ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume VI-4/W2-2020, 2020
5th International Conference on Smart Data and Smart Cities, 30 September – 2 October 2020, Nice, France

"I KNOW HOW YOU FEEL" – PREDICTING EMOTIONS FROM SENSORS FOR ASSISTED PEDELEC EXPERIENCES IN SMART CITIES

S. Schneider*, H. Dastageeri, P. Rodrigues, V. Coors

University of Applied Sciences Stuttgart
Schellingstraße 24
70174 Stuttgart, Germany
(sven.schneider, habiburrahman.dastageeri, preston.rodrigues, volker.coors) @hft-stuttgart.de

Commission IV

KEY WORDS: experience sampling, mobile sensors, urban emotions, machine learning

ABSTRACT:

Emotions are one of the manner humans use to indicate how they feel about a particular event, place or things. To date there is no consensus about the correlation of measured data to an unambiguously defined emotional state. The selection of parameters, their weight and range, which derive at an emotion, are not clearly defined. Especially, if measurements took place outdoors and during a physical activity. This work is based on previous work and focuses on the parameters and methods to classify measured data to an emotional state. We took a closer look to the values, defined ranges for parameters and performed further pre-processing steps. Furthermore, we revised the assignment of an emotion, analyzed the parameter weights and their correlation. Moreover, we compared our previous approach with further Machine Learning (ML) methods. The results are in line with previous work, however, indicate the need for more and heterogeneous data to endorse the outcome. Further results from the parameter analysis suggest an importance of the skin conductance level (SCL) depending on the method used.

1. INTRODUCTION

The attractiveness of a city depends on many components. City history, green spaces and infrastructure are some relevant aspects to name a few. However, these aspects vary strongly and depend on an individual's point of view. Leading eventually to one key factor namely *feelings*. A person's emotion for a location defines the personal relationship to a place, area, city and even country. Increasing the attractiveness of a city is generally aimed by enriching public life and enhancing urban spaces. But these endeavors are usually based on presumptions. There are only few proper ways of measuring emotions on a personal level, especially if the aim is to find locations in need of improvements or to affirm carried out improvement strategies.

So what is a good way to explore a city widely under the aforementioned aspects and from different vantage points, for example, by different users and user groups? An optimal solution would be to integrate these observations without actually asking[1] a user to actively participating in providing feedback but to *passively* measure the state of emotions of citizens. With *smart wearables* such as *smart watches*, *fitness trackers* and *smart phones*, there is already a wide range of user accepted sensory information available. Ways of commuting through a city are also diverse and let citizens experience a city on different scales or level of details. Classical means of commuting range from *high-level*[2] means of transport such as buses, trams and light rail systems to *low-level* means of transportation such as cycling, using e-scooters and walking. In this study, we focus on a particular *low-level* mode of transport, namely cycling.

[1] of course consent of a user to use their data anonymously is mandatory.
[2] note: these terms are introduced in the paper for abstraction of different levels of traveling and do not refer to some kind of standard.

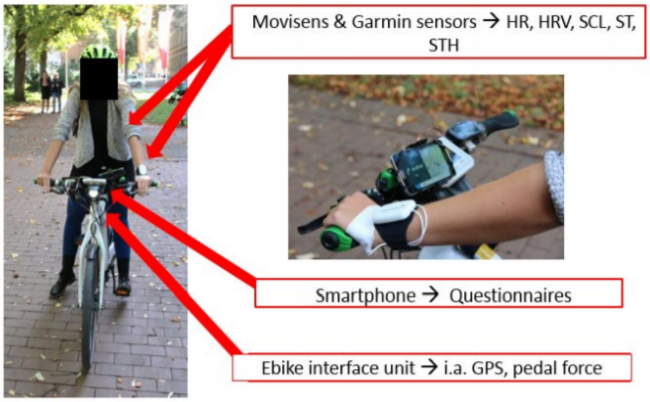

Figure 1. Ebike and proband with measurement instruments and sensor (Double-Blind-RiviewREF, 2019).

This is because commuting with a bike and in particular a *pedelec*, a relatively large area can be explored in very little time. Although the commute in some cases may be brief, the level of detail a user can experience is relatively high. Furthermore, our *pedelec* (Figure 1) is equipped with a series of additional sensors required for our study.

Furthermore, pedelecs as a means to foster a low-carbon mobility have versatile advantages. The physical exercise for instance is one main advantage that also supports well-being. According to the Healthy Cities Vision of World Health Organization (WHO), health and well-being are the basic factors for urban health. A healthy city ensures that health and well-being of both the people and the planet are at the heart of the city's internal and external policies. In this context, one of our visions is to provide citizens with a *pedelec* sharing system that can guide a user based on their physical and psychological (emotional)

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-VI-4-W2-2020-149-2020 | © Authors 2020. CC BY 4.0 License.

condition. For example, a given scenario may guide the user towards a longer but less steep route according to the fitness level (e.g. determined by the heart rate) and the remaining battery charge of the *pedelec*. Another scenario, closely linked to this study, is to guide the user along routes which reduce stress by avoiding high traffic routes. Furthermore, by detection emotions during a user's ride, places of (dis-) comfort can be identified. This kind of information may aid officials at city administrations to reproduce places of comfort and improve places of discomfort (e.g. by taking measures to reduce traffic or planting trees).

Our aim is to recognize emotions on a personal level using sensors in an outdoor environment and during a physical activity. This work is based on previous work published in (Kohn et al., 2018, Dastageeri et al., 2019). Building up on the results, this study focuses on improving the methods of emotion recognition and increasing the data quality by an additional preprocessing step.

