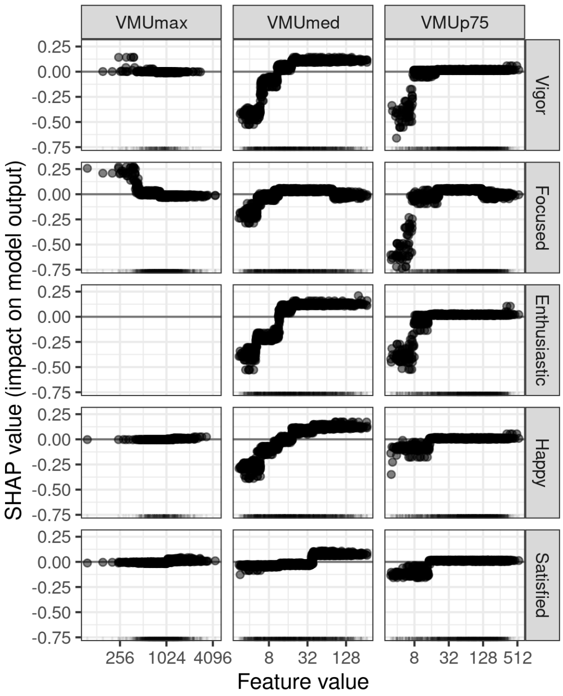


Figure 2

SHAP dependence plots for selected ACC features representing movement, and positive affects. Feature values are in units of mg and are scaled by a base 2 logarithm on for visualization.



Note. Empty dependence plot for feature VMUmax for affect Enthusiastic indicates that XGBoost models found no splits using this feature.

Discussion

The aim of this thesis was to evaluate the possibility of affect detection in daily life using wearable sensors and EMAs – a context of study that has received little attention despite its importance for affective computing applications. In this study, a signal processing pipeline was implemented to deal with artefacts and other problems in data, a variety of features were extracted and machine learning models predicting self-reported affective state scorings were trained. Model performance was evaluated using leave-one-subject-out cross-validation, and achieved performances were compared against a baseline. Permutation importance and SHAP values were used identify important features and the relationships governing model performance.

The models were able to improve performance over baseline when predicting high activation positive affective states vigor, focused and enthusiastic. Permutation importance highlighted the significance of inertial and HRV features. Examination of

Figure 3

SHAP dependence plots for selected HRV (CVSD, MADRR, SD2/SD1) and HR (HRmin) features and positive affects. Feature values are in normalized units.

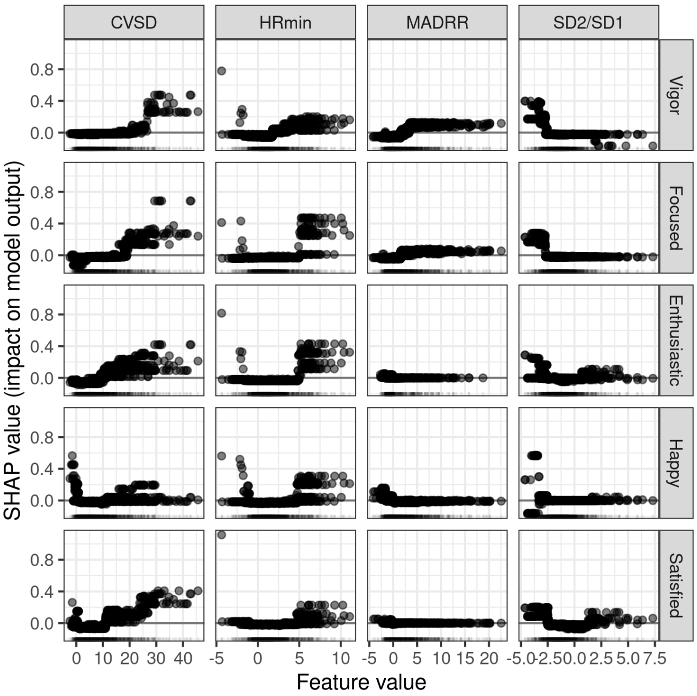
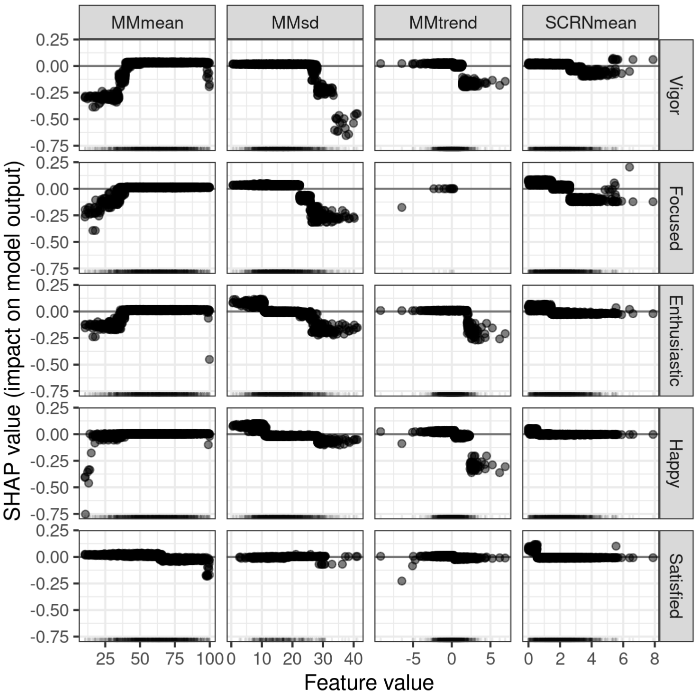


Figure 4

SHAP dependence plots for selected EDA-based features and positive affects. Feature values are in original units.



SHAP values indicated that the model performance mainly relied on connecting sedentary periods of low movement and SNS activation with decreased positive affect predictions, while high HRV, minimum HR and movement slightly increased positive affect predictions. Although the achieved performances are modest and the used feature importance analyses exploratory, these results are valuable in guiding future research in a little investigate context of study.

Model performance

In many regards the observed results were within expectations based on previous research. First, the observed predominance of high intensity and frequency positive affective states over low frequency and intensity negative affects corresponded to previous surveys in non-clinical samples (Komulainen et al., 2014; Zelenski & Larsen, 2000). This imbalanced negative affective state score distribution combined with the issue of missing values caused significantly smaller samples sizes for negative affect models. Together these factors likely severely limited predicting negative affective state scores.

The achieved performance in detecting the positive affective states seemed to reflect placement on an underlying dimension of arousal or degree of associated physiological activation. Vigor for example could be placed highly on Russels circumplex models (Russell, 1980) activation dimension, followed by focused or enthusiastic, happy and satisfied. This correlation between placement on arousal axis and performance makes sense, since both quantifiable physiological signals and arousal can be seen arising from SNS activation (Bota et al., 2019). The affective states where consistent improvement over baseline was noted (vigor, focused, enthusiastic) also closely correspond to the Positive and Negative Affect Schedule subscales of active, attentive and enthusiastic, indicating that the underlying dimension could possibly be characterized as positive activation (Watson et al., 1988; Watson et al., 1999).

Although not straightforward, the percentage of folds where improvements were noted over baseline model could be used as a proxy for classification accuracy (Table 2).

Based on these values ranging from 59.4 to 71.0 % in affects vigor, focused and enthusiastic, the achieved performance is comparable to other affect detection studies predicting EMA-based affective states in naturalistic environments. Compared to Jaques et al. (2016) and Taylor et al. (2020), which use the only known comparably sized EMA-dataset for affect detection, the performance here could be interpreted as being good since all affective state scores were used, instead of excluding ambiguous values in the middle of the scale.

Model interpretation

Based on examination of SHAP values, the XGBoost models utilized three primary effects to predict the affective states linked with positive activation: periods of low movement, low EDA level and high HRV. These periods of low movement, defined by VMUmed and VMUp75 features below 8-16 mg, could be possibly characterized as sedentary states like sleeping or low effort activities without much associated movement (Figure 2). Moodmetric EDA measurements impacted predictions similarly to low movement, with MMmean values below 40 described as corresponding to states of deep to regular relaxation (Krupić et al., 2021) lowering predicted positive affect scores. Similar reduction of positive affect scores was noted with high dispersion and trend of the Moodmetric index (Figure 4). This could be due to transient episodes of sympathetic arousal over low level baseline EDA, and therefore these features could be informative of periods of low EDA level. High HRV marked by high CVSD, MADRR and low SD2/SD1 ratio showed increases in the predicted positive affect scores (Figure 3), although many more inconsistent effects were noted in the various other HRV features extracted. Additionally high minimum heart rate was found to increase positive affect predictions.

The importance of low movement and EDA level could possibly be interpreted as behavioral and physiological manifestations of low positive activation. However the increase in positive affect predictions in high HRV could be seen as conflicting evidence, since HRV is mainly linked with PNS activation, often described corresponding to "rest-and-digest" types of activities. The relationship between positive affect and HRV is however complicated, with high activation positive affects found to increase HRV when examined as between subject traits, but showing the opposite effect when examined within subjects (Papousek et al., 2010; Schwerdtfeger & Gerteis, 2014). One possibility here is that the personalized nighttime normalization procedure didn't succeed in removing between subject variations in HRV measures, which the models were able to utilize to predict higher trait positive affectivity levels. Similar relationships of increased movement, HRV and HR related to positive affectivity in daily life has been found by Määttänen et al. (2021). Although here positive affect was calculated from relatively lower-activation positive affects content, happy and joyful, and the examined HRV variable was 5 minute SDNN aggregated over 45 minute periods. Still the used measures are closely conceptually associated and can be seen representing the same underlying phenomena.

Finally it should be reminded that examination of SHAP values is a novel approach in interpreting model performance and that no statistical testing procedure to identify significant relationships is performed. The examined SHAP dependence plot features were selected by manual inspection of SHAP summary plots and thus subject to individual choice, although the selection of features showing the largest and most consistent SHAP values was prioritized. Regardless, the corroboration of identified relationships to previous research and ability visualize detailed non-linear relationships holds great promise in guiding feature extraction for affect detection and highlighting interesting phenomena for future study.

Limitations and future research

In many regards the issue of ground truth measurement and conceptualization is one of the most important problems in any supervised machine learning study, which also be noted here. For example, the issue of missing values, and their exclusion due to the reasons outlined in introduction likely played major part in influencing the observed results. For the negative affects which already showed a positively skewed, unipolar distribution this exclusion of 66 - 88 % of samples likely meant the loss of signal. The positive affect predictions likely suffered also, since the largest prediction accuracies were noted in the lower end of positive affect scorings. Future work is needed to identify a method for including the missing values in analyses and to prevent responder ambiguity between lowest value of the scale and answer omission.

Although measurement of affective states by EMAs guarantees a high degree of ecological validity, it also introduces difficulties of measuring and interpreting the measured phenomena. In this study affective states were queried in the form of accuracies of affect claims of the form "In the last 30 minutes, I've felt like:" in Finnish. This in principle allowed conceptualization of the measured phenomena as core affect in the lines of Russell (1980), mood according to Watson et al. (1988) or even as sequences of emotions, since Ekman (1992) had reported of an instance where subject reported multiple emotions in close proximity as one. The questions also left room for interpretation whether they queried the existence of an instance of an affective state, or constant presence of an affect during last 30 minutes. Also some of the affects queried might have different meanings in Finnish and English. For example the Finnish word *virkeä* has a meaning of being well rested which is not so explicitly present in the translation *vigor*. Additionally, interpersonal variability of experiencing and reporting affective states further complicates the picture. These effects cause ambiguity in what the studied phenomena is, requiring additional contextual information about the antecedent events and timings of experienced affects to identify different components. Presence of this ambiguity likely limits the expected performance using a supervised machine learning approach.

The poor signal quality in this and other ambulatory affect detection studies is currently one of the major limiting factors for performance. Low signal quality can however be compensated by adding more signal types: First by examining raw PPG signal amplitude measures of vasoconstriction associated with SNS activation could be extracted. Additional features could be also extracted from EDA such as SCR amplitudes or including the SCL which was omitted here due to missing data. There were also signals present in this dataset that could likely increase performance, but which were omitted to limit the scope of the study. For example activity classification categories were provided by the PulseOn wristband and sleep, circadian and calendar features could be engineered from timestamps and ACC, in addition the application used for EMA questionnaires recorded phone activity which could be analyzed. New unobtrusive wearable sensors could also be introduced, such as skin temperature to measure circadian effects, or gyroscopes, magnetometers, barometers and GPS to provide more accurate activity detection and contextual data.

Another way to increase performance is by reformulating the approach. For example personalized or multi task learning methods could be used to capture the interpersonal variation present in affective state self-reports (Jaques et al., 2017). However, in these cases it is important to have sufficient sample sizes from each subject and to evaluate model performance against a personalized baseline in the lines of Zenonos et al. (2016). The application of personalized approaches could also help to disambiguate between within subject and between subject effects, e.g. between HRV and positive affect. A related but broader approach would be to personalize models on the level of clusters of subjects having similar physiological reactivity. Clustering is an example of unsupervised machine learning approach, which could also be utilized in this context with high potential. For example in this study the collection of EMAs practically limited the samples sizes and model performance, while unsupervised methods could also utilize the abundant amounts of unlabeled data. This approach has already been used in stress detection by Tervonen et al. (2020) achieving around 60 % accuracy using in-field data. Still another approach would be to apply semi-supervised learning making use of the limited amount of labeled data as well as the large amount of unlabeled data.

Conclusions

This thesis examined the possibility of affect detection in the context of daily life using machine learning and wearable devices. Self-reported affective states from EMAs were predicted using three weeks of PPG, EDA and ACC recordings. A signal processing pipeline to deal with movement artefacts and other issues in the data is presented, a variety of features are extracted and the performance of machine learning models predicting the self reported affective states is evaluated. Above baseline performance is achieved in predicting affect vigor with statistical significance, while improvements are also noted for affects focused and enthusiastic. Examination of SHAP values is utilized to identify low movement, low EDA level and high HRV as predictive for high activation positive affective states. These results contribute to the field of affect detection by focusing on the sparsely investigated context of daily life, providing ecologically valid estimates of performance outside the laboratory. Additionally, the usage of SHAP values to examine the relationships utilized by the trained models presents a novel and robust method for guiding feature extraction and highlighting interesting relationships for future study.

References

Accuracy of Beat-to-Beat Heart Rate Estimation Using the PulseOn Optical Heart Rate Monitor. (2018). https://pulseon.com/wp-content/uploads/2021/10/IBI-whitepaper_2018.pdf

Benedek, M. & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, *190*(1), 80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing [Publisher: Wiley Online Library]. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Bota, P. J., Wang, C., Fred, A. L. N. & Placido Da Silva, H. (2019). A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals. *IEEE Access*, *7*, 140990–141020. https://doi.org/10.1109/ACCESS.2019.2944001

Boucsein, W. (2012). *Electrodermal Activity*. Springer US. https://doi.org/10.1007/978-1-4614-1126-0

Breiman, L. (2001). Random forests [Publisher: Springer]. *Machine learning*, *45*(1), 5–32.

Cacioppo, J. T., Tassinary, L. G. & Berntson, G. (2007). *Handbook of psychophysiology*. Cambridge university press.

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y. & Li, Y. (2021). *Xgboost: Extreme Gradient Boosting*. https://CRAN.R-project.org/package=xgboost

Ciccone, A. B., Siedlik, J. A., Wecht, J. M., Deckert, J. A., Nguyen, N. D. & Weir, J. P. (2017). Reminder: RMSSD and SD1 are identical heart rate variability metrics. *Muscle Nerve*, *56*(4), 674–678. https://doi.org/10.1002/mus.25573

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., Barral, O., Kosunen, I., Valtonen, T., Huotilainen, M., Ravaja, N. & Jacucci, G. (2016). The Psychophysiology Primer: A Guide to Methods and a Broad Review with a Focus on Human–Computer Interaction. *FNT in Human–Computer Interaction*, *9*(3-4), 151–308. https://doi.org/10.1561/1100000065

Delgado-Gonzalo, R., Parak, J., Tarniceriu, A., Renevey, P., Bertschi, M. & Korhonen, I. (2015). Evaluation of accuracy and reliability of PulseOn optical heart rate monitoring device. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 430–433. https://doi.org/10.1109/EMBC.2015.7318391

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Ekkekakis, P. (2013). *The Measurement of Affect, Mood, and Emotion: A Guide for Health-Behavioral Research* [Publication Title: The Measurement of Affect, Mood, and Emotion]. Cambridge University Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3-4), 169–200. https://doi.org/10.1080/02699939208411068

Ernst, G. (2014). *Heart Rate Variability*. Springer London. https://doi.org/10.1007/978-1-4471-4309-3

Franchini, K. G. & Cowley Jr, A. W. (2004). Autonomic control of cardiac function. *Primer on the autonomic nervous system* (pp. 134–138). Elsevier.

Fridja, N. H. (2009). Mood. In D. E. Sander & K. R. Scherer (Eds.), *The Oxford companion to emotion and the affective sciences*. Oxford University Press.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, *29*(5). https://doi.org/10.1214/aos/1013203451

Godfrey, A., Conway, R., Meagher, D. & ÓLaighin, G. (2008). Direct measurement of human movement by accelerometry. *Medical Engineering & Physics*, *30*(10), 1364–1386. https://doi.org/10.1016/j.medengphy.2008.09.005

Hees, V. T. v., Fang, Z., Zhao, J. H., Heywood, J., Mirkes, E., Sabia, S. & Migueles, J. H. (2021). *GGIR: Raw Accelerometer Data Analysis.* https://doi.org/10.5281/zenodo.1051064

Heikkilä, P., Honka, A., Mach, S., Schmalfuß, F., Kaasinen, E. & Väänänen, K. (2018). Quantified Factory Worker - Expert Evaluation and Ethical Considerations of Wearable Self-tracking Devices. *Proceedings of the 22nd International Academic Mindtrek Conference*, 202–211. https://doi.org/10.1145/3275116.3275119

Jaques, N., Taylor, S., Nosakhare, E., Sano, A. & Picard, R. (2016). Multi-task learning for predicting health, stress, and happiness. *NIPS Workshop on Machine Learning for Healthcare.*

Jaques, N., Taylor, S., Sano, A., Picard, R. et al. (2017). Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. *IJCAI 2017 Workshop on artificial intelligence in affective computing*, 17–33.

Jeppesen, J., Beniczky, S., Johansen, P., Sidenius, P. & Fuglsang-Frederiksen, A. (2014). Using Lorenz plot and Cardiac Sympathetic Index of heart rate variability for detecting seizures for patients with epilepsy. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 4563–4566. https://doi.org/10.1109/EMBC.2014.6944639

Jussila, J., Venho, N., Salonius, H., Moilanen, J., Liukkonen, J. & Rinnetmäki, M. (2018). Towards ecosystem for research and development of electrodermal activity applications. *Proceedings of the 22nd International Academic Mindtrek Conference*, 79–87. https://doi.org/10.1145/3275116.3275141

Komulainen, E., Meskanen, K., Lipsanen, J., Lahti, J. M., Jylhä, P., Melartin, T., Wichers, M., Isometsä, E. & Ekelund, J. (2014). The Effect of Personality on

Daily Life Emotional Processes (J. Yuan, Ed.). *PLoS ONE*, *9*(10), e110907. https://doi.org/10.1371/journal.pone.0110907

Krupić, D., Žuro, B. & Corr, P. J. (2021). Anxiety and threat magnification in subjective and physiological responses of fear of heights induced by virtual reality. *Personality and Individual Differences*, *169*, 109720. https://doi.org/10.1016/j.paid.2019.109720

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L. & Bischl, B. (2019). Mlr3: A modern object-oriented machine learning framework in R. *JOSS*, *4*(44), 1903. https://doi.org/10.21105/joss.01903

Levenson, R. W. (2014). The Autonomic Nervous System and Emotion. *Emotion Review*, *6*(2), 100–112. https://doi.org/10.1177/1754073913512003

Lundberg, S. M., Erion, G. G. & Lee, S.-I. (2019). Consistent Individualized Feature Attribution for Tree Ensembles [arXiv: 1802.03888]. *arXiv:1802.03888 [cs, stat]*. Retrieved March 5, 2022, from http://arxiv.org/abs/1802.03888

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Määttänen, I., Henttonen, P., Väliaho, J., Palomäki, J., Thibault, M., Kallio, J., Mäntyjärvi, J., Harviainen, T. & Jokela, M. (2021). Positive affect state is a good predictor of movement and stress: Combining data from ESM/EMA, mobile HRV measurements and trait questionnaires. *Heliyon*, *7*(2), e06243. https://doi.org/10.1016/j.heliyon.2021.e06243

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book

Nicodemus, K. K., Malley, J. D., Strobl, C. & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*(1), 110. https://doi.org/10.1186/1471-2105-11-110

Pankaj, Kumar, A., Komaragiri, R. & Kumar, M. (2021). A Review on Computation Methods Used in Photoplethysmography Signal Analysis for Heart Rate Estimation. *Arch Computat Methods Eng.* https://doi.org/10.1007/s11831-021-09597-4

Papousek, I., Nauschnegg, K., Paechter, M., Lackner, H. K., Goswami, N. & Schulter, G. (2010). Trait and state positive affect and cardiovascular recovery from experimental academic stress. *Biological Psychology*, *83*(2), 108–115. https://doi.org/10.1016/j.biopsycho.2009.11.008

Parak, J., Tarniceriu, A., Renevey, P., Bertschi, M., Delgado-Gonzalo, R. & Korhonen, I. (2015). Evaluation of the beat-to-beat detection accuracy of PulseOn wearable optical heart rate monitor. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 8099–8102. https://doi.org/10.1109/EMBC.2015.7320273

Pettersson, K., Tervonen, J., Narvainen, J., Henttonen, P., Maattanen, I. & Mantyjarvi, J. (2020). Selecting Feature Sets and Comparing Classification Methods for Cognitive State Estimation. *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 683–690. https://doi.org/10.1109/BIBE50027.2020.00115

Picard, R. W. (2000). *Affective computing.*

R Core Team. (2022). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Roseman, I. J., Wiest, C. & Swartz, T. S. (1994). Phenomenology, behaviors, and goals differentiate discrete emotions. [Publisher: American Psychological Association]. *Journal of personality and social psychology*, *67*(2), 206.

Russell, J. A. (1980). A circumplex model of affect. [Publisher: American Psychological Association]. *Journal of personality and social psychology*, *39*(6), 1161.

Russell, J. A. (1991). In defense of a prototype approach to emotion concepts. [Publisher: American Psychological Association]. *Journal of personality and social psychology*, *60*(1), 37.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145

Russell, J. A. (2005). Emotion in human consciousness is built on core affect [Publisher: Imprint Academic]. *Journal of consciousness studies*, *12*(8-9), 26–42.

Russell, J. A. & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. [Publisher: American Psychological Association]. *Journal of personality and social psychology*, *76*(5), 805.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. https://doi.org/10.1177/0539018405058216

Schmidt, P., Reiss, A., Dürichen, R. & Laerhoven, K. V. (2019). Wearable-Based Affect Recognition—A Review. *Sensors*, *19*(19), 4079. https://doi.org/10.3390/s19194079

Schwerdtfeger, A. R. & Gerteis, A. K. S. (2014). The manifold effects of positive affect on heart rate variability in everyday life: Distinguishing within-person and between-person associations. *Health Psychology*, *33*(9), 1065–1073. https://doi.org/10.1037/hea0000079

Shaffer, F. & Ginsberg, J. P. (2017). An Overview of Heart Rate Variability Metrics and Norms. *Front. Public Health*, *5*, 258. https://doi.org/10.3389/fpubh.2017.00258

Shiffman, S., Stone, A. A. & Hufford, M. R. (2008). Ecological Momentary Assessment. *Annu. Rev. Clin. Psychol.*, *4*(1), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Tamura, T., Maeda, Y., Sekine, M. & Yoshida, M. (2014). Wearable Photoplethysmographic Sensors—Past and Present. *Electronics*, *3*(2), 282–302. https://doi.org/10.3390/electronics3020282

Taylor, S., Jaques, N., Nosakhare, E., Sano, A. & Picard, R. (2020). Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Trans. Affective Comput.*, *11*(2), 200–213. https://doi.org/10.1109/TAFFC.2017.2784832

Tervonen, J., Puttonen, S., Sillanpää, M. J., Hopsu, L., Homorodi, Z., Keränen, J., Pajukanta, J., Tolonen, A., Lämsä, A. & Mäntyjärvi, J. (2020). Personalized mental stress detection with self-organizing map: From laboratory to the field. *Computers in Biology and Medicine*, *124*, 103935. https://doi.org/10.1016/j.compbiomed.2020.103935

Tohka, J. & van Gils, M. (2021). Evaluation of machine learning algorithms for health and wellness applications: A tutorial. *Computers in Biology and Medicine*, *132*, 104324. https://doi.org/10.1016/j.compbiomed.2021.104324

Torniainen, J., Cowley, B., Henelius, A., Lukander, K. & Pakarinen, S. (2015). Feasibility of an electrodermal activity ring prototype as a research tool [ISSN: 1558-4615]. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6433–6436. https://doi.org/10.1109/EMBC.2015.7319865

van Hees, V. T., Fang, Z., Langford, J., Assah, F., Mohammad, A., da Silva, I. C. M., Trenell, M. I., White, T., Wareham, N. J. & Brage, S. (2014). Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: An evaluation on four continents. *Journal of Applied Physiology*, *117*(7), 738–744. https://doi.org/10.1152/japplphysiol.00421.2014

van Hees, V. T., Gorzelniak, L., Dean León, E. C., Eder, M., Pias, M., Taherian, S., Ekelund, U., Renström, F., Franks, P. W., Horsch, A. & Brage, S. (2013). Separating Movement and Gravity Components in an Acceleration Signal and Implications for the Assessment of Human Daily Physical Activity (M. Müller, Ed.). *PLoS ONE*, *8*(4), e61691. https://doi.org/10.1371/journal.pone.0061691

Vinik, A. I. (2012). The Conductor of the Autonomic Orchestra. *Front. Endocrin.*, *3*. https://doi.org/10.3389/fendo.2012.00071

Watson, D., Clark, L. A. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. [Publisher: American Psychological Association]. *Journal of personality and social psychology*, *54*(6), 1063.

Watson, D., Wiese, D., Vaidya, J. & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. [Publisher: American Psychological Association]. *Journal of personality and social psychology*, *76*(5), 820.

Wright, M. N. & Ziegler, A. (2017). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Yasuma, F. & Hayano, J.-i. (2004). Respiratory Sinus Arrhythmia. *Chest*, *125*(2), 683–690. https://doi.org/10.1378/chest.125.2.683

Zelenski, J. M. & Larsen, R. J. (2000). The Distribution of Basic Emotions in Everyday Life: A State and Trait Perspective from Experience Sampling Data. *Journal of Research in Personality*, *34*(2), 178–197. https://doi.org/10.1006/jrpe.1999.2275

Zenonos, A., Khan, A., Kalogridis, G., Vatsikas, S., Lewis, T. & Sooriyabandara, M. (2016). HealthyOffice: Mood recognition at work using smartphones and wearable sensors. *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 1–6. https://doi.org/10.1109/PERCOMW.2016.7457166

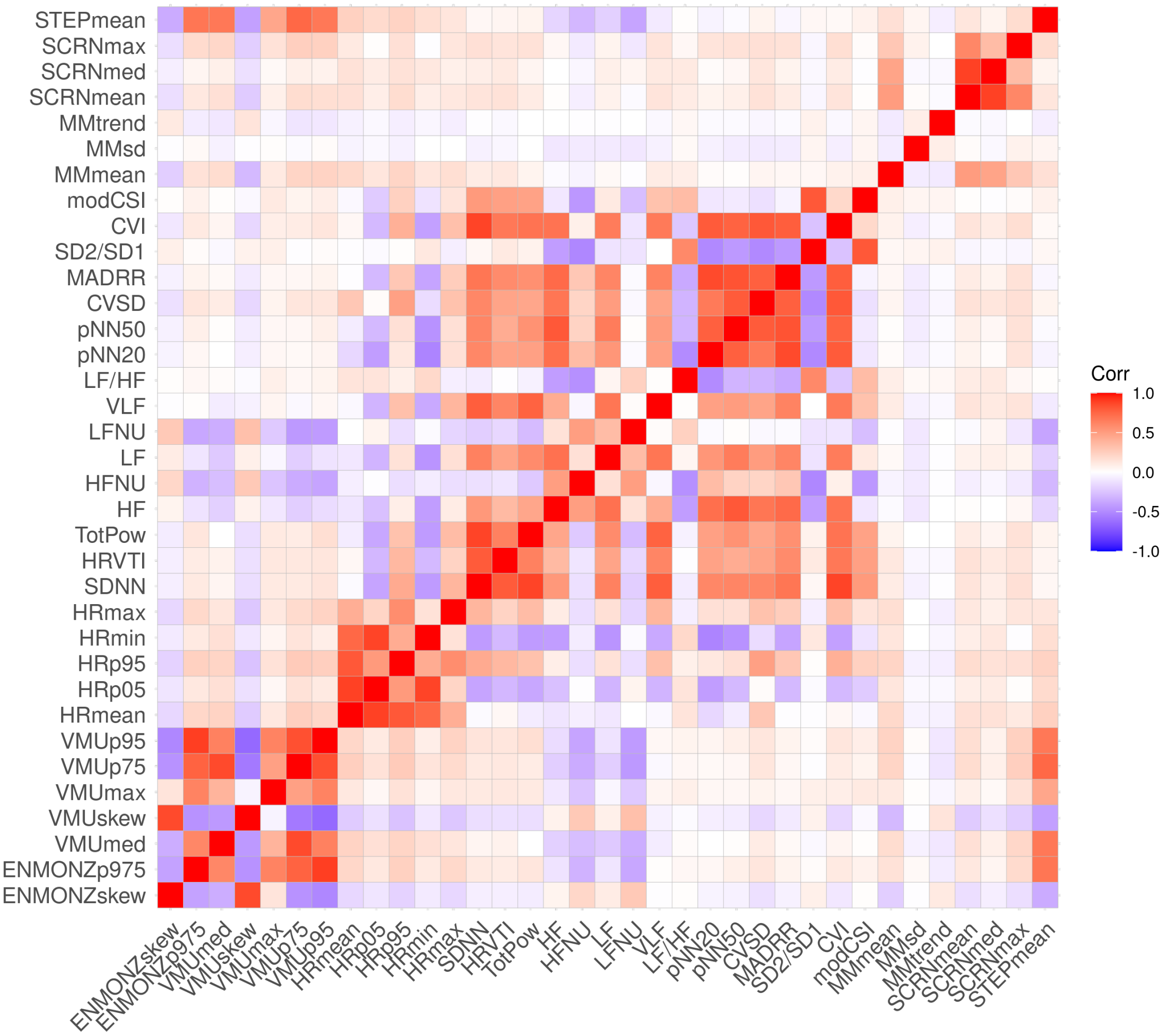
Zhu, Z., Satizabal, H. F., Blanke, U., Perez-Uribe, A. & Troster, G. (2016). Naturalistic Recognition of Activities and Mood Using Wearable Electronics. *IEEE Trans. Affective Comput.*, *7*(3), 272–285. https://doi.org/10.1109/TAFFC.2015.2491927

Appendix A

Feature correlation matrix

Figure A1

Final dataset feature correlations. Note the four primary signal types clustering along the identity line: ACC, HR, HRV and EDA.



Appendix B

Model predictions

The following figures visualize boxplots of predicted values for RF and XGBoost models for affects angry, anxious, enthusiastic, focused, happy, sad, satisfied and vigor. Number of observations belonging to each true value has been marked above x-axis tickmarks.

Figure B1

Boxplots for RF predicted values.

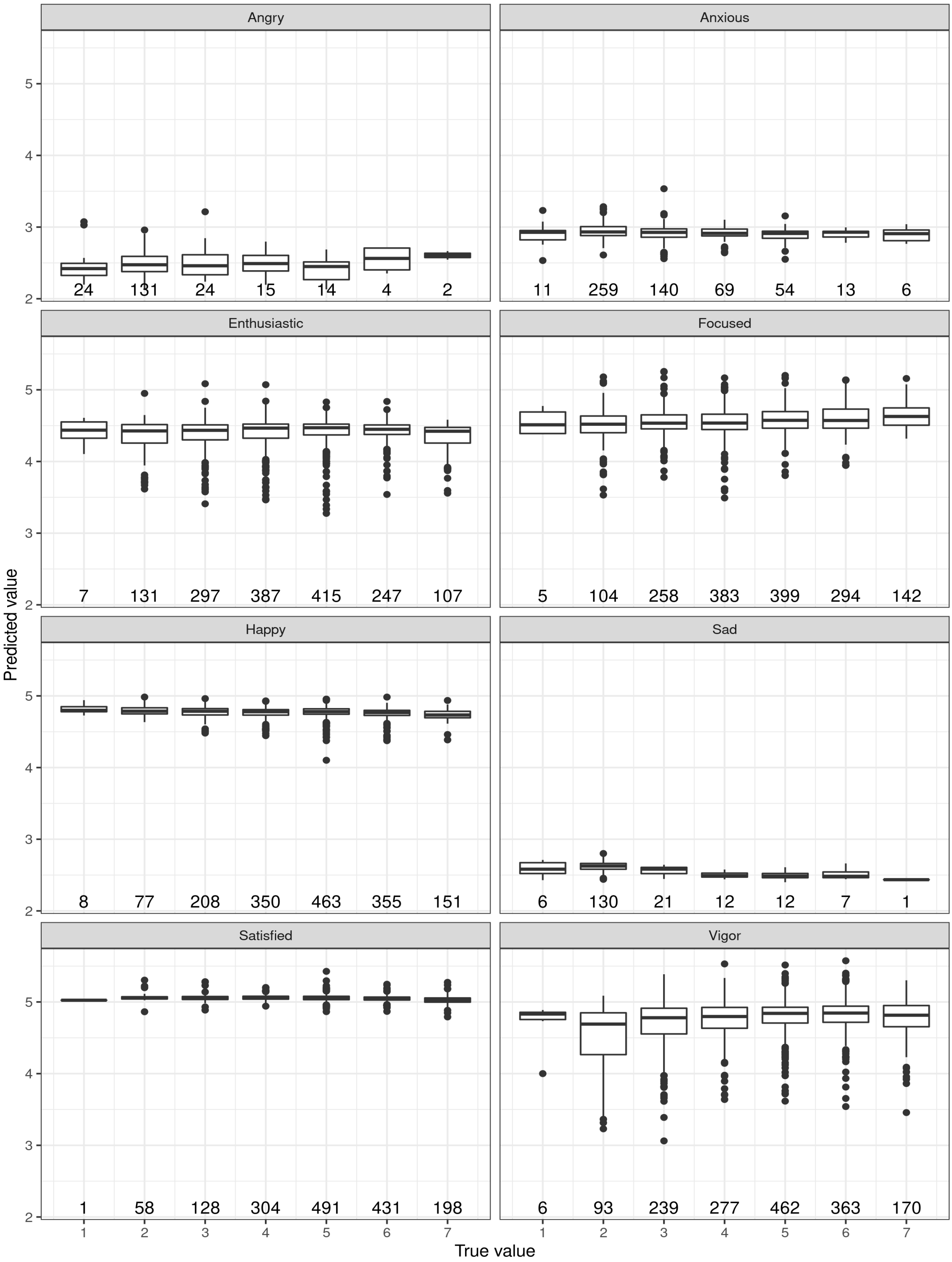
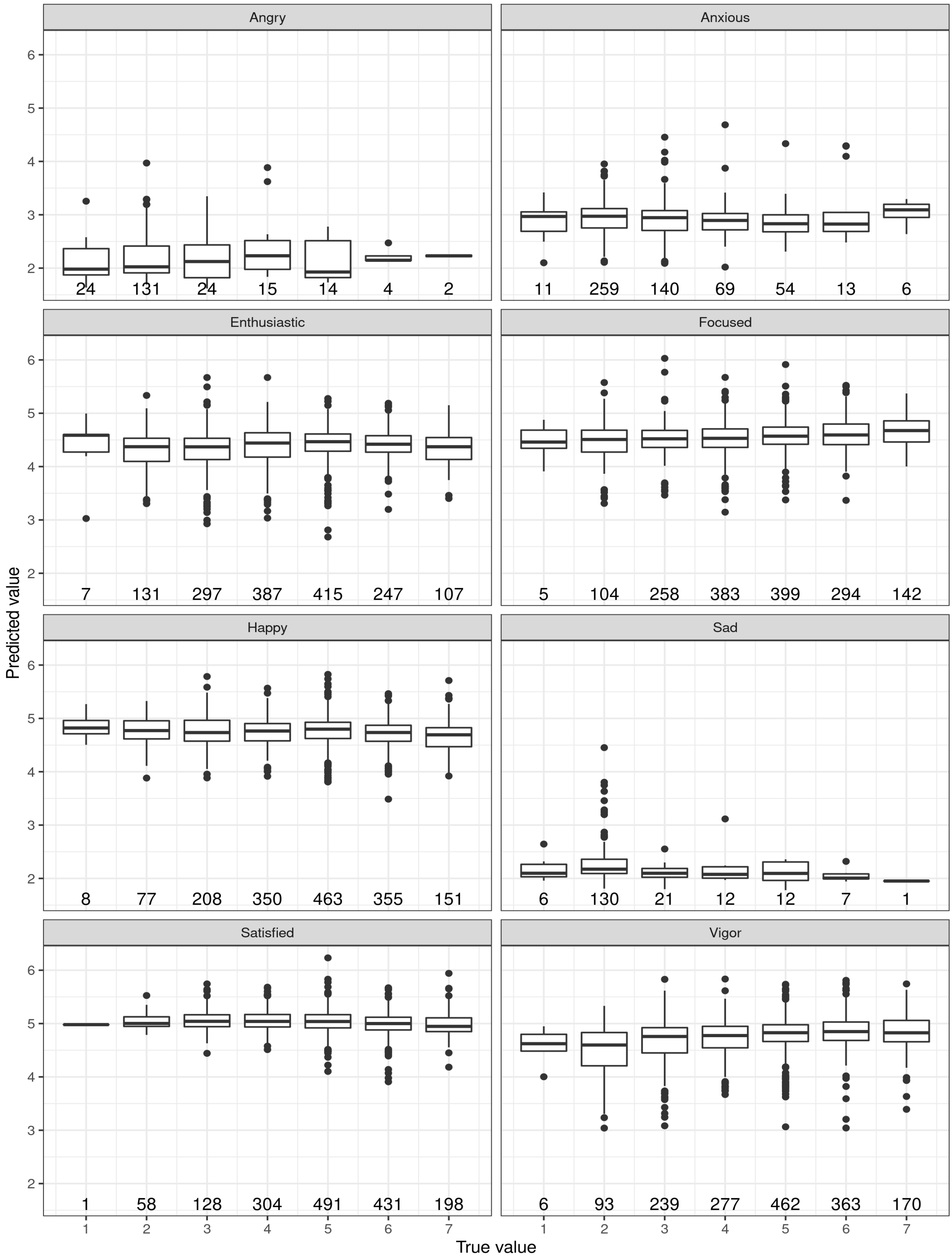


Figure B2

Boxplots for XGBoost predicted values.



Appendix C

SHAP summary plots

The following visualizations detail complete SHAP summary plots for XGBoost models for affects angry, anxious, enthusiastic, focused, happy, sad, satisfied and vigor. Point placement on the x-axis measures SHAP value magnitude (impact on model output) for a single prediction. Features have been ranked on the y-axis according to the mean absolute SHAP value, marked on the right side of feature identifiers. Normalized feature values are represented on a colour scale ranging from z-score of 2 and above (dark blue) to -2 and below (bright yellow). SHAP values have been jittered on the y-axis based on the density of values.

Figure C1

SHAP summary plot for affect angry.

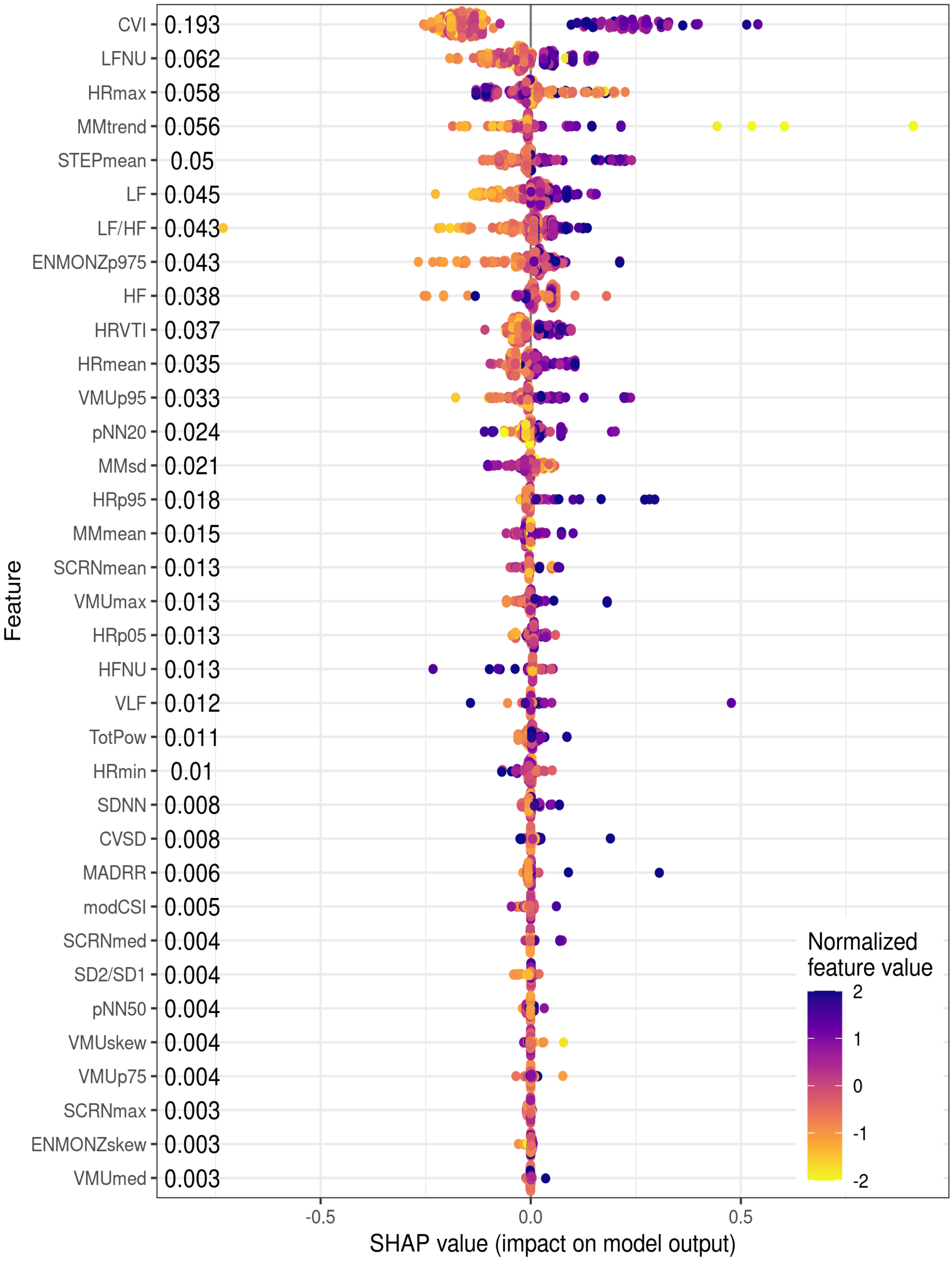


Figure C2

SHAP summary plot for affect anxious.

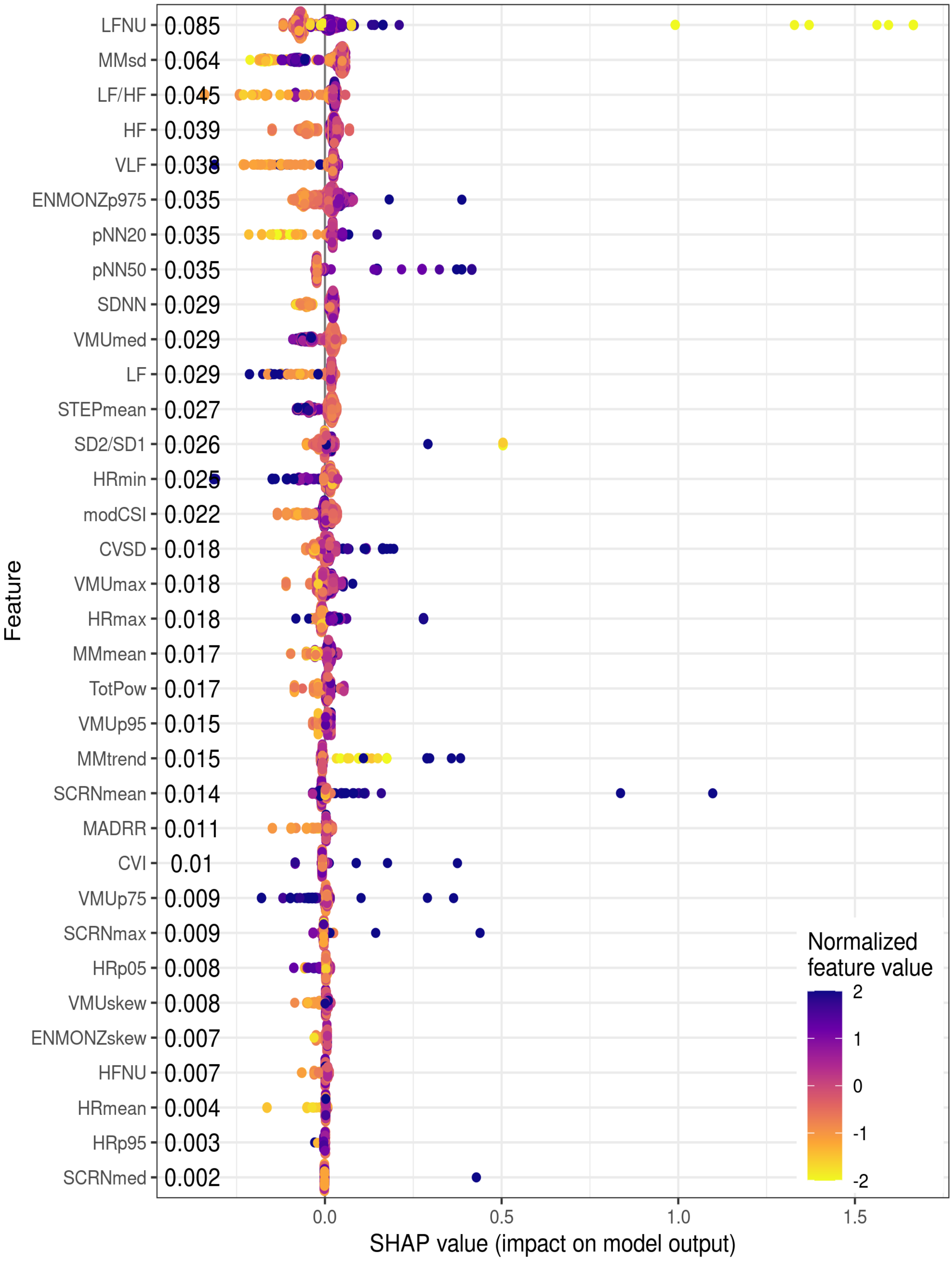


Figure C3

SHAP summary plot for affect enthusiastic.

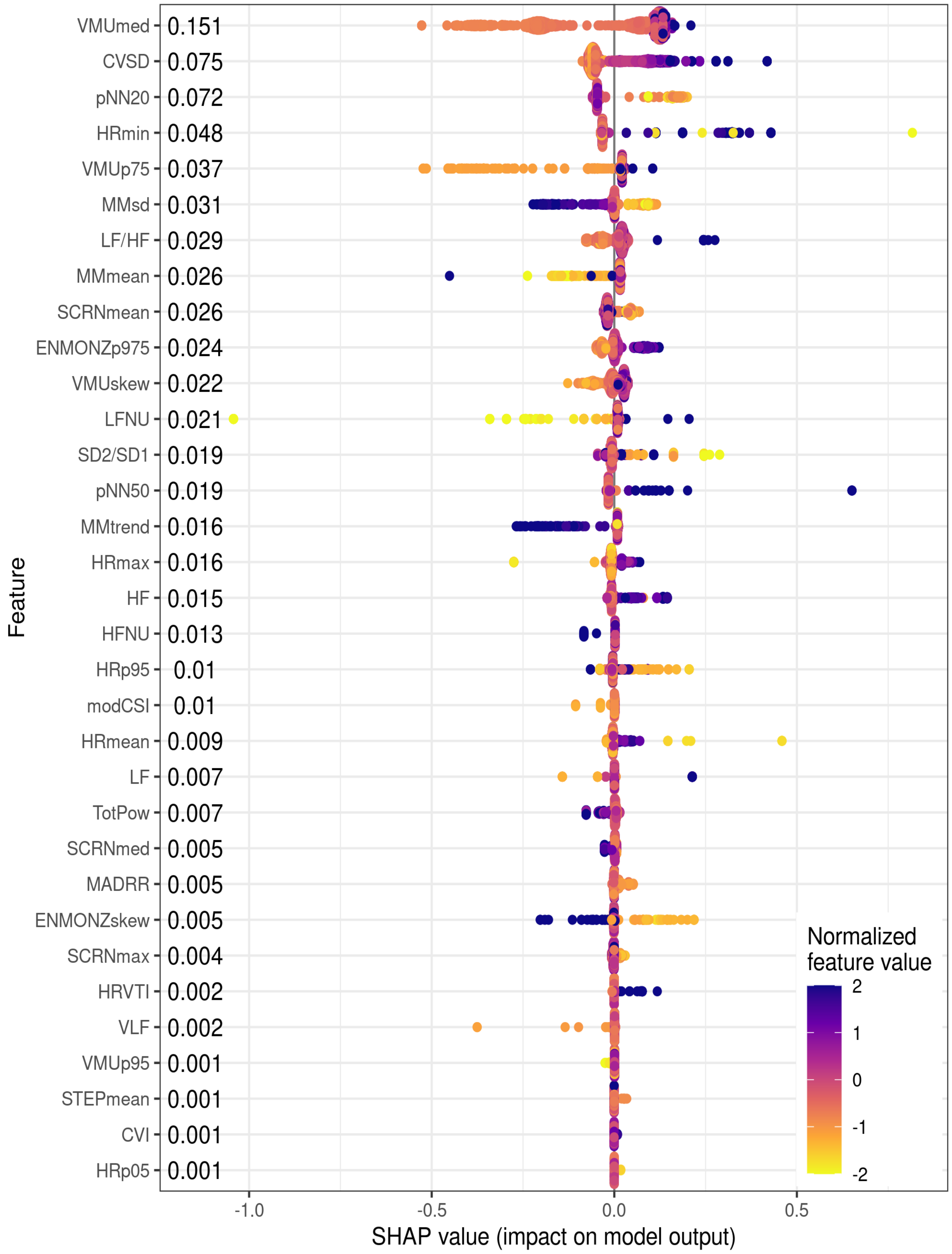


Figure C4

SHAP summary plot for affect focused.

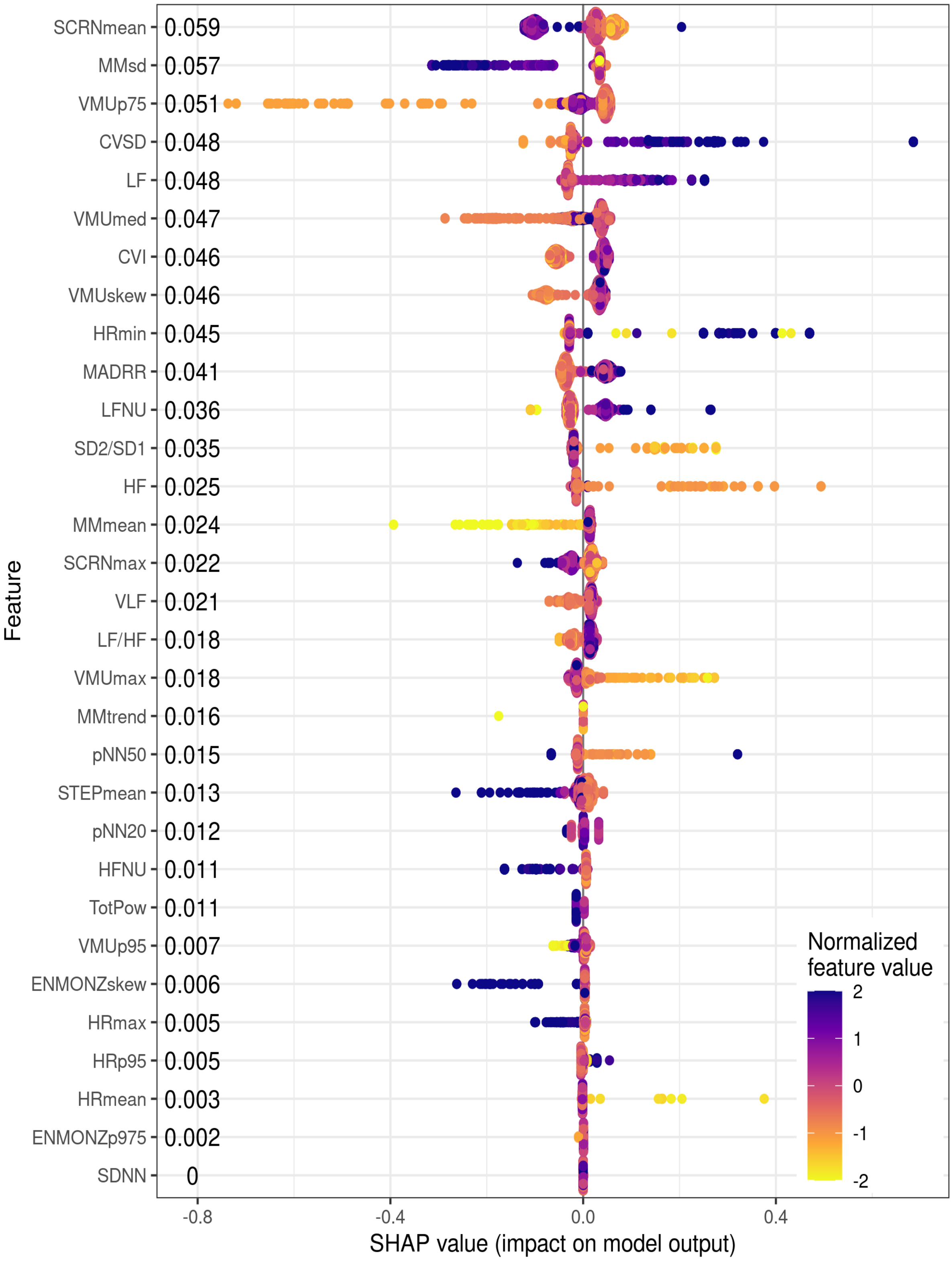


Figure C5

SHAP summary plot for affect happy.

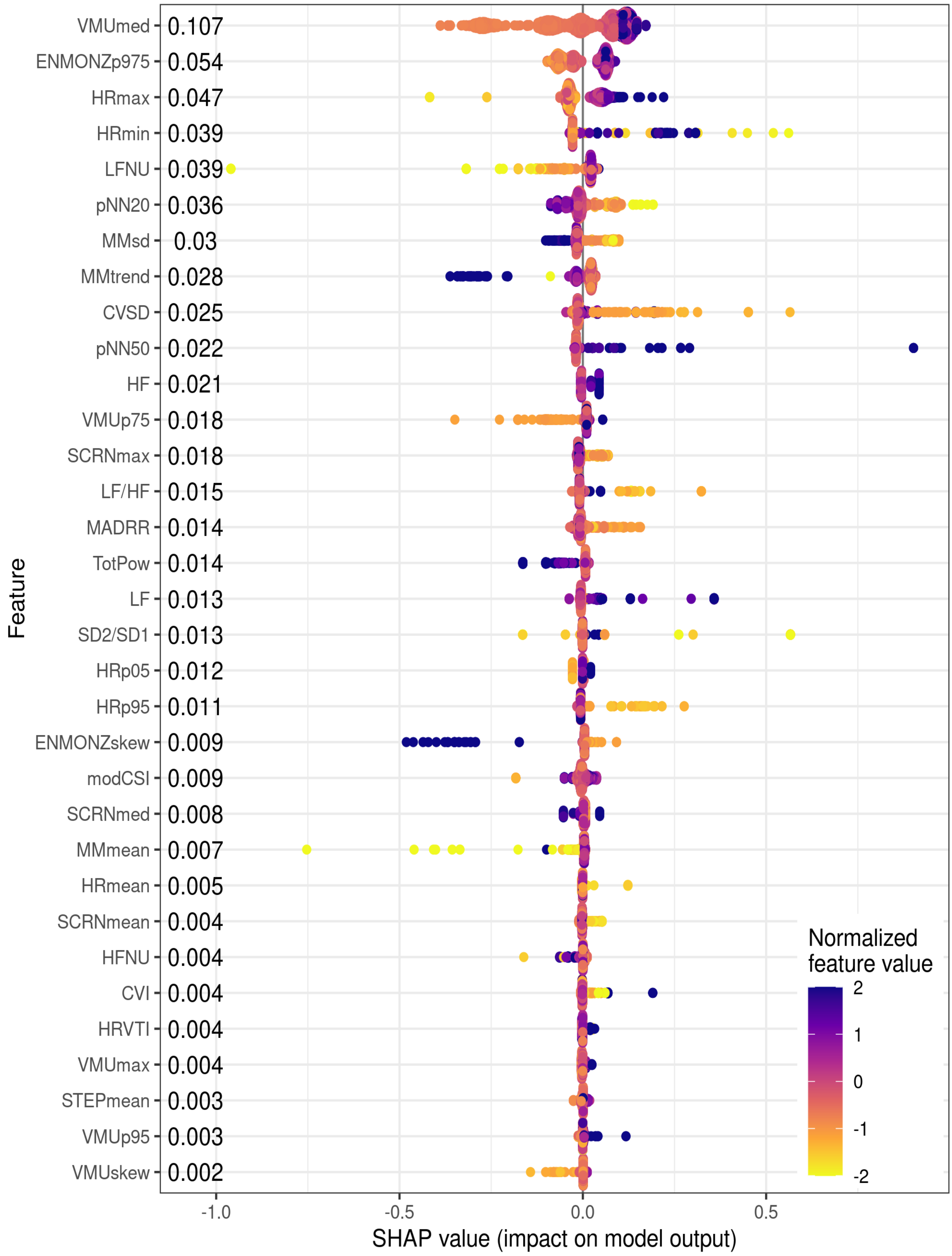


Figure C6

SHAP summary plot for affect sad.

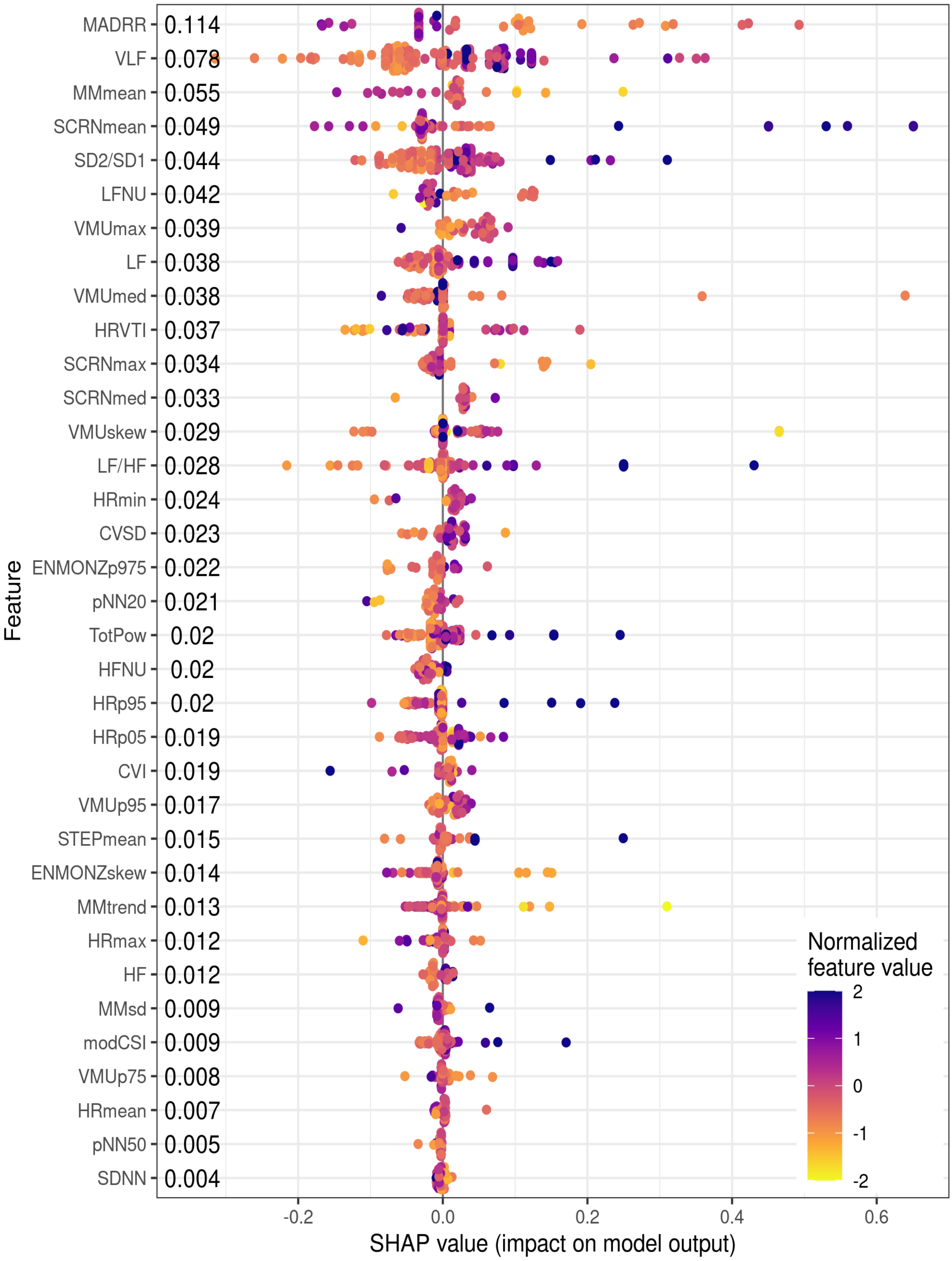


Figure C7

SHAP summary plot for affect satisfied.

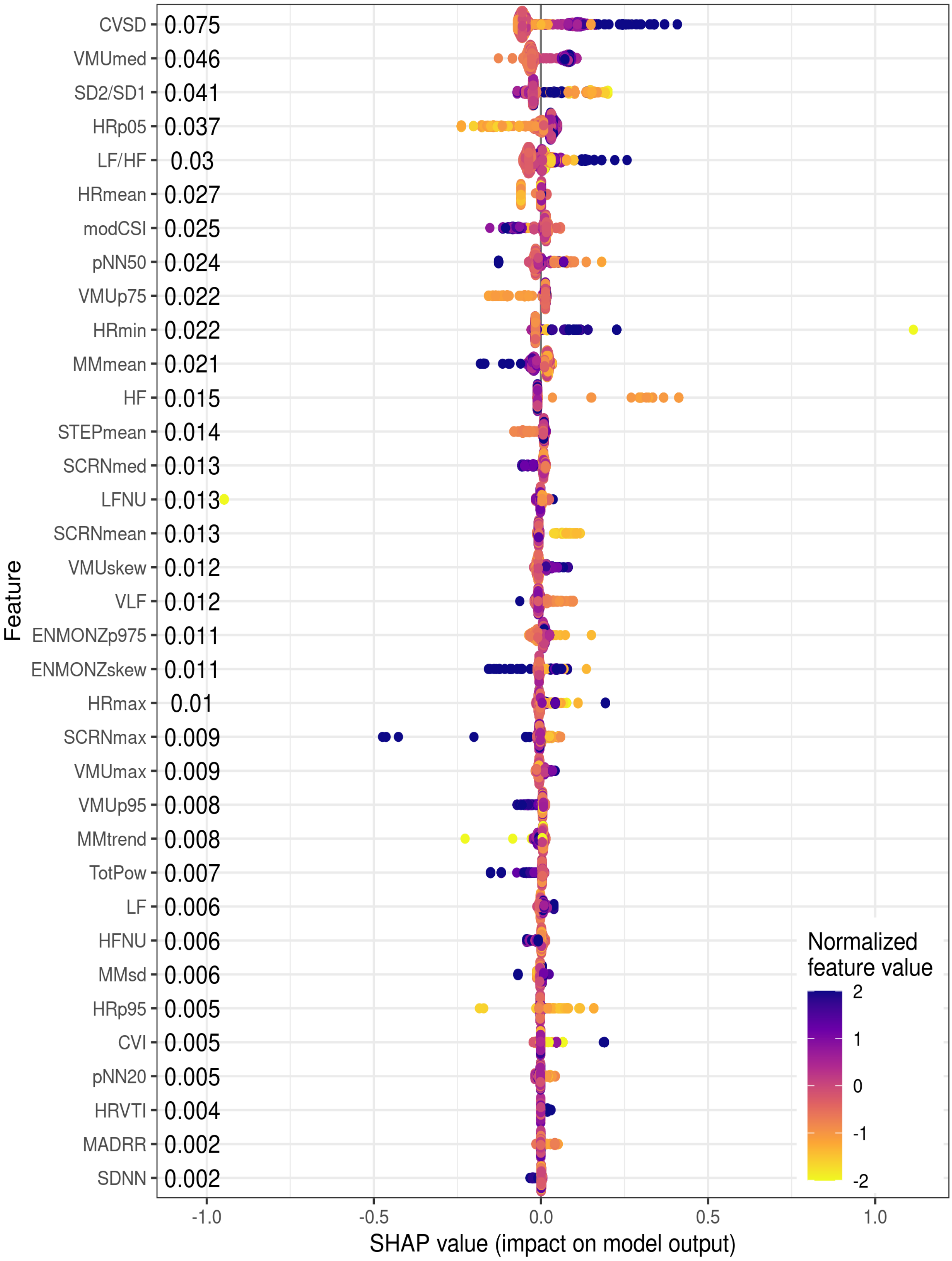
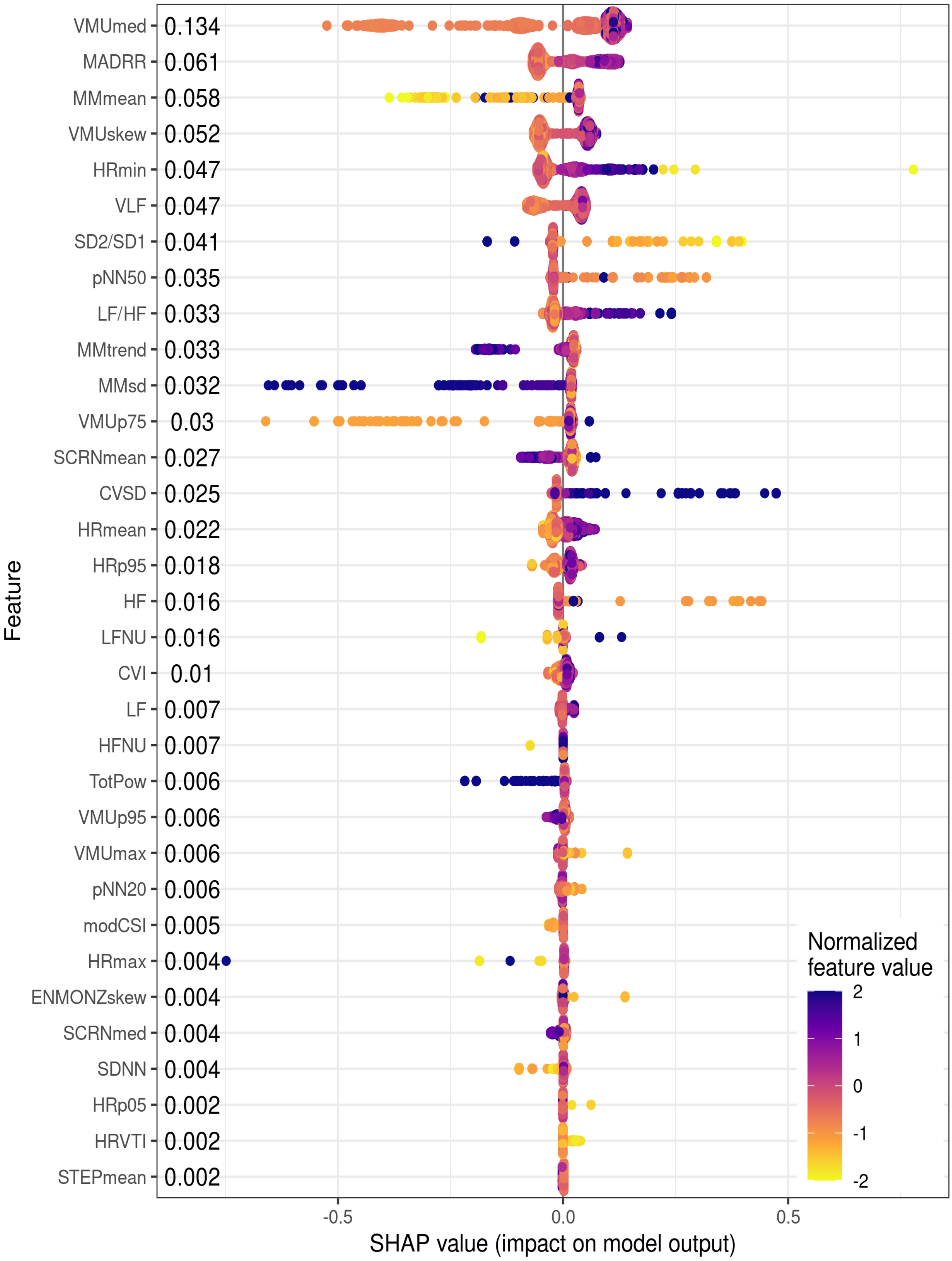


Figure C8

SHAP summary plot for affect vigor.


Feature
VMUmed 0.134
MADRR 0.061
MMmean 0.058
VMUskew 0.052
HRmin 0.047
VLF 0.047
SD2/SD1 0.041
pNN50 0.035
LF/HF 0.033
MMtrend 0.033
MMsd 0.032
VMUp75 0.03
SCRNmean 0.027
CVSD 0.025
HRmean 0.022
HRp95 0.018
HF 0.016
LFNU 0.016
CVI 0.01
LF 0.007
HFNU 0.007
TotPow 0.006
VMUp95 0.006
VMUmax 0.006
pNN20 0.006
modCSI 0.005
HRmax 0.004
ENMONZskew 0.004
SCRNmed 0.004
SDNN 0.004
HRp05 0.002
HRVTI 0.002
STEPmean 0.002
SHAP value (impact on model output)
Normalized feature value