# "Emotions are the Great Captains of our Lives": Measuring Moods through the Power of Physiological and Environmental Sensing

Keith April Araño, Peter Gloor, Carlotta Orsenigo, and Carlo Vercellis

**Abstract**—The paper proposes the use of a smartwatch-based system for measuring the emotions of individuals in a classroom setting with respect to five mood variables: Activation, Tiredness, Pleasance, Quality of Presentation and Understanding. Internal (body) and external (environment) data such as movement, heart rate, noise, temperature and humidity were collected through the built-in sensors of the smartwatch. The system was verified by means of a longitudinal study that has been carried out in a series of workshops and lectures. Through experience-based sampling, participants were polled at periodic time intervals asking them to enter a self-assessment of the aforementioned mood states directly on the smartwatch. The goal was to demonstrate whether sensor data can be used to effectively predict the five moods. By resorting to a machine learning approach our system was able to predict the moods with an accuracy ranging between 89-95% for single-output classification, 92-99% for the chain classification task and of approximately 93% for the multi-output analysis. Our results showed also that body signals are better predictors compared to the external environmental variables. These results demonstrate and verify the potential of smartwatches in collecting and predicting human emotions, enabling dynamic feedback loops to enhance user experience.

**Index Terms**—Machine learning, affective computing, modelling human emotion, affect sensing and analysis

## 1 Introduction

According to van Gogh, "little emotions are the great captains of our lives and we obey them without realizing it". What if we had a system that empowers us to be the great captains of our emotions? The aim of this paper is to evaluate the predictive ability of a set of body-related and environmental sensor data with respect to different emotional states, in order to build a tool which automatically provides dynamic feedbacks that can be used to enhance people's comfort when engaged in a given scenario.

Technological innovation underpins nowadays most, if not all, products and services. From one side, artificial intelligence (AI) is shaping nearly every aspect of life enabling, for instance, the emergence of virtual assistants or agents able to interact with humans, or of autonomous vehicles embedding autopilot capabilities. From the other, the prevalence of low-cost sensor technology and the increasing popularity of machine learning (ML) methods led to the development of wearable devices which can measure, track and predict different variables in a wide range of application areas, such as healthcare [1], [2], security [3], [4] and education [5], [6], just to name a few. In particular, a growing interest in the application of AI in this last sector has been observed, due to its potential to identify the main drivers of students performance and to design adaptive and personalized learning paths [7].

Recent studies on the close relationship between emotions and environmental factors have shown the predictive power of sensor data and location patterns in analyzing human emotions [8], [9]. Based on the hypothesis that certain emotions can be effectively predicted through internal (body) and external (environment) sensor data, in this paper we focus on the application of a smartwatch-based system first introduced in [10] for measuring mood variables, in order to collect immediate feedbacks and let individuals dynamically and intelligently adapt their responses according to different situations. Specifically, the paper is an extension of [8] by introducing the use of smartwatches in a real-world scenario represented by a classroom, and contains the following main novel contributions.

First, the proposed approach resorts to a non-intrusive and scalable smartwatch-based system to collect the measurements of sensor data and of five mood variables (i.e. *Activation*, *Tiredness*, *Pleasance*, *Quality of Presentation* and *Understanding*) for the purpose of mood prediction in a classroom setting. Previous works on the same subject were focused on Activation and Pleasance alone. Furthermore, to the best of our knowledge the combined use of internal (body) and external (environment) sensor data for the prediction of all the aforementioned moods has not been previously considered in similar studies. Second, the paper analyses a real-world dataset gathered in a classroom setting from participants wearing smartwatches, who were periodically polled about the five moods under investigation. As opposed to most of the publicly available datasets collected in controlled laboratory settings, our measurements were recorded in the participants' natural environment and are therefore more truthful. Moreover, performing experiments in a real-world scenario allows the results to be better generalized to similar cases. Third, besides more traditional

- *K. Araño, C. Orsenigo and C. Vercellis are with the Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Milan, Italy 20156.*
  *E-mail: keithapril.arano@polimi.it*
- *P. Gloor is with the Center for Collective Intelligence, Massachusetts Institute of Technology, Boston, United States 02139.*

single-output prediction tasks, the paper illustrates a multi-output classification analysis which has never been considered previously, with the aim of exploiting the potential correlation among the mood variables. Finally, it also implements a classifier chain model to further explore such correlation. As far as we know, none of the previous works performed a similar study.

To assess the predictive ability of the set of internal and external input variables several machine learning experiments were carried out by comparing the performances of different classifiers, ranging from classical methods such as K-Nearest Neighbor, Decision Trees and Support Vector Machines to more sophisticated neural networks-based techniques. The set of algorithms considered in this paper is broader compared to similar studies which explored the prediction ability of fewer supervised learning methods.

The remainder of the paper is organized as follows. Section 2 provides an overview of previous research studies relying on machine learning and sensor data for the purpose of mood prediction, or for specific predictive analysis, regarding students academic performance. Our data collection process and the ML problem statement are outlined in Section 3. Section 4 describes the data pre-processing and the experimental setup, while Section 5 illustrates the results achieved for each machine learning task. Conclusions and future developments are finally discussed in Section 6.

## 2 Related Works

The increasing availability of low-cost sensor technology embedded in wearable devices which enable the continuous collection of large amounts of data has strongly revealed the strategic role of machine learning in a variety of domains, such as healthcare, security, transportation, disaster management, education. It is indeed an exciting time for the research community as more and more groundbreaking technologies are developed and several novel application areas will emerge in the future. This represents an opportunity for researchers to develop new approaches, to harness both the technologies and the data currently available.

Emotions have been studied in psychology for a long time. While there is no universally accepted definition of emotions, they are frequently studied in connection with mood, creativity, and motivation [11]. Emotions are defined as "positive and negative experiences that are associated with a particular pattern of activity" [12]. One of the most accepted frameworks distinguishes six key emotional states: anger, disgust, fear, happiness, sadness and surprise [13], while one of the most influential emotional theories has been brought forward by Lazarus [14], who discerns three phases in the execution of an emotion. First, there is the cognitive appraisal, where the individual assesses the event triggering the emotion. Second, there are physiological changes, such as change in the heartrate, or production of hormones, and third, action, where the individual feels the emotion and reacts to it.

The majority of the studies on human emotions resorting to a machine learning framework focused on the analysis of emotions inferred from written texts, facial expressions and speech. Recently, with the growing popularity and diffusion of wearable sensor technologies, physiological signals received more attention and a significant number of researchers started to investigate the relationship between sensor data and human emotions. For example, in [9] a data-driven approach based on sensor data is presented with the purpose of assessing the impact of a person's external environment and physiological changes in predicting emotions. The authors employed statistical methods for feature analysis and various machine learning algorithms such as Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbor (KNN) and Naive Bayes (NB) as classifiers. The same supervised learning methods were shown to be effective for classifying emotions using body sensor data as features [15], [16], [17].

More sophisticated studies in which different modalities are combined to predict emotions have been also developed. For example, in [18] authors analyzed jointly computer interactions, facial expressions, body postures and physiological signals collected via body sensors to detect work stress in office environments. Specifically, they employed alternative machine learning methods available in Weka [19], such as Linear Regression (LR), KNN, SVM, Regression Trees and Multilayer Perceptron (MLP). Another study by Brady et al. [20] used a publicly available dataset for emotion prediction by combining audio, video and physiological sensor data. In this case, SVM, Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) were used as classifiers for the audio, video and physiological data, respectively.

While such studies were devoted to investigate how sensor data can be exploited to predict emotions, a limited number of works focused on the classroom setting use-case. In this application domain the potential impact of adaptive learning and personalized teaching using machine learning can drastically improve the current traditional educational systems [21], hence motivating researchers to make significant contributions in this emerging field. In a recent study by Herrera-Alcántara et al. [22], a student activity recognition system that relies on ML to analyze data from smartwatches and smartphones was developed, and is planned to be used in the future to build a recommendation system to enhance students academic performance. [6] proposed an early recognition system to identify students who are at risk of failing a course. In [23] authors collected students electroencephalography (EEG) brain activity and trained classifiers to detect confusion. More advanced studies were also developed such as in [24], where reinforcement learning was employed to recommend effective learning activities from the analysis of students' heart beats, quiz scores, blinks and facial expressions.