2. MEASURES

One key aspect is to select and assess the parameters that are necessary to provide our aimed output, namely an emotion. For data collection the Department of Business Psychology, HFT - University of Applied Sciences Stuttgart stated an experiment for pedelec drivers on a given route in Stuttgart (Figure 1). For further details of the experiment and to reason the selection of the used parameters, we want to refer the reader to our previous paper (Dastageeri et al., 2019). In this work, we focus on biosensor data, its relevance and weight as well as the interaction between parameters (Section 4.1).

2.1 Heart Rate Variability

The HRV is a crucial indicator for defining the well-being and stress resistance of the body. A healthy heart does not beat evenly (Arbeitsmedizin, Forschung, 2014). There is a variation of time from one heartbeat to the next called HRV. This effect results from the interaction between the sympathetic nervous system (SNS) and the parasympathetic nervous system (PSNS). In times of a physical or mental stress, the SNS regulates the physical organs according to the dangerous situation and increases the capability of physical action. Stress hormones were released, the heart rate and blood pressure raises. That way the body is able to react faster. This situation is called fight or flight. The PSNS fulfils the contrary function. After the stressful situation, the body needs to rest and recover. The PSNS decreases the heart rate and returns to its standard value using the body's own resources. This reaction is called *rest and digest* (Jelinek et al., 2017, Hoffman, 2020).

Contrary to the heart rate the HRV is low in stressful situations and high during the resting phase. Physical or mental stressors can cause the stressful situation. Therefore, the HRV allows to draw conclusions about pleasant and unpleasant situations. The requirement to analyse HRV accordingly is a generally healthy and fit person. Genes define 30% of the general HRV level. But live style defines the remaining 70%. Therefore during preprocessing the gender, height, weight and age of the probands were taken into consideration (Hoffman, 2020). The HRV is not tangible but can be measured by detecting the time from one wave to the next. There are different methods of detection. The most common way of commercially available wearables is using photoplethysmogram (PPG).

2.2 Heart Rate

The heart rate of a healthy adult is between 60 until 80 beats per minute (Martens, 2018). Influencing factors are among other things age, level of fitness, emotions, temperature as well as physical and mental stress (Lexikon der Biologie, Herzfrequenz, 1999). Any form of stress immediately raises the heart rate. There are several calculation methods to define the maximum heart rate. The common calculation formula by Haskell and Fox is dated early 1970s (Kolata, 2001):

$$HR_{maximum} = 220 - age \tag{1}$$

Tough 2001 Tanaka, Monahan and Seals developed a formula based on 18.000 probands, which reached better results and was used in this approach (Tanaka et al., 2001):

$$HR_{maximum} = 208 - (Age * 0,7) \tag{2}$$

The heart rate is recorded using an optical recording mostly on the ear or the wrist. The time is measured in milliseconds. The most common used time parameters are minimum heart rate, maximum heart rate, total of heart rate in 24 hours, SDNN, SDANN, RMSSD or pNN50 (Autonom Health, 2017). Our approach uses RMSSD by applying the Movisens EcgMove 3 on a chest strap.

2.3 Skin Temperature

The body temperature, which usually refers to the body core temperature varies and depends on many aspects. Amongst other things age, gender, menstrual cycle, weight and natural rhythms affect the body temperature. During 24 hours, the lowest level is around 4 a.m. and the highest between 4 and 6 p.m., assuming the person sleeps at night and is awake during the day (Mackowiak, 1992). For adults there is a wide range, which is considered a normal temperature and varies between 33.2–38.2 °C (Sund-Levander et al., 2002). However, the skin temperature (ST) differs from the body core temperature. Besides the above mention aspects, ST varies and depends on each body part. Furthermore, the temperature of the environment, weather and clothing has a higher impact. Usually the normal ST of the trunk varies between 33.5 – 36.9°C (Bierman, 1936). Over the protruding parts, the ST is lower. After prolonged exposure to the environment, ST may also be as low as 14°C (Benedict et al., 1919).

In our approach, the ST is measured on two body parts. First on the chest beneath the clothes using a chest belt. Second on the wrist with a smart watch. For verifying, the data and finding outliers it is needed to consider that the measurements took place during a physical activity with various levels of difficulties. In addition, the probands faced headwind, which has a high impact on ST. The measured data coincide with assumed presumption. The overall minimum ST on the chest was 19 °C, the maximum 36°C, on the wrist the minimum was 13.4 °C, the maximum was 29.3 °C.

2.4 Skin conductance level (SCL)

Stimulation such as stress or emotions results to an increased production of *eccrine* sweat controlled by the SNS. Therefore,

sweat cannot be controlled consciously and is predestined to be a good objective source of information to find out if someone is stressed or triggered emotionally. However, it is important where sweat secretion takes place. Usually hands and feet are also triggered whenever someone is emotionally stimulated. Especially on the palm, sweat responds to psychological stimuli (Edelberg, 1972). This reaction can be measured by two electrodes placed on the palm. By applying a constant voltage of 0.5 volt, the resulting conductance varies which can be measured (Benedek, Kaernbach, 2010). This value is named Skin Conductance Level (SCL), which is also referred to as Galvanic Skin Response. SCL is measured in μS (microsiemens) and has a range of typically 2 – 20 μS. Tough, the SCL value differs strongly from person to person. A comparison of SCL alone is not significant. In addition, the SCL depends highly on the surrounding temperature as well as the physical activity of the proband. It is recommended to measure indoors with a temperature of 23 °Celsius and a constant humidity (Boucsein, 2012).

In our approach, the proband is performing a bicycle ride outdoors, which required various levels of effort. The temperature fluctuated and the proband faced headwind. To counter that challenging task our approach also took into consideration the skin surface temperature on the wrist and under clothes on the chest, the heart rate and heart rate variability. Our aim was to define the relevance of SCL in combination with these parameters and possibly correlations. In addition, the calculation was done with and without considering SCL to estimate its effect under these circumstances.

3. ARTIFICIAL INTELLIGENCE FRAMEWORKS

The use of data science in different discipline has resulted in extraction of knowledge and unseen patterns from data in that domain. However, it is cumbersome to manage domain specific data. The complexity further increases when we need to integrate and analyze large volume of domain specific heterogeneous data in real-time. To address these issues, we are analyzing two frameworks. The first framework is Apache Spark (spark.apache.org). Apache Spark is an agnostic data processing engine. This framework has a wide spread use in academia and was also used in the authors previous work (Dastageeri et al., 2019). The second framework is a Gaussian Processes (GPs). It operates in a fully Bayesian manner. The GP algorithm was implemented on a local machine using standard *Matlab2018b* software. However, integration into the Spark framework may be viable.

3.1 Neural Network using Spark

Apache Spark is an agnostic data analytics engine that is used for large-scale data analytic and processing. The core engine consists of four main libraries(`SQL`, `MLlib` for machine learning, `GraphX`, and `Spark Streaming`) each with a specific functionality. These libraries can be integrated seamlessly in an application to develop a fully distriduted solution. For this work, we are using `MLlib` and `Spark streaming` library. The main reason to use `MLlib` is because it supports *ML Pipelines*. *ML Pipelines* helps to compose multiple algorithms into a single *pipeline*. A *Pipeline* is used to define a sequence of indipendent steps called *stages*. Each *stage* can either be defined as a *Transformer* or an *Estimator*. A successful execution processes input data as it passes through different steps. A sequence of multiple *Transformers* and *Estimators* constitutes a *ML workflow*.

The `MLlib` library in Spark supports many classification methods. ❶ binary classification, ❷ multiclass classification and ❸ regression are some of the well methods that have support for distributed processing. In this paper, we will be use multiclass classification. This will help to us to identify complex nonlinear relationships in measure data.

Multilayer perceptron Classifier is a network of interconnected nodes called *neurons*. Nodes that belong to the same group forms a layer. Each node in one layer is connect to every node in the next layer. In addition, each node is also assigned a value called *weights* and an output function called *activation function*. The purpose of the *activation function* is to produces an nonlinear output for the sum of all the inputs to that node. This results in a connected network that represents a model of nonlinear mapping between input and output. A multilayer perceptron typically consists of an input layer, one or more hidden layers and an output layer. We use two *activation function*, one for the hidden layer and the other for the output layer. The hidden layers use *sigmoid (logistic)* since we have non negative inputs while the output layer uses *softmax* since we want to identify multiple classes. In order to identify complex nonlinear relationships a multilayer perceptron has to learn. This process called network *traning*. During this process the nodes in the network is provided with input data. On each iteration the *weights* of the nodes are adjusted until the input is mapped to the known output. In mobile environment, multilayer perception is an is good candidate as it makes no prior assumptions concerning the distribution of data. Furthermore, the model can be trained to accurately generalize unseen data.

3.1.1 Model preparation A Neural Net(NNET) needs a sample of input data for training and testing the model. We separated the input data randomly in a 60:40 ratio. The latter split was used to test the performance of the *model* while the former split was used as training data. In order to use the *ML Pipeline* in Spark, data has to be transformed as *label* and *features*. Spark has support for different transformation tools. As illustrated in Figure 2 our *ML workflow* consisted of `String Indexer` and `Vector Assembler` as *transformer* and `Multilayer Perceptron` as *estimator*.

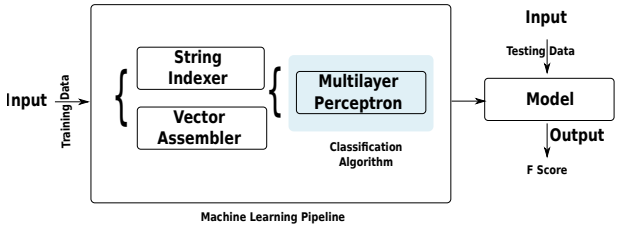

Figure 2. Machine Learning Workflow

For transforming data into *label* we use the `String Indexer`. It encodes a column consisting of string data to a column of indices(integer). Furthermore, to transform column into *features* we use the `Vector Assembler`. It is useful for combining raw features and features generated by different feature transformers into a single feature vector. The *estimator* is then fed with the output of `String Indexer` i.e. *label* and `Vector Assembler`, i.e. *features* to produce a results. The whole chain produces a *model* which can then be used to classify unseen data.

3.2 Gaussian Processes (GPs)

This section provides background about GPs which is taken from the author's previous work (Schneider et al., 2010) which

is based on the GP textbook (Rasmussen, Williams, 2006).

A GP is a collection of any finite number of random variables which have a joint Gaussian distribution. The supervised learning problem requires a training set $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$ consisting of N input points $x_i \in \mathbb{R}^D$ (where D, represents the dimensionality of the data) and the outputs $y_i \in \mathbb{R}$ to compute the predictive distribution $f(x_*)$ at any new test point x_*. A Gaussian Process model employs a multi-variate Gaussian distribution over the space of function variables $f(\mathbf{x})$, mapping input to output spaces. A GP is specified by its mean function $\mu(\mathbf{x})$ and kernel / covariance function $k(\mathbf{x}, \mathbf{x}')$, so that $f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. Denoting $(X, \mathbf{f}, \mathbf{y}) = (\{x_i\}, \{f_i\}, \{y_i\})_{i=1}^{N}$ for the training set and $(X_*, \mathbf{f}_*, \mathbf{y}_*) = (\{x_{*i}\}, \{f_{*i}\}, \{y_{*i}\})_{i=1}^{N}$ for the test data, the joint Gaussian distribution with mean zero ($\mu(x) = 0$) becomes

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \quad \begin{bmatrix} K(X,X) + \sigma^2 I & K(X, X_*) \\ K(X, X_*) & K(X_*, X_*) \end{bmatrix}\right). \quad (3)$$

In (3), $\mathcal{N}(\mu, \mathbf{\Sigma})$ is a multi-variate Gaussian distribution with mean μ and covariance $\mathbf{\Sigma}$ and K is the covariance matrix computed between all samples in the data set. The predictive distribution for new data points can be obtained as $p(f_i|X_*, X, \mathbf{y}) = \mathcal{N}(\mu_*, \mathbf{\Sigma}_*)$ where

$$\mu_* = K(X_*, X)[K(X,X) + \sigma^2 I]^{-1}\mathbf{y}, \quad (4)$$

, by conditioning on the observed training points.

$$\mathbf{\Sigma}_* = K(X_*, X_*) - K(X_*, X)[K(X,X) + \sigma^2 I]^{-1} K(X, X_*) + \sigma^2 I. \quad (5)$$

Learning the hyper-parameters of the covariance function from a dataset is equivalent to learning a GP model. This can be performed by maximizing the log of the marginal likelihood with respect to θ in a Bayesian framework such as:

$$\log p(\mathbf{y}|X, \theta) = -\frac{1}{2}\mathbf{y}^\intercal [K(X,X) + \sigma^2 I]^{-1}\mathbf{y} - \frac{1}{2}\log|K(X,X) + \sigma^2 I| - \frac{N}{2}\log 2\pi \quad (6)$$

On the right side of equation (6), the three terms represent (from left to right) the data fit term, complexity penalty term and a normalization constant.

3.2.1 Training The hyper-parameters which control the GP (Rasmussen, Williams, 2006), can be estimated from the data using the maximum likelihood approach (Eq.6). Initially, the hyper-parameters were initialized with random values and then used in a gradient descent based method to search for the optimal hyper-parameters. In order to avoid a local minimum, the hyper-parameter search is repeated several times with different random starting values (Guyon, Elisseeff, 2003). After this, the best hyper-parameter set is obtained by comparison of the magnitude of the log marginal likelihood iteration of the search and finally selecting the one with the highest value.

3.2.2 Prediction The predictive distribution for a test point x_* is obtained from the joint Gaussian distribution of the training data and the test data by conditioning on the observed / known targets in the training set. The predictive distribution is generally Gaussian with a mean and a covariance given by (Eq. 4 and 5). The prediction thus yields the most likely label for a new test point x_* with a given variance around it.

3.3 Classification

So far the predictions are real valued numbers as this method describes a regression problem, thus far. However, in the context of GPs, classification and regression can be viewed as *function approximations*. This means, the GP yields a guess for a class label and does not take the classical form of *class predictions* (Rasmussen, 2004). The set of numbers for the targets and the predictions are different, i.e $y \in \mathbb{Z}$ and $p(f_*|X_*, X, \mathbf{y}) \in \mathbb{R}$; thus, a decision boundary or classes need to be defined. This is done by defining the boundary as the equal distance from two class labels $\{-1, +1\}$ which yields a decision boundary of zero:

$$C = \begin{cases} -1, & \text{if } \mathbb{E}(f_*|X_*, X, \mathbf{y}) < 0 \\ 1, & \text{if } \mathbb{E}(f_*|X_*, X, \mathbf{y}) > 0 \\ 0, & \text{if } \mathbb{E}(f_*|X_*, X, \mathbf{y}) = 0 \end{cases}, \quad (7)$$

where zero defines the *null-class* which practically, never occurs. $\mathbb{E}(f_*|X_*, X, \mathbf{y})$ stands for the expected value for f_*. The algorithm was implemented in MatLab (R2018b) according to (Rasmussen, Williams, 2006).

3.4 'One vs. All' approach

In an 'One vs. All' (Rifkin, Klautau, 2004) paradigm, all observations of one particular emotion are extracted from the data set and labeled as class '-1', the remaining emotions was labeled as class '1'. The extracted emotion was substituted with another emotion from the data set after the GP was applied to this data set and the new configuration was applied again using the GP. This process was repeated until all emotions were classified using the one vs. all approach, i.e. for the mentioned data set (Section 4.1), it was repeated six times.

3.5 Kernels

To classify the data, two different kernels (also termed covariance functions) were used. The first kernel was the very commonly used squared exponential (SE) covariance function, the second was a newer kernel - the Observation Angle Dependent kernel (OAD) (Melkumyan, Nettleton, 2009) - used in material science and hyperspectral image processing for modeling discontinuous functions (Schneider et al., 2014). Both kernels were used in an Automatic Relevance Determination (ARD) approach (Hastie et al., 2009, Guyon, Elisseeff, 2003) by which the hyper-parameters and their importance in the model are learned by the GP model during the training stage. For simplicity, the *characteristic length-scale matrix* Ω was assumed to be diagonal, meaning that each dimension of the data was represented by an individual *characteristic length-scale* parameter $l_{1\ldots D}$. Thus $\Omega = \text{diag}(D)$:

$$\Omega = \begin{bmatrix} l_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & l_D \end{bmatrix} \quad (8)$$

3.5.1 The Squared Exponential kernel (SE) is infinitely differentiable because of its exponential term. Thus, for relatively smooth functions (e.g. without no large discontinuities),

this is a useful property, yielding also a quite smooth GP output. Further, it is *stationary* i.e. invariant against translation as suggested by the $\mathbf{x} - \mathbf{x}'$ term. The signal variance σ_0^2 is one of the hyper-parameters of this covariance function. As the kernel was used in an ARD setting, the parameter l normally used was replaced by the Ω, thus the kernel can account for the variation in each direction / dimension. The SE kernel is then written as:

$$K(\mathbf{x}, \mathbf{x}') = \sigma_0^2 \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^\intercal \, \Omega(\mathbf{x} - \mathbf{x}')\right] \quad (9)$$

3.5.2 The Observation Angle Dependent kernel is a *non-stationary* covariance functions, which depends only on the angle at which points are observed (Melkumyan, Nettleton, 2009). The angle is a measure of similarity between two data entries (i.e. two observations) by considering each multi-dimensional data point as a
D-dimensional unit vector in $\mathbb{R}^D$. Because this kernel uses an angular metric rather than a distance metric, the norm of the vector does not affect the result, i.e. it is scale invariant. The OAD covariance function is defined as:

$$K(\mathbf{x}, \mathbf{x}') = \sigma_0^2 \left(1 - \frac{1 - \sin\varphi}{\pi} \cdot \frac{\mathbf{x}^\intercal \Omega \mathbf{x}'}{\sqrt{\mathbf{x}^\intercal \Omega \mathbf{x}} \cdot \sqrt{\mathbf{x}'^\intercal \Omega \mathbf{x}'}}\right) \quad (10)$$

where σ_0, φ are two scalar hyper-parameters (in addition to Ω) of the covariance function.

4. DATA AND RESULTS

4.1 Data set and pre-processing

The data set used in this study was produced in previous work and explained in detail (Dastageeri et al., 2019). In total 32 female and 5 male business psychology students participated in the studies. The mean age was (22), the mean height (173) cm and the mean weight (65) Kg. Missing data was replaced by the method of linear interpolation. As methods of analysis, explanatory and predictive approaches were chosen. The aim of previous work was to distinguish between two emotions - happy or scared - of pedelec drivers to locate areas or situations of discomfort in the city of Stuttgart. In this work, the focus is on improving the detection of emotions and to distinguish between more emotional states as mentioned above. The main difference between the data set in this study and the previous one is, that the classification of emotions is more diverse and specific, i.e. ❶ angry, ❷ comfortable, ❸ enthusiastic, ❹ happy, ❺ scared, ❻ stressed.

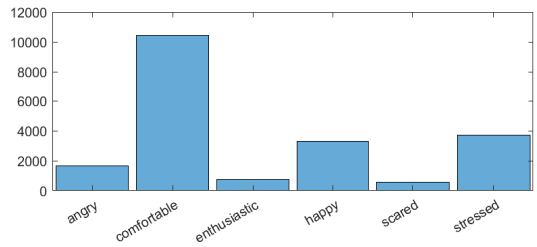

Figure 3. Distribution of classes within the data set.

Data were cleaned by removing outlier in data, e.g. heart rate values below and above 50 and 205 bpm, respectively, were removed. Also entries with a heart rate variability which were larger than the heart rate were removed. Similar corrections were made for body temperature, skin conductance (i.e. values ≤ 0). Generally, entries which were not represented in a large number of observations (i.e. ≤ 20) were removed. Furthermore, all data entries with gender male were removed, as there were only few entries in the data set and the features and distribution of emotions was distributed differently for this gender - hence the feature 'gender' was not used. After cleaning there were 17168 data entries with eight features per entry. The features that remained were: ① heart rate, ② heart rate variability, ③ body temperature (chest belt), ④ skin conductance, ⑤ body temperature (wrist band), ⑥ age, ⑦ height and ⑧ weight.

The GPs used only 1,717 samples across the six classes as training data, the NNET used 10,300 training samples. Thus, the test samples for GPs and NNET were 15,451 and 6,868, respectively. This is a huge difference in the amount of training and test data that was used between the two methods.

4.2 Results

In the following section, results of the ML methods are presented using F-scores only. We did reject the use of accuracies considering a preference bias towards the minority class examples in the data set. Having an imbalanced data sets as being the case in this study, accuracy is not suitable because the impact of the least represented, but potentially more important examples, is reduced when compared to that of the majority class (Branco et al., 2016). The F-score on the other hand is more informative about the effectiveness of a classifier on predicting correctly the cases that matter to the user (Estabrooks, Japkowicz, 2001). Only when both recall (a measure of completeness) and precision (a measure of exactness) are high, the resulting F-score value, is also high (Branco et al., 2016).

The average F-score values for the classification of the six different emotions used in this study are on average very high with values above 95% for the GP method using either kernel. There is little variance between the classification performance of the different emotional states, hence the small standard deviation across the the six emotions can be obtained (Table 1). The lowest F-scores were observed for emotions *angry, scared* and *enthusiastic* for both kernels and also using the NNET. There is little to chose between the OAD and the SE kernel within the GP method, as the difference in performance mostly shows in the third decimal place when comparing F-scores. The NNET also performs well, with overall F-scores of 0.73, however, compared to the GPs, it performed rather poorly. The weakest performance of the NNET was achieved for the emotion *stressed* which corresponds to the class with a relatively low number of (training) samples (cf. Figure 3).

Table 1. Summary of F-scores using the different methods. Mean and standard deviation (SD) was calculated across the 6 different emotions.

F-scores	GP-SE	GP-OAD	NNET
angry	0.869	0.876	0.716
comfortable	0.966	0.968	0.857
enthusiastic	0.961	0.920	0.896
happy	0.909	0.935	0.689
scared	0.973	0.931	0.596
stressed	0.921	0.918	0.642
mean	0.933	0.925	0.733
SD	0.041	0.030	0.119

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-VI-4-W2-2020-149-2020 | © Authors 2020. CC BY 4.0 License.

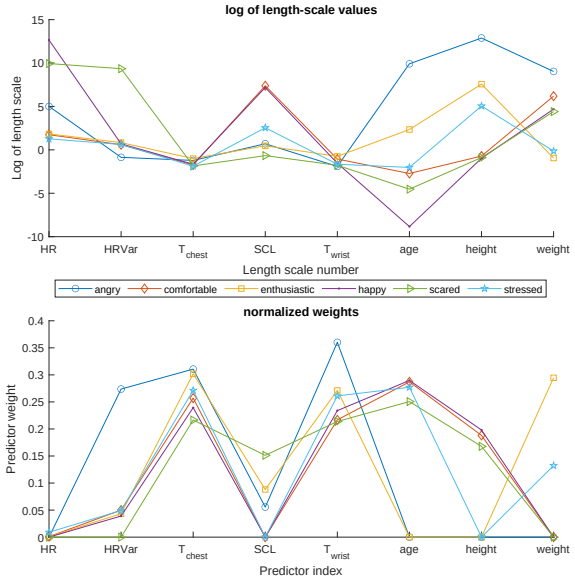

Figure 4. SE length scale parameters (log values) and parameter weights with SCL parameter. Note that for values smaller zero, the absolute value of l was used. This applies to the following figures of this type.

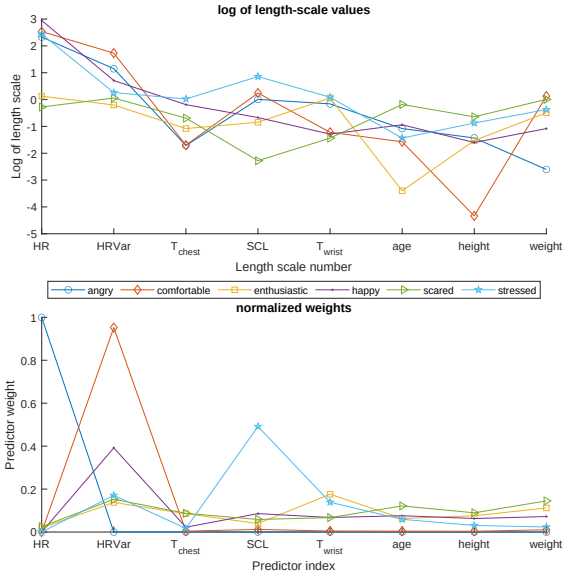

Figure 5. OAD length scale parameters (log values) and parameter weights with SCL parameter.

4.2.1 Parameter analysis The weights of different physiological parameters such as HR, SCL and so forth showed different contributions to the model depending on the emotion and the covariance function that was used. For the SE kernel, the parameter weights were quite variable (Figure 4, bottom), ranging from zero for HR, age, height and weight (for most of the emotions) to around 0.3 to 0.4, for the chest and wrist temperatures, respectively. The latter was true for all emotions, while HRvar and weight seemed to be important for *stressed* and *enthusiastic*.

For the OAD kernel, the weights were rather unspecific and had low values, from close to zero around 0.2 for most parameters and emotions (Figure 5). However, for HRvar, the weights were much higher than for the other parameters, even ranging close to 1 for *comfortable*. The weight for HR for *angry* was also very high, however, close to zero for the other emotions. SCL yielded weights around 0.1 for all emotions, except for *stressed* where the weight was around 0.5.

4.2.2 Results after removing SCL As many studies suggest that SCL is of great importance to the determination of the emotion, SCL was collected. Though the circumstances of the experiment regarding environmental temperature, physical activity and general mental distraction of the traffic differed from the recommended laboratory environment.To probe its overall relevance and impact despite all adversity, SCL was removed in the aftermath. For this, the classification task was redone and the results were reported in Table 2. Interestingly, the results were on averag,e and for the specific classes (i.e. emotions), very similar to the ones obtained with SCL. The performance of the NNET without SCL was much poorer, than for the GPs where practically not real difference was observed. Further investigation is needed to find out the cause. Presumably, the weights and relevance for SCL will change in a laboratory environment. Furthermore, the NNET yielded F-scores of 0 - 0.075 for the same classes it performed the poorest in the previous set of results (i.e. with SCL), namely *angry, scared* and *enthusiastic*. Again, this is most likely, due to the low number of (training-) samples for these classes.

Table 2. Summary of F-scores using same data as previously, however, without SCL parameter.

F-scores	GP-SE	GP-OAD	NNET
angry	0.879	0.862	0.075
comfortable	0.966	0.968	0.771
enthusiastic	0.920	0.878	0.000
happy	0.909	0.923	0.509
scared	0.920	0.912	0.000
stressed	0.921	0.932	0.440
mean	0.919	0.913	0.299
SD	0.028	0.038	0.321

5. DISCUSSION

The results of this study suggest at first glance that it is possible to distinguish between human emotions based on physiological properties of probands during exercise. Classification results using the machine learning methods were high throughout for GPs and variable for NNETs. While NNET yielded the lowest average performance of around 0.76, the GP methods, irrespective of the kernel used, yielded a performance beyond 0.9. While these result look very promising, they have to be take with caution as discussed in the following paragraphs. In fact, there are several aspects in the obtained results, that suggest that much more research is needed to truely define human emotions based on physiological sensor readings.

❶ Humans, especially young adults, have difficulties to exactly determine their state of emotion. It is relatively easy to distinguish between positive and negative feelings, however, to differentiate between pure emotions are difficult as that state is relativeley rare and short lived (Cowie et al., 2002).

❷ From a data point of view, weak classifications rates (using GPs) for some emotions, for the emotion *angry, scared*

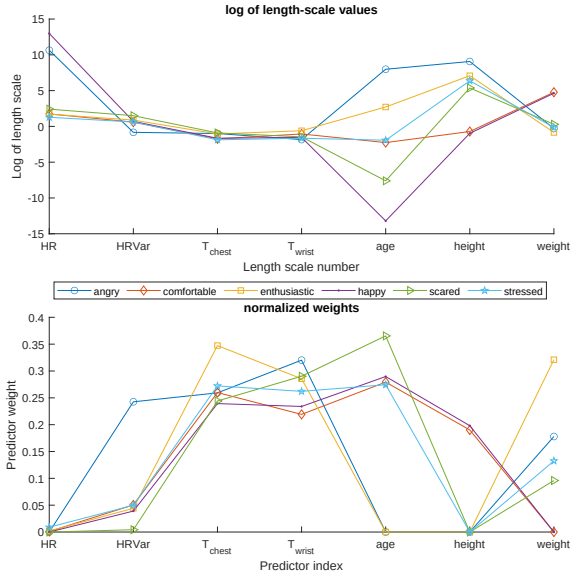

Figure 6. SE length scale parameters (log values) and parameter weights for the data set without SCL.

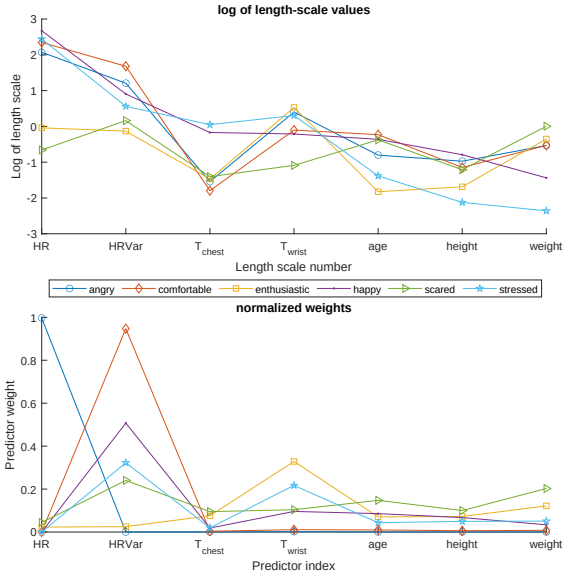

Figure 7. OAD length scale parameters (log values) and parameter weights for the data set without SCL.

and *enthusiastic* can be explained with a very small number of samples for this class. This point is becomes even more obvious, when examining the effects of SCL on the NNET's performance as the same classes are effected severely. However, the entire performance of the NNET drops severely, indicating that this method requires clearly much more training data with more features than the GPs. One reasons why *stressed* yielded low performance may be that this emotion is difficult to gauge for humans and often mixes with other emotions as there can be negative but also positive stress. However, this in a way shows, that SCL does in fact has some impact and relevance, at least for some methods which is in line with related literature. However, what this really means for developing an *emotion predictor application*, must be investigated further and cannot be covered in this paper.

❸ The analysis of SCL is ambiguous. The comparison of the methods with and without the usage of SCL show only little importance to the results using GPs. Several reasons might cause this issue. Measurement of SCL during a physical activity might be one. Another subject is the effect of the measurement environment concerning the humidity and temperature outside which differs strongly from the laboratory environment. Further studies needs to specify each reason with regards to their respective effect.

❹ This approach used data from a exceedingly homogeneous group of probands. The considered values were derived from females exclusively with similar age and weight. This might be a reason for the above average results. The next step would be to continue experiments with much more and heterogeneous group of probands. Male and female should be considered with different age, weight and fitness level. Furthermore, the experiments should be repeated under a laboratory environment to compare the impact of each parameter and especially of SCL.

❺ Notwithstanding all those points of caution, this is a first attempt to differentiate between *six* different emotions and results certainly show promise to accomplish this task. In future, more sophisticated methods (e.g. Deep Neural Networks) and further experiments will improve upon the results presented in this study.

6. CONCLUSION

One focus of this work was to weigh the parameters in relation to an emotional state. The results regarding go along with current findings. For the SE kernel as well as the OAD kernel HRvar is important for the emotion *stressed*. For the OAD kernel HR was significant for the emotional state of *angry*. The effect of SCL was generally not high as the experiment environment did not fit with the recommended requirements for measurements. But for the small difference observed, SCL correlated with the emotion *stressed*.

The results seems promising as they broadly fit the expected outcome. Though, further studies are required to verify weight for the parameters HRvar, HR and SCL and to obtain the weights for body temperature (chest and wrist), age, height and weight. Further, experiments in a more controlled environment need to be conducted and more ground-truth or at least a *surrogate* ground-truth needs to be generated during the drive e.g. using facial emotional recognition based on Deep Convolutional networks.

ACKNOWLEDGEMENT

Removed due to double blind review process. This work has been jointly developed in the project Simstadt 2.0 (Funding number: 03ET1459A) and i_city (Fundingnumber: 03FH9I01IA). The project i_city is supported by the German Federal Ministry of Education and Research (BMBF), while project Simstadt 2.0 is supported by the German Ministry of Economics (BMWi). The authors are responsible for the content of this publication. The authors would also like to thank the anonymous reviewers who have helped to increase the quality of the paper.

The project team would like to thank Prof. Dr. Thomas Bäumer, Prof. Dr. Patrick Müller and Jan Silberer of the Department of Business Psychology, HFT - University of Applied Sciences Stuttgart for their support and encouragement. Furthermore, the authors would like to express gratitude to the project team of Daimler TSS for their aid and assistance for providing and setting up the pedelecs as well as for sharing their expertise and experience.

REFERENCES

Arbeitsmedizin, B., Forschung, S. W., 2014. Leitlinie Nutzung der Herzschlagfrequenz und der Herzfrequenzvariabilit in der Arbeitsmedizin und der Arbeitswissenschaft. 1–60.

Autonom Health, 2017. Kennen Sie die Grundlagen der Herzratenvariabilität?

Benedek, M., Kaernbach, C., 2010. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1), 80–91.

Benedict, F. G., Miles, W. R., Johnson, A., 1919. The Temperature of the Human Skin. *Proceedings of the National Academy of Sciences*, 5(6), 218–222.

Bierman, W., 1936. The temperature of the skin surface. *Journal of the American Medical Association*, 106(14), 1158.

Boucsein, W., 2012. *Electrodermal Activity*.

Branco, P., Torgo, L., Ribeiro, R. P., 2016. A Survey of Predictive Modeling on Imbalanced Domains. *ACM Comput. Surv.*, 49(2).

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. G., 2002. Emotion Recognition in HCI. *IEEE Signal Processing Magazine*, 19(6), 2–4.

Dastageeri, H., Rodrigues, P., Silberer, J., 2019. Happy or scared - Detecting Emotions of pedelec drivers in urban areas. IV-4/W9, 27–33.

Edelberg, R., 1972. Electrical activity of the skin: Its measurement and uses in psychophysiology. *Handbook of Psychophysiology*, 12, 1011.

Estabrooks, A., Japkowicz, N., 2001. A mixture-of-experts framework for learning from imbalanced data sets. *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, IDA '01, Springer-Verlag, Berlin, Heidelberg, 34–43.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3, 1157–1182.

Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Hoffman, T., 2020. Was ist die Herzratenvariabilität (HRV) und wieso ist sie wichtig?

Jelinek, H. F., Cornforth, D. J., Khandoker, A. H., 2017. Introduction to ECG time series variability analysis: A simple overview. *ECG Time Series Variability Analysis: Engineering and Medicine*, 1–12.

Kohn, L., Dastageeri, H., Bäumer, T., Moulin, S., Müller, P., Coors, V., 2018. Hot or not – identifying emotional " hot spots " in the city.

Kolata, G., 2001. 'Maximum' Heart Rate Theory Is Challenged.

Lexikon der Biologie, Herzfrequenz, 1999.

Mackowiak, P. A., 1992. A Critical Appraisal of 98.6 F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich. *JAMA: The Journal of the American Medical Association*, 268(12), 1578.

Martens, 2018. Herzfrequenz: Das ist der Normalwert des Ruhepulses.

Melkumyan, A., Nettleton, E., 2009. An Observation Angle Dependent Nonstationary Covariance Function for Gaussian Process Regression. *Lecture Notes in Computer Science*, Neural Inf, Springer, 331–339.

Rasmussen, C., 2004. Gaussian processes in machine learning. O. Bousquet, U. Luxburg, G. Rätsch (eds), *Lecture Notes in Artificial Intelligence 3176 - Advanced Lectures on Machine Learning*, Springer-Verlag, Heidelberg, chapter Gaussian P, 63–71.

Rasmussen, C. E., Williams, C. K. I., 2006. *Gaussian Process for Machine Learning*. The MIT Press, Cambridge, Massachusetts.

Rifkin, R., Klautau, A., 2004. In Defense of One-Vs-All Classification. *J. Mach. Learn. Res.*, 5, 101–141.

Schneider, S., Melkumyan, A., Murphy, R. J., Nettleton, E. W., 2010. Gaussian Processes with OAD covariance function for hyperspectral data classification. *Proceedings of IEEE International Conference on Tools with Artificial Intelligence*, Arras, France.

Schneider, S., Murphy, R. J., Melkumyan, A., 2014. Evaluating the performance of a new classifier – the GP-OAD: A comparison with existing methods for classifying rock type and mineralogy from hyperspectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98, 145–156.

Sund-Levander, M., Forsberg, C., Wahren, L. K., 2002. Normal oral, rectal, tympanic and axillary body temperature in adult men and women: a systematic literature review. *Scandinavian Journal of Caring Sciences*, 16(2), 122–128.

Tanaka, H., Monahan, K. D., Seals, D. R., 2001. Age-predicted maximal heart rate revisited. *Journal of the American College of Cardiology*, 37(1), 153–156.