In these state-of-the-art works different machine learning methods were applied to improve the student learning process by analyzing data coming from body sensors and other similar devices. Yet they seldom investigated the effects of external environmental variables such as noise, weather conditions, room temperature, humidity and location. Prior research on the relationships between human emotions and the surrounding environment [9], [10], [25] showed the potential influence of such environmental factors in predicting emotions, therefore motivating their inclusion in the analysis. Moreover, to the best of our knowledge predicting the mood variables Activation, Pleasance, Tiredness, Quality of Presentation and Understanding has not been explored in previous studies on the same subject. In this paper we therefore propose a mood variable predictive system which is based on both internal (body) and external (environment) sensor data, collected by periodically polling participants attending several lectures and workshops about the five aforementioned emotional states. The general framework of our approach in terms of input variables and predicted mood states is depicted in Figure 1.

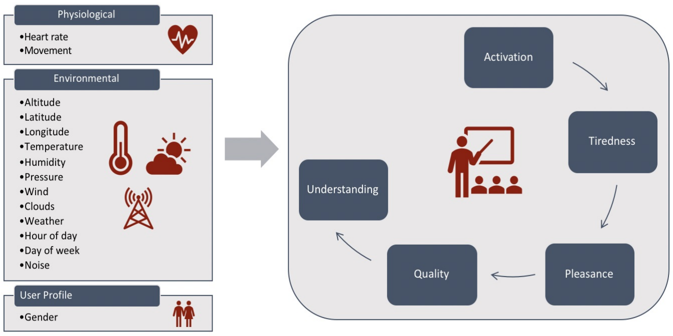


Fig. 1. Framework of the proposed approach

## 3 Methodology

### 3.1 Data Collection

Several body-sensing, environment and mood-related data were recorded for 30 people during five different COINS (Collaborative Innovation Networks) lectures/workshops held in China, Germany and Switzerland in September and October 2018. To this aim, individuals were provided with commercial smartwatches which were connected to their smartphones. Thanks to an app developed ad-hoc and called *Happimeter*, it was then possible to collect data from the smartwatch built-in accelerometer, light sensor, microphone and heart rate sensor, while the location was detected through the GPS sensor of the smartphone, as described in [8].

Mood-related variables were similarly measured through the smartwatch-based system by periodically polling each participant. Specifically, individuals were asked to answer the following questions about the five emotional states, in a Likert scale between 0 and 2:

1. Activation (How active do you feel?)
2. Tiredness (How tired do you feel?)
3. Pleasance (How pleasant do you feel?)
4. Quality (What is the quality of the presentation?)
5. Understanding (Do you understand the presentation?)

The choice of the mood variables was based on the findings of previous behavioral, cognitive neuroscience and developmental studies of affect, and of works focused on the factors that, among others, exert an influence on students learning and growth in achievement.

The inclusion of Activation and Pleasance was primarily motivated by the well-known and widely adopted Circumplex Model of Affect, according to which affective states can be described as a linear combination of two neurophysiological dimensions, one related to valence and the other to arousal, or as a varying degree of both [26]. Each emotional state (i.e. excited, elated, sad, happy, etc.) can be, in particular, expressed as the product of a positive/negative valence (in a pleasure-displeasure continuum) and of the level of alertness (or activation), and can be drawn depending on the Likert scale values inputted: an activation level of 2 and a pleasance level of 1, for example, elicit an excited emotion. The mood variable Tiredness, on the other hand, can be linked to learning capacity and academic performance since its increase is intimately related to students' lower self-discipline and reduced ability to persist in achieving goals [27]. The Quality of Presentation deserves similar attention, as teaching quality has been consistently identified as an important factor when it comes to students' learning [28], [29]. Finally, Understanding plays a prominent role as a formative assessment since it is often used to gauge learning in a classroom environment [30]. For these last mood variables increasing values on the Likert scale were associated to more intense emotional states: adverse for Tiredness, desirable for Quality of Presentation and Understanding.

These ground truth mood data, that were used to build the prediction models, were based on the subjective evaluation of the participants. We deemed this a reasonable approach since individuals are potentially the best judges of their own emotional states. Observe that, this is in contrast to most of the publicly available emotion datasets which were commonly labeled by various external annotators, who weren't the same individuals who elicited the labeled emotions. In the LIRIS-ACCEDE dataset [31], for example, the ground truth annotations were collected through crowdsourcing from an heterogenous group of raters, which could have led to inconsistencies in the labels assignment.

The set of features periodically collected from each attendee is described in Table 1 and includes both numeric and categorical attributes. For each measurement a timestamp in the form (Day, Hour) was also recorded. Notice that, the feature *acceleration* describing the acceleration of the participant in the space was originally tracked in terms of average and variance, i.e. was actually represented by six features corresponding to the average and the variance of the measurements in the three-dimensional ($x$, $y$, $z$) space.

### 3.2 Machine Learning Problem Statement

In the context of machine learning predicting the mood variables can be cast in the form of a multi-category classification problem.

Formally, let $D$ denote a set of $t$ examples $(\mathbf{x}_i, y_i)$, $i \in \mathcal{T} = \{1, 2, \ldots, t\}$, in the $(n+1)$-dimensional real space $\mathbb{R}^{n+1}$, where $\mathbf{x}_i$ is a vector of $n$ attributes (or features) and $y_i$ is a scalar representing the label (or class) associated to $\mathbf{x}_i$. In our analysis, the input vector $\mathbf{x}_i$ corresponds to the set of sensor, timestamp and profile data recorded for participant $i$, where the label $y_i$ is associated to one of the mood variables and, therefore, takes its values in the set $Y = \{0, 1, 2\}$, according to the Likert scale used.

Let also $F$ denote a set of functions $f^{\alpha}(\mathbf{x}) : \mathbb{R}^n \mapsto Y$ representing hypothetical relationships between input vectors and labels, where $\alpha$ is a set of adjustable parameters.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3003736, IEEE Transactions on Affective Computing



TABLE 1
Description of the variables collected

| Category | Variables | Description |
|---|---|---|
| Sensor | Heart Rate | Average number of heart beats per minute |
| | Acceleration | Magnitude of the acceleration of the user movement in the physical 3D space ($x$, $y$ and $z$ axes) |
| | Altitude | Elevation of the location of the lecture/workshop |
| | Latitude and Longitude | Geographic coordinates of the location |
| | Temperature | Indoor temperature |
| | Humidity | Indoor humidity |
| | Pressure | Indoor pressure |
| | Wind | Outdoor wind level |
| | Clouds | Outdoor cloud level |
| | Weather | Outdoor general weather (clear, clouds, drizzle, rain) |
| | Noise | Indoor noise level |
| Timestamp | Hour of day | Hour of the day of the lecture/workshop |
| | Day of week | Day of the week of the lecture/workshop |
| Profile | Gender | Gender of the participant |
| Mood | Activation | Self-reported score for Activation, range $[0, 2]$ |
| | Tiredness | Self-reported score for Tiredness, range $[0, 2]$ |
| | Pleasance | Self-reported score for Pleasance, range $[0, 2]$ |
| | Quality of Presentation | Self-reported score for Quality of Presentation, range $[0, 2]$ |
| | Understanding | Self-reported score for Understanding, range $[0, 2]$ |

TABLE 2
Descriptive summary statistics of the numerical features recorded by the smartwatch

| | acc_avg_x ($m/s^2$) | acc_avg_y ($m/s^2$) | acc_avg_z ($m/s^2$) | acc_var_x ($(m/s^2)^2$) | acc_var_y ($(m/s^2)^2$) | acc_var_z ($(m/s^2)^2$) | altitude ($masl$) | heart rate ($bpm$) | clouds (%) | humidity (%) | latitude ($degree$) | longitude ($degree$) | noise ($Pa$) | pressure ($mbar$) | temperature ($Kelvin$) | wind ($mph$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | -7.3 | -467.1 | 233.4 | 96377.0 | 70262.2 | 93743.1 | 99.2 | 81.5 | 24.7 | 64.8 | 41.7 | 58.9 | 0.00539 | 1023.2 | 291.5 | 3.4 |
| std | 411.4 | 359.3 | 440.3 | 119309.6 | 83368.8 | 105207.5 | 149.3 | 22.8 | 34.6 | 12.9 | 9.4 | 54.5 | 0.00503 | 8.8 | 7.1 | 1.9 |
| min | -1034.0 | -1002.0 | -993.0 | -617534.0 | -710942.0 | -715435.0 | -260.3 | 0.0 | 0.0 | 44.0 | 30.0 | 6.8 | 0.00012 | 991.8 | 278.91 | 0.5 |
| 25% | -267.0 | -728.0 | -29.0 | 6561.0 | 9216.0 | 12100.0 | 3.1 | 69.0 | 0.0 | 51.0 | 31.4 | 10.9 | 0.00206 | 1021.0 | 285.0 | 2.0 |
| 50% | -13.0 | -552.0 | 290.0 | 51529.0 | 44521.0 | 59049.0 | 5.6 | 81.0 | 0.0 | 63.0 | 49.9 | 10.9 | 0.00389 | 1026.6 | 289.4 | 2.8 |
| 75% | 234.0 | -284.0 | 555.0 | 145924.0 | 101124.0 | 143641.0 | 237.0 | 94.0 | 64.0 | 72.0 | 49.9 | 120.6 | 0.00804 | 1028.0 | 299.0 | 4.9 |
| max | 1175.0 | 983.0 | 995.0 | 1024144.0 | 889249.0 | 736164.0 | 526.1 | 200.0 | 92.0 | 93.0 | 51.6 | 120.6 | 0.06069 | 1029.0 | 301.1 | 7.2 |

We are then required to determine a function $f^* \in F$ which optimally describes the relationship between $\mathbf{x}_i$ and $y_i$, i.e. for which a suitable measure of discrepancy between $f^\alpha(\mathbf{x}_i)$ and $y_i$ is minimized over $i \in \mathcal{T}$.

In our experimental framework three different multi-category classification tasks were addressed. The goal of the first was to evaluate the effectiveness of the set of input features for predicting each mood variable considered individually. To this aim, we solved separate classification problems by taking each mood in turn as the target variable to predict. This led to five different single-output multi-category classification models, obtained by deriving the best relationship between the input features $\mathbf{x}_i, i \in \mathcal{T}$, and each target as follows

$$\mathbf{x}_i = (x_{\text{i}1}, x_{\text{i}2}, \dots, x_{\text{i}n}) \xrightarrow{f^*_m(\mathbf{x}_i)} y_{\text{i}m} \in Y, \tag{1}$$

where $m \in \mathcal{M} = \{1, 2, \dots, 5\}$ indicates one of the mood variables Activation, Tiredness, Pleasance, Quality of Presentation and Understanding.

In the second task the emotional states were considered jointly and a multi-output multi-category classification analysis was performed. The purpose in this case was to explore the potential correlations among the targets and evaluate whether multi-output learning could achieve better classification results over the single-output counterpart by leveraging the information embedded in the mood variables taken as a whole. In a multi-output analysis it is required to build a predictive model which simultaneously generates a set of labels expressing different concepts, represented in our case by the emotional state levels, so that separate, although related, classification problems are solved at once. For the second task, therefore, the final aim was to identify the best association between individual inputs and their multiple target variables as follows

$$\mathbf{x}_i = (x_{\text{i}1}, x_{\text{i}2}, \dots, x_{\text{i}n}) \xrightarrow{f^*(\mathbf{x}_i)} \mathbf{y}_i = (y_{\text{i}1}, y_{\text{i}2}, \dots, y_{\text{i}5}), \tag{2}$$

where the label $y_{\text{i}m} \in Y, \forall i \in \mathcal{T}, \forall m \in \mathcal{M}$.

The third task was finally devoted to build a classifier chain model with the purpose of capturing the label dependencies and, therefore, exploring further the correlations among the mood variables. Based on the classifier chain paradigm, for each target a model is trained on the input vectors $\mathbf{x}_i, i \in \mathcal{T}$, augmented by the targets which have been formerly predicted along the chain. Specifically, the model generated for the $k$-th target in the chain formalizes the best relationship between inputs and output as follows

$$\mathbf{x}_i = (x_{\text{i}1}, \dots, x_{\text{i}n}, y_{\text{i}1}, \dots, y_{\text{i}k\text{-}1}) \xrightarrow{f^*_k(\mathbf{x}_i)} y_{\text{i}k} \in Y. \tag{3}$$

Notice that, the classification performance is in this case sensitive to the choice of the target order since different chain models involve a sequence of classifiers trained on different sets of input features. Finding the globally optimal chain is computationally expensive even for a small number $q$ of targets, as it would require to search over a space of $q!$ alternatives. For this reason, in the present study we confined the attention to one possible configuration by selecting a plausible order among the mood variables, as described in Section 5.3.

In our computational setting single, multi-output and chain models were generated by combining the data from all the participants. We are aware of the heterogeneity of the population involved in our experiments. However, the limited amount of available information prevented us from training robust personalized models for each individual.

## 4 Experiments

### 4.1 Data Pre-processing and Analysis

Prior to the experiments some pre-processing steps were applied to clean the data and arrange them in a form suitable for the subsequent classification task.

First, outliers were removed by computing the $Z$-score [32] for each numeric feature and discarding those examples whose score fell outside the range $[-3, 3]$. Second, missing data were resolved. Notice that, in our dataset missing values affected only two attributes, one numeric and the other categorical, represented by *noise* and *weather*, respectively. For the feature *noise* missing values were replaced by the mean of the known measurements recorded at each timestamp from the participants attending the same event. For the feature *weather*, instead, the most frequent value among the measurements recorded via the weather API, as described in [10], at the same timestamp for the same event was used, since the most common value was deemed a reasonable proxy. Third, certain sensor data were transformed into more meaningful features. In particular, Hour of day (*hour_of_day*) and Day of week (*day_of_week*) were extracted from the timestamps, as more useful insights could be obtained from these time encodings compared to the original datetime format (e.g. individuals are more active in the morning rather than in the afternoon). At the same time, from the raw latitude and longitude sensor data the *city* and the *country* were derived through reverse geocoding using Google's Geocoding API [33]. Finally, the accelerometer measurements originally expressed in a three-axes representation $(x, y, z)$ were replaced by a single feature called *movement*, defined by the root mean square of the three components ($\sqrt{x^2 + y^2 + z^2}$), as suggested in [9].

The final set of numeric features collected from the sensors, and submitted to further analysis together with the categorical predictors *gender*, *weather*, *hour_of_day*, *day_of_week*, *city* and *country*, is described in Table 2, which also reports for each attribute some descriptive statistics (i.e. mean, standard deviation, lower and upper quartiles, minimum and maximum values) computed after the pre-processing phase.

An exploratory data analysis was also performed to assess the significance of the internal (body) and external (environment) sensor data for the mood prediction task and to better set up the subsequent ML experiments. In particular, a correlation matrix for all sensor and mood-related variables was computed to measure the collinearity between features (Figure 2), with the aim of removing redundant predictors and reducing the training time. Theil's U, Pearson's R and the Correlation Ratio were used to measure the association between two categorical variables, two numeric variables and between a pair of categorical and numeric variables, respectively (Table 4). The correlation matrix plotted in Figure 2 showed a high correlation of categorical attributes such as *clouds* and *weather* as well as of *country* with *city* and *city* with *altitude*. Among these features, *weather*, *country* and *altitude* were eliminated. The correlation matrix also showed a certain degree of correlation between the mood variables Quality of Presentation and Understanding. Formulating the analysis as a multi-output classification problem, therefore, turned out to be a reasonable approach to explore such possible correlation. As a further step we also analyzed the scatter plots of the numeric predictors, from which no structured relationships emerged for most variables apart from weather-related attributes (i.e. temperature, humidity, wind and pressure) which showed some obvious linear relationships.

At the end of the pre-processing phase the final dataset used to build the classification models was composed by approximately 6000 examples described in terms of twelve features: heart rate, movement, temperature, humidity, pressure, wind, clouds, noise, city, hour_of_day, day_of_week and gender. The distribution of these examples among the three intensity levels (class values) for each emotional state is indicated in Table 3.

TABLE 3
Distribution of the examples among the class values

| Mood | Class 0 | Class 1 | Class 2 | Total |
|---|---|---|---|---|
| Activation | 944 | 3453 | 2030 | 6427 |
| Tiredness | 3600 | 1985 | 384 | 5969 |
| Pleasance | 232 | 2518 | 3677 | 6427 |
| Quality of Presentation | 169 | 2447 | 3353 | 5969 |
| Understanding | 181 | 2013 | 3775 | 5969 |

### 4.2 Experimental Settings

For the purpose of mood prediction we resorted to nine alternative supervised learning techniques implemented in scikit-learn [34] and Keras [35] and represented by Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, Multilayer Perceptron (MLP), Support Vector Machines (SVM), Gradient Boost, XGBoost and Long Short-Term Memory (LSTM) Networks.

The performance of these classifiers for each prediction task (single-output, multi-output, chain classification) was evaluated according to the following validation scheme. From the set of available examples we first derived by means of a stratified split a training and a test set containing, respectively, 70% and 30% of the instances. To select the most promising combination of parameters for each algorithm we then applied a randomized search in the parameters space by extracting 20 random combinations and evaluating them by means of 10-fold cross-validation applied on the training set. The parameters tuned for each technique, together with the interval of values randomly explored, is described in Table 5. Finally, we built the classification model on the training set with the best parameters achieved for each classifier, and evaluated its performance on the test set in terms of accuracy.

Prior to the experiments all the categorical input features were one-hot encoded whereas the numeric variables were standardized by means of the scikit-learn *StandardScaler* filter. Moreover, to account for the unbalanced ground truth labels affecting our datasets, as shown in Table 3, we leveraged the parameter *class weight* available for many algorithms (i.e. Decision Tree, Logistic Regression, Random Forest and SVM), in order to penalize more heavily the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3003736, IEEE Transactions on Affective Computing



TABLE 4
Measures used for correlation analysis

| Measure | Variable Type | Mathematical Formulation |
|---|---|---|
| Theil's U | Categorical vs Categorical | $U(X\|Y) = \frac{H(X)-H(X\|Y)}{H(X)}$ <br> where $H(X) = -\sum_x P_X(x)logP_X(x)$ <br> and $H(X\|Y) = -\sum_{x,y} P_{X,Y}(x,y)logP_{X\|Y}(x\|y)$ |
| Pearson's R | Numerical vs Numerical | $r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$ <br> where $n$ is the sample size, $x_i, y_i$ are the individual sample points $i$ <br> $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sample mean), and the same for $\bar{y}$ |
| Correlation Ratio | Categorical vs Numerical | $\eta^2 = \frac{\sum_x n_x(\overline{y}_x - \overline{y})^2}{\sum_{x,i}(y_{xi} - \overline{y})^2}$ <br> where $\overline{y}_x = \frac{\sum_i y_{xi}}{n_x}$ and $\overline{y} = \frac{\sum_x n_x \overline{y}_x}{\sum_x n_x}$, $n_x$ is the number of observations in category $x$ |

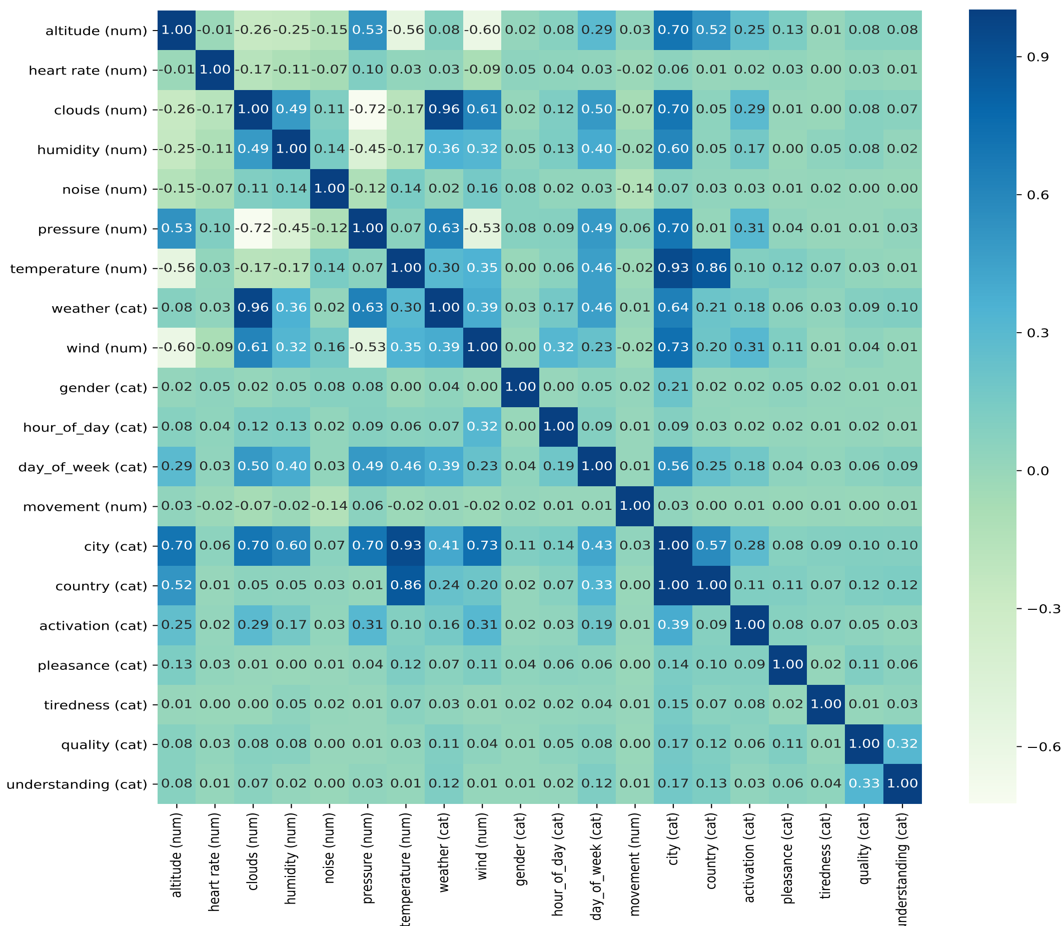


Fig. 2. Correlation matrix for all variables

misclassification of the minority classes examples. For the other classifiers we instead applied the Synthetic Minority Oversampling Technique (*SMOTE*), which oversamples the minority classes instances by synthesizing new observations from the existing ones.

Notice that, the number of available examples for each mood variable after the data cleaning phase determined a slight difference in the size of the training and the test sets in the prediction tasks. More specifically, the single-output models were obtained by using 6427 instances for Activation and Pleasance and 5969 observations for the other emotional states. The multi-output and the chain models were instead generated on the set of 5969 examples for which the measurements of all the input predictors were available for all the moods.

# 5 Results and Discussion

## 5.1 Single-Output Multi-Category Classification

The first prediction task was devoted to analyze the effectiveness of the set of body-related and environmental features in predicting the mood variables considered in-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3003736, IEEE Transactions on Affective Computing



TABLE 5
Parameters space for each classifier

| Method | Parameters |
|---|---|
| Decision Tree | min samples split: integer $[2, 10]$ |
| | max depth: integer $[2, 20]$ |
| KNN | n neighbors: integer $[2, 20]$ |
| Logistic Regression | penalty: $['l1', 'l2']$ |
| | C: float $[0.1, 2.0]$ |
| Random Forest | n estimators: integer $[10, 100]$ |
| | min samples split: integer $[2, 10]$ |
| | max depth: integer $[20, 70]$ |
| MLP | hidden layer sizes: $[(50, 50, 50), (50, 100, 50), (100, )]$ |
| | activation: ['tanh', 'relu'] |
| | solver: ['sgd', 'adam'] |
| | alpha: float $[0.00001, 0.001]$ |
| | learning rate: ['constant', 'adaptive'] |
| SVM | gamma: float $[0.1, 2.0]$ |
| | C: float $[0.1, 2.0]$ |
| | kernel: ['linear', 'rbf', 'poly'] |
| Gradient Boost | max depth: integer $[2, 10]$ |
| | max features: $["log2", "sqrt", None]$ |
| XGBoost | max depth: integer $[2, 10]$ |
| LSTM | hidden nodes: $[128, 256, 384, 512]$ |
| | optimizers: ['adam', 'rmsprop'] |
| | batch size: $[16, 32, 48, 64]$ |

dividually, therefore leading to five separate single-output multi-category classification problems.

The accuracy achieved on the test sets by training the classifiers for each emotional state taken independently is indicated in Figure 3 and Table 6. Among the alternative algorithms Logistic Regression consistently provided the worst outcomes. This result seems to suggest the non-linearity of the underlined classification regions, which can be hardly detected by classifiers, such as Logistic Regression, for which the decision boundary is linear. On the contrary, ensemble tree-based classifiers like Random Forest and Gradient Boost outperformed the other techniques. This achievement is in line with prior research studies which showed the usefulness of such algorithms for emotion recognition tasks [9], [15], [16]. More complex methods such as LSTM, which were proven to work well with the temporal dependencies characterizing sensor data [25], were unable to reach the performance achieved by the tree-based classifiers. The lower classification accuracy might be in this case motivated by the lack of a large amount of training examples, which is usually required by deep learning algorithms to yield superior results.

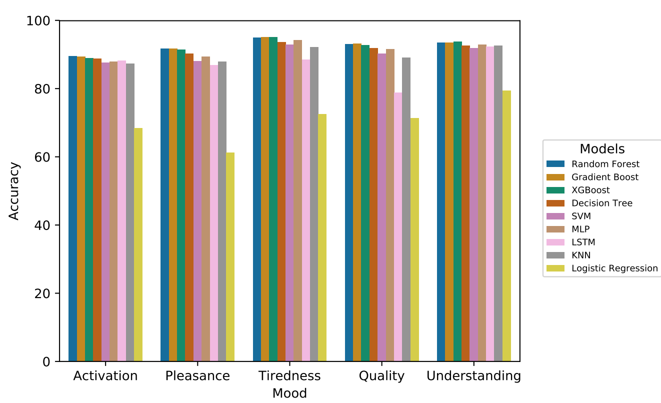


Fig. 3. Comparison of the accuracy of the single-output models

TABLE 6
Accuracy values for the single-output models

| Model | Activation | Tiredness | Pleasance | Quality | Understanding |
|---|---|---|---|---|---|
| Random Forest | 89.53% | 94.97% | 91.65% | 93.02% | 93.41% |
| Gradient Boost | 89.42% | 95.14% | 91.65% | 93.13% | 93.52% |
| XGBoost | 88.85% | 95.09% | 91.45% | 92.74% | 93.80% |
| MLP | 87.82% | 94.25% | 89.32% | 91.51% | 92.85% |
| Decision Tree | 88.75% | 93.63% | 90.25% | 91.79% | 92.52% |
| SVM | 87.66% | 92.91% | 87.97% | 90.23% | 91.79% |
| LSTM | 88.18% | 93.30% | 84.50% | 81.80% | 92.07% |
| KNN | 87.35% | 92.18% | 87.87% | 89.11% | 92.52% |
| Logistic Regression | 68.38% | 72.42% | 61.17% | 71.30% | 79.45% |

In terms of computational effort required, unsurprisingly tree-based methods were among the fastest algorithms as shown by Table 7, which indicates the average training time for building the models. Notice that, all experiments were performed on a machine with 8 GB RAM and 4 core CPU.

TABLE 7
Average training time (minutes) of the models

| Method | Single-output | Multi-output |
|---|---|---|
| Decision Tree | 0.00014427 | 0.00098226 |
| Random Forest | 0.00110438 | 0.00768893 |
| KNN | 0.00113072 | 0.00160625 |
| Logistic Regression | 0.00611656 | 0.05920169 |
| XGBoost | 0.01181077 | 0.03406380 |
| Gradient Boost | 0.01224895 | 0.04758920 |
| MLP | 0.03696476 | 0.14795474 |
| SVM | 0.11835161 | 0.17417982 |
| LSTM | 19.4107374 | 26.94515855 |

In order to evaluate the predictive ability of the input features involved in the well-performing ensemble tree-based classifiers we resorted to the Mean Decrease in Impurity (MDI) index, which measures the importance of a predictor in estimating the value of the target variable across all the trees in the ensemble. For the Random Forest classifier, for example, the MDI of a given input variable is defined as the total decrease in node impurity imputable to that feature, weighted by the probability of reaching that node and averaged over all the trees in the forest. The higher MDI, the more relevant the input variable.

Figure 4 displays the MDI of the features for each of the five moods computed for the Random Forest models. As one may observe *movement* and *heart rate* turned out to be, at a large extent, the most influential variables for all the prediction tasks. This result demonstrates that body-related (internal) signals are better predictors compared to external factors such as those referred to the individuals' surrounding scenario. Moreover, it confirms the appropriateness of including body signals when dealing with classroom setting use-cases as suggested by prior research studies [22], [23], [24], which also showed that on-body modalities are more robust sources for emotion classification "in the wild" [25], as in our case where data were collected directly in the participants natural environment.

Despite playing a less prominent role, *hour_of_day*, *humidity* and *temperature* deserve attention too, being the most important features among the external predictors, especially for what concerns Tiredness and Understanding. The discriminatory ability of *hour_of_day* is not surprising, since students tend to be more active and focused in the early

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3003736, IEEE Transactions on Affective Computing



hours of the day. On the other hand, the evidence on the environmental factors *humidity* and *temperature* is aligned with previous works, which associated inadequate classroom thermal conditions, such as high levels of humidity and high temperatures, with impaired academic performance due to lowered students concentration and increasing sleepiness.

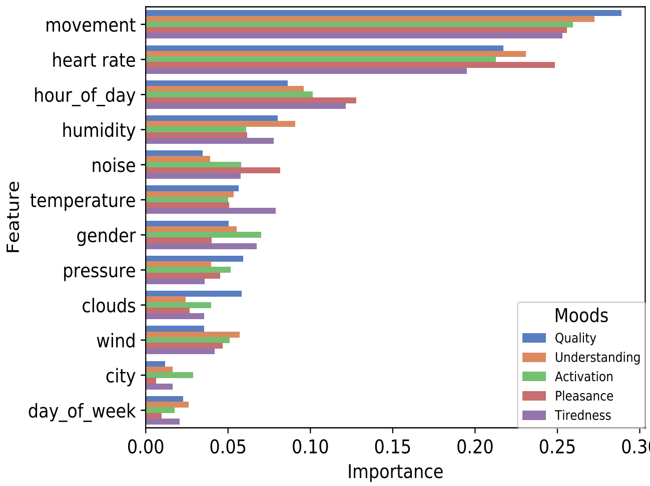


Fig. 4. Feature importance for the Random Forest single-output models

## 5.2 Multi-Output Multi-Category Classification

The outcome of the correlation analysis performed in the pre-processing phase suggested the existence of possible correlations between the target variables. To exploit such correlations a multi-output multi-category classification model was trained for each algorithm. As described in Section 3.2, in multi-output learning several classification tasks are solved at the same time, with the purpose of improving the learning process by leveraging their similarities [36]. The output is in our case a unique classification model which generates the predictions for all the mood variables simultaneously.

The results of this second analysis are illustrated in Table 8, which indicates the accuracy obtained on the test set by each classifier, together with the optimal parameters pointed out by the randomized grid search and used for training. Likewise in the single-output framework tree-based ensemble techniques, in general, and Random Forest, in particular, provided the highest accuracy among the competing techniques. Notice that, for these classifiers the models trained individually provided overall better predictions compared to the multi-output counterparts. The lower accuracy might be ascribed to the higher complexity of the learning process carried out in the multi-output scheme, in which the models are generated to concurrently predict all the moods, compared to the single-output framework in which they are specifically trained to fitting a given target.

Consistently with the former analysis, we investigated the relative importance of the input features in the outperforming tree-based multi-output models. The MDI scores sorted in decreasing order for the Random Forest algorithm are plotted in Figure 5. In accordance with the results of the individual models, *movement*, *heart rate* and *hour_of_day*

TABLE 8
Accuracy values for the multi-output models

| Model | Optimal Parameters | Accuracy |
|---|---|---|
| Random Forest | n_estimators=83<br>min_samples_split=4<br>max_depth=49 | 93.31% |
| Gradient Boost | max_depth=7<br>max_features='log2' | 93.24% |
| XGBoost | max_depth=9 | 92.91% |
| MLP | hidden_layer_sizes=(50, 100, 50)<br>activation='tanh'<br>solver='adam'<br>alpha=0.000895<br>learning_rate='constant | 92.28% |
| Decision Tree | min_samples_split=2<br>max_depth=15 | 91.96% |
| SVM | gamma=1.143<br>C=1.967<br>kernel='rbf' | 90.93% |
| LSTM | hidden nodes=512<br>optimizer='rmsprop'<br>batch size=48 | 88.51% |
| KNN | n_neighbors=5 | 75.53% |
| Logistic Regression | penalty='l1'<br>C=0.180 | 72.36% |

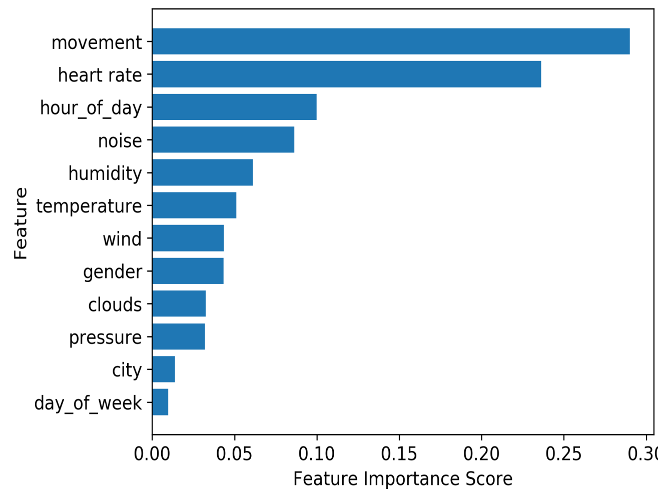


Fig. 5. Feature importance for the Random Forest multi-output model

proven to be highly significant in concurrently predicting the five emotional states. These last insights may help teachers in adapting their lecturing styles by taking, for instance, a break to perform some physical exercises and promote movement or by holding the lecture during the most favorable hours of the day, and could be also integrated into gamification activities in learning environments, which have been shown to improve students engagement and learning [37], [38], [39]. Moreover, they confirm the relevance attributed to these drivers by several studies, which demonstrated that sustained physical activity stimulates plasticity in the brain and, hence, enhance the learning capacity [40], [41], [42], or showed that most of the students enjoy taking classes in which active learning activities are organized [43], [44].

## 5.3 Multi-Category Chain Classification

To further explore the potential correlations among the mood variables a classifier chain model, depicted in Fig-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3003736, IEEE Transactions on Affective Computing



ure 6, was also generated, in which the first model in the chain was trained by using just the sensor data as input features and each subsequent model was built on the sensor data plus the mood variables formerly predicted along the chain.

Being aware of the computational complexity of finding the optimal sorting among the emotional states, as motivated in Section 3.2, we confined the attention to one plausible order, described in Figure 1. Specifically, the first model in the chain was generated to predict, based on the sensor data, the level of Activation, which was used to estimate the degree of Tiredness provided their inherent relationship. Tiredness was then exploited to predict Pleasance, given the reasonable influence that weariness exerts on the sense of happy satisfaction and enjoyment, which in its turn was included to explain the participants feedback on the Quality of Presentation. This last mood variable, finally, was deemed fundamental to estimate the levels of Understanding.

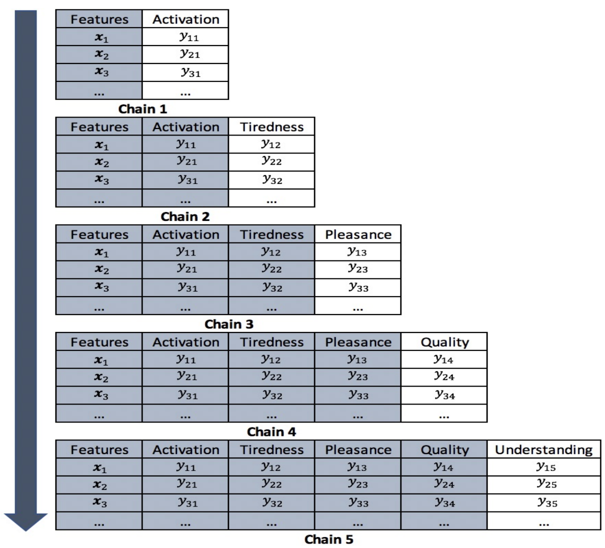


Fig. 6. Mood variables prediction by classifiers chain

The results achieved along the chain by the algorithms are depicted in Figure 7 and Table 9. As on one may notice, the classification accuracy on the test set consistently improves along the chain for all the methods. Moreover, yet starting from the models trained to predict the first mood (Activation) the accuracy obtained in the chain scheme by most of the algorithms are superior compared to the results of the multi-output framework. This achievement validates the plausibility of the order chosen for the emotional states (i.e. each mood included in the chain helps in the prediction of the following one) and confirms the existence of possible correlations among the targets which are effectively exploited by the chain classification model. The ensemble techniques based on decision trees emerged once again as best performers.

The feature importance in the chain models for the Random Forest algorithm is shown in Figure 8, which highlights the relevant contribution of Tiredness. This emerges also by the boosting in accuracy that this mood, used as predictor, brings for estimating the level of the subsequent Pleasance mood (Figure 6). In general, Tiredness and Activation turned out to be more informative features compared to environmental factors such as *temperature* and *noise*. Moreover, consistently with the correlation analysis Quality of Presentation was shown to be an important variable for explaining the level of Understanding.

TABLE 9
Accuracy values for the chain models

| Model | Chain 1 | Chain 2 | Chain 3 | Chain 4 | Chain 5 |
|---|---|---|---|---|---|
| Random Forest | 92.71% | 93.68% | 96.56% | 97.48% | 98.77% |
| Gradient Boost | 92.04% | 93.63% | 96.56% | 97.74% | 98.61% |
| XGBoost | 92.19% | 93.32% | 96.82% | 97.28% | 98.72% |
| MLP | 92.24% | 94.25% | 95.74% | 98.00% | 99.23% |
| Decision Tree | 91.88% | 91.68% | 94.61% | 96.56% | 97.33% |
| SVM | 92.30% | 93.58% | 95.79% | 97.02% | 98.72% |
| LSTM | 90.91% | 93.89% | 93.68% | 97.02% | 99.13% |
| KNN | 92.04% | 92.76% | 94.66% | 96.97% | 98.72% |
| Logistic Regression | 71.96% | 70.93% | 76.58% | 77.20% | 93.32% |

## 6 Conclusions

The paper described the use of a smartwatch-based system to collect sensor data and predict various mood variables that are relevant in a classroom lecture setting. Our analysis relied on three classification tasks resorting to alternative frameworks, represented by the single-output, multi-output and the classification chain paradigms.

Our results showed that body signals such as *heart rate* and *movement* are better predictors compared to the external environmental variables. In terms of machine learning algorithms used for classification, ensemble tree-based methods yielded superior performances over more complex methods such as LSTM. At a more general level, our work demonstrated that the combined use of sensors and machine learning is a promising approach for the study of emotions in a classroom setting and can open new opportunities in other application areas of emotion recognition.

Despite the outcome of our research empirically showed the opportunity of collecting physiological and environmental sensor data in a classroom setting to effectively predict relevant emotions through machine learning, we are aware of certain limitations. From one side, the dataset collected and analyzed is relatively small in terms of number of participants (30 people) and study period (5 different days of lectures/workshops). Due to the limited number of samples per attendee, we were constrained to build a general predictive model by combining the data from all the participants, rather than a personalized model suited to each individual. In future works we intend to cover a larger sample size and run larger scale experiments, in order to generate models customized to specific users or groups and investigate the differences in their behavior and learning process. The availability of a larger sample will also allow us to more adequately apply and explore deep learning algorithms. On the other hand, our research considered only the sensor modality. As future developments we plan to include other forms of modalities such as voice, body language and facial expressions, besides physiological and environmental sensor data which have been the focus of the present work.

While the aforementioned limitations exist, our study has nevertheless shown a novel approach for predicting mood states in a classroom setting which can be extended to

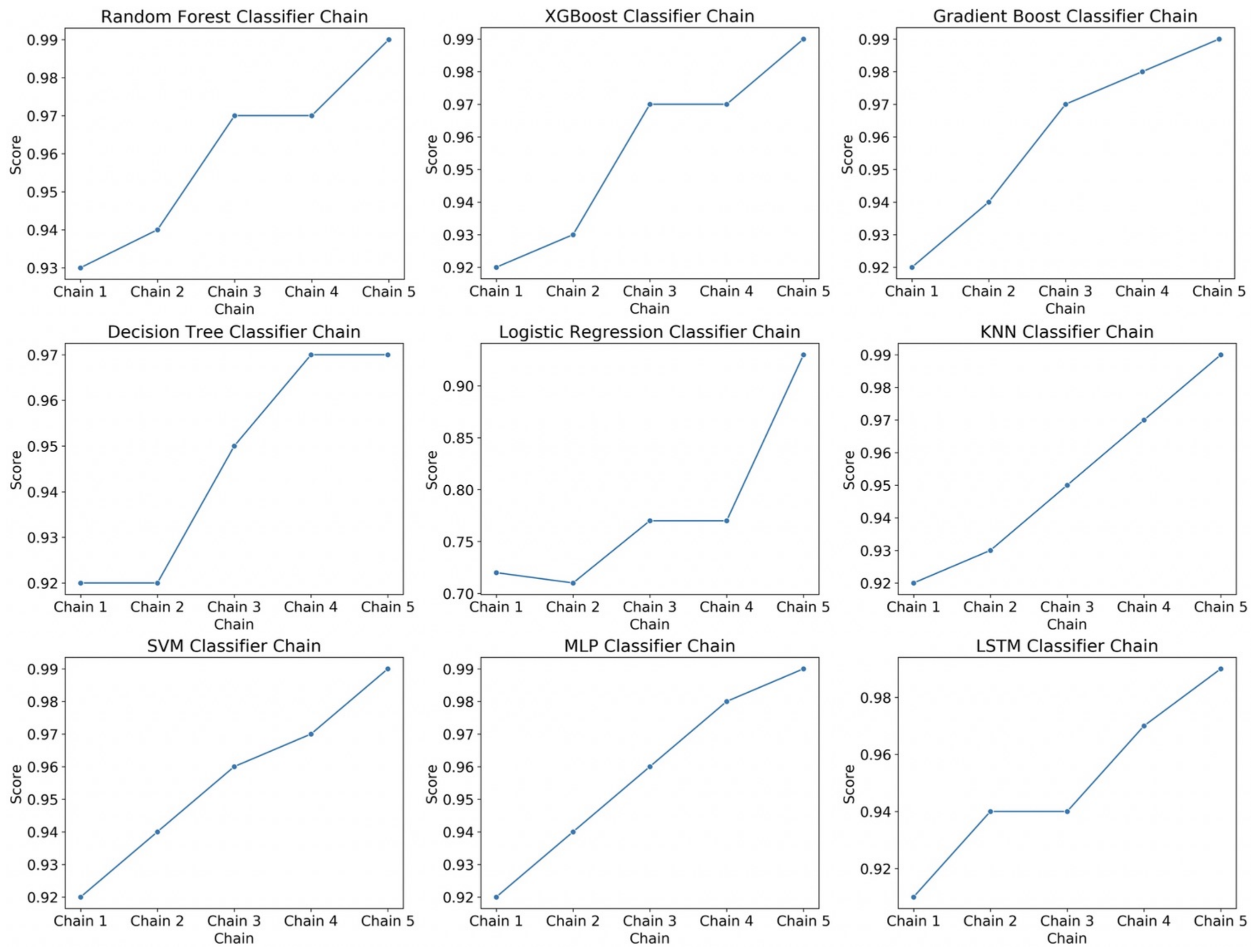


Fig. 7. Accuracy of the algorithms along the classification chain

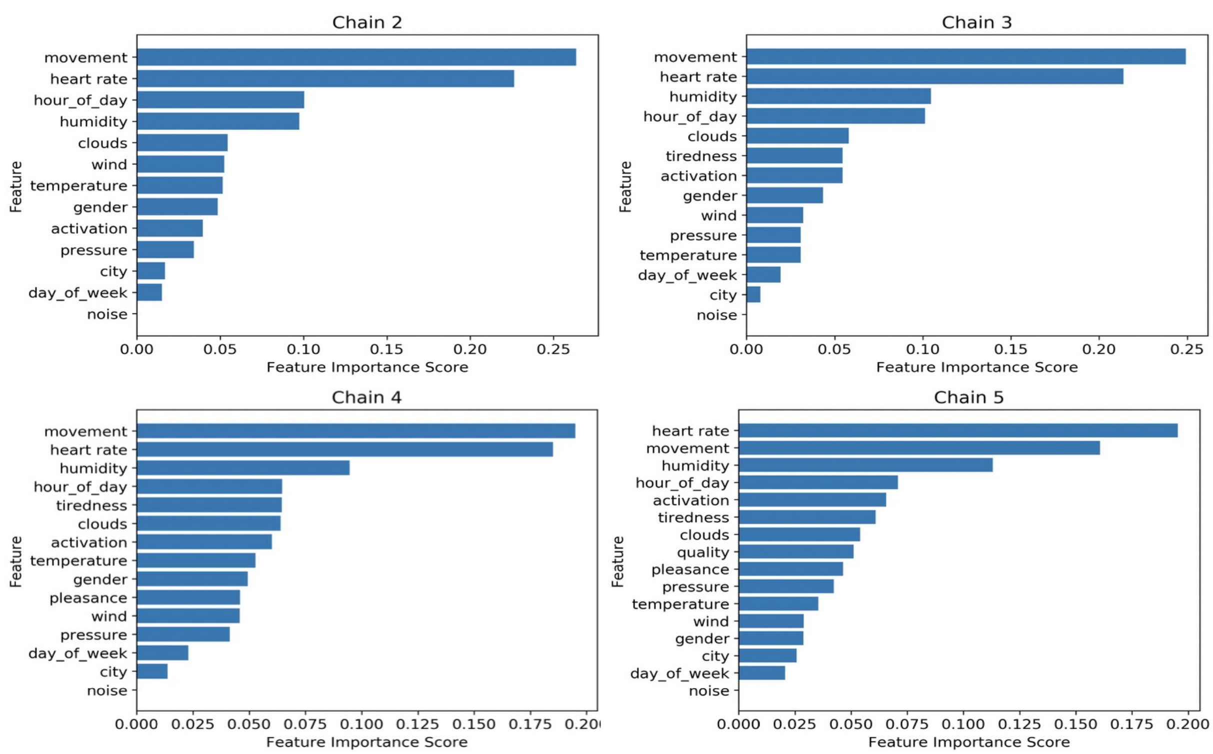


Fig. 8. Feature importance for the Random Forest chain model

other use-cases. Measuring emotions and becoming aware of one's emotions can be tremendously useful in other applications. On the one hand, call center agents can get direct input about the emotional state of their customers on the phone as well as of their own state to dynamically adjust their response [45]. On the other hand, becoming aware and knowing one's own emotions will fundamentally change the behavior of an individual. For instance, people frequently are not aware of their own anger building up; getting an early warning sign alerting them to this build-up will help them calm down [46]. These are just two examples of what we call the "virtual mirroring" effect, where making people aware of their own behavior will dramatically change this behavior [47]. Our study is therefore an important building block for developing such capabilities. With the imperative need to address the ever-increasing complexity of tasks and problems in various application domains nowadays, significant contributions are required and our work provides a step in this direction.

## Acknowledgments

This work has been supported by Philips Lighting as part of an overall MIT "Grand Challenges in Lighting" research grant.

## References


[1] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, pp. 1192–1209, 2013.

[2] O. Aouedi, M. A. Bach Tobji, and A. Abraham, "An ensemble of deep auto-encoders for healthcare monitoring," in *Hybrid Intelligent Systems* (A. M. Madureira, A. Abraham, N. Gandhi, and M. L. Varela, eds.), pp. 96–105, Springer International Publishing, 2020.

[3] S. Dey, Q. Ye, and S. Sampalli, "A machine learning based intrusion detection scheme for data fusion in mobile clouds involving heterogeneous client networks," *Information Fusion*, vol. 49, pp. 205 – 215, 2019.

[4] A. Elsaeidy, K. S. Munasinghe, D. Sharma, and A. Jamalipour, "Intrusion detection in smart cities using restricted boltzmann machines," *Journal of Network and Computer Applications*, vol. 135, pp. 76 – 83, 2019.

[5] M. Montebello, "Assisting education through real-time learner analytics," in *2018 IEEE Frontiers in Education Conference (FIE)*, (Los Alamitos, CA, USA), pp. 1–7, IEEE Computer Society, Oct. 2018.

[6] M. Ciolacu, A. Tehrani, L. Binder, and P. Svasta, "Education 4.0 - artificial intelligence assisted higher education: Early recognition system with machine learning to support students' success," *2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging, SIITME 2018 - Proceedings*, pp. 23–30, 2019.

[7] M. Ciolacu, A. F. Tehrani, R. Beer, and H. Popp, "Education 4.0 — fostering student's performance with machine learning methods," *2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME)*, pp. 438–443, 2017.

[8] P. A. Gloor, A. F. Colladon, F. Grippa, P. Budner, and J. Eirich, "Aristotle said "happiness is a state of activity" — predicting mood through body sensing with smartwatches," *Journal of Systems Science and Systems Engineering*, vol. 27, pp. 586–612, Oct. 2018.

[9] E. Kanjo, E. M. Younis, and N. Sherkat, "Towards unravelling the relationship between on-body, environmental and emotion data using sensor information fusion approach," *Information Fusion*, vol. 40, pp. 18 – 31, 2018.

[10] P. Budner, J. Eirich, and P. A. Gloor, "Making you happy makes me happy - measuring individual mood with smartwatches," *Computing Research Repository*, 2017.

[11] M. Cabanac, "What is emotion?," *Behavioural Processes*, vol. 60, no. 2, pp. 69 – 83, 2002.

[12] D. Schacter, D. T. Gilbert, and D. M. Wegner, *Psychology (2nd Edition)*. New York: Worth, 2011.

[13] P. Ekman, "Basic emotions.," in *Handbook of cognition and emotion.*, pp. 45–60, New York, NY, US: John Wiley & Sons Ltd, 1999.

[14] R. Lazarus and B. Lazarus, *Passion and Reason: Making Sense of Our Emotions*. Oxford University Press, 1994.

[15] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, (New York, NY, USA), p. 3–10, Association for Computing Machinery, 2016.

[16] M. Adibuzzaman, N. Jain, N. Steinhafel, M. Haque, F. Ahmed, S. I. Ahamed, and R. Love, "Towards in situ affect detection in mobile devices: A multimodal approach," in *Proceedings of the 2013 Research in Adaptive and Convergent Systems*, RACS '13, (New York, NY, USA), p. 454–460, Association for Computing Machinery, 2013.

[17] Z. Guendil, Z. Lachiri, C. Maaoui, and A. Pruski, "Multiresolution framework for emotion sensing in physiological signals," *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pp. 793–797, 2016.

[18] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," *IEEE Transactions on Affective Computing*, vol. 9, pp. 227–239, Apr. 2018.

[19] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham, "Weka: Practical machine learning tools and techniques with java implementations," 1999.

[20] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, (New York, NY, USA), pp. 97–104, ACM, 2016.

[21] J. Bughin, E. Hazan, S. Ramaswamy, M. Chui, T. Allas, P. Dahlstrom, N. Henke, and M. Trench, "Artificial intelligence: the next digital frontier?," 2017.

[22] O. Herrera-Alcántara, A. Y. Barrera-Animas, M. González-Mendoza, and F. Castro-Espinoza, "Monitoring Student Activities with Smartwatches: On the Academic Performance Enhancement," *Sensors*, vol. 19, p. 1605, Apr. 2019.

[23] H. Wang, Y. Li, X. Hu, Y. Yang, Z. Meng, and K.-M. Chang, "Using eeg to improve massive open online courses feedback interaction," *CEUR Workshop Proceedings*, vol. 1009, pp. 59–66, Jan. 2013.

[24] S. Liu, Y. Chen, H. Huang, L. Xiao, and X. Hei, "Towards smart educational recommendations with reinforcement learning in classroom," in *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pp. 1079–1084, Dec. 2018.

[25] E. Kanjo, E. M. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Information Fusion*, vol. 49, pp. 46 – 56, 2019.

[26] J. E. Posner, A. A. Prekates, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, vol. 17, no. 3, p. 715–734, 2005.

[27] N. Perkinson-Gloor, S. Lemola, and A. Grob, "Sleep duration, positive attitude toward life, and academic achievement: The role of daytime tiredness, behavioral persistence, and school start times," *Journal of Adolescence*, vol. 36, no. 2, pp. 311 – 318, 2013.

[28] D. F. McCaffrey, J. R. Lockwood, D. M. Koretz, and L. S. Hamilton, *Evaluating Value-Added Models for Teacher Accountability*. RAND Corporation, 2003.

[29] B. Rowan, R. Correnti, and R. Miller, *What Large-scale, Survey Research Tells Us about Teacher Effects on Student Achievement: Insights from the Prospects Study of Elementary Schools*. CPRE research reports series, Consortium for Policy Research in Education, 2002.

[30] "Measuring Student Learning." Accessed on June 30, 2019 [Online]. Available: https://teaching.cornell.edu/teaching-resources/assessment-evaluation/measuring-student-learning.

[31] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, pp. 43–55, Jan. 2015.

[32] P. Lavrakas, *Encyclopedia of Survey Research Methods*. A SAGE reference publication, SAGE Publications, 2008.

[33] "Reverse Geocoding (Address Lookup)." Accessed on Feb. 13, 2019 [Online]. Available: https://developers.google.com/maps/documentation/geocoding.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg,

J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] F. Chollet *et al.*, "Keras." https://keras.io, 2015.

[36] D. Xu, Y. Shi, I. W.-H. Tsang, Y.-S. Ong, C. Gong, and X. Shen, "A survey on multi-output learning," *IEEE Transactions on neural networks and learning systems*, 2019.

[37] M. Kuo, T. Chuang, Y. Su, T. Chen, and B. Cheng, "What are the better gamification tools for elementary school teachers?," in *2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 346–349, July 2018.

[38] A. Domínguez, J. S. de Navarrete, L. de Marcos, L. Fernández-Sanz, C. Pagés, and J.-J. Martínez-Herráiz, "Gamifying learning experiences: Practical implications and outcomes," *Computers and Education*, vol. 63, pp. 380 – 392, 2013.

[39] G. Barata, S. Gama, J. Jorge, and D. Gonçalves, "Gamification for smarter learning: tales from the trenches," *Smart Learning Environments*, vol. 2, p. 10, May 2015.

[40] S. Vaynman and F. Gomez-Pinilla, "License to run: Exercise impacts functional plasticity in the intact and injured central nervous system by using neurotrophins," *Neurorehabilitation and Neural Repair*, vol. 19, no. 4, pp. 283–295, 2005.

[41] A. Kandola, J. Hendrikse, P. J. Lucassen, and M. Yücel, "Aerobic exercise as a tool to improve hippocampal plasticity and function in humans: Practical implications for mental health treatment," *Frontiers in Human Neuroscience*, vol. 10, p. 373, 2016.

[42] H. W. K. III and H. D. Cook, *Educating the Student Body: Taking Physical Activity and Physical Education to School*. Washington, DC: The National Academies Press, 2013.

[43] I. Janna, K. Christopher, S. Barbara, J. Kalle, K. Joseph, L. Jari, and S.-A. Katariina, "Science classroom activities and student situational engagement," *International Journal of Science Education*, vol. 41, no. 3, pp. 316–329, 2019.

[44] X. Gordy, L. Zhang, A. Sullivan, J. Bailey, and E. Carr, "Teaching and learning in an active learning classroom: A mixed-methods empirical cohort study of dental hygiene students," *Journal of dental education*, vol. 83, no. 3, pp. 342–350, 2019.

[45] V. A. Petrushin, "Detecting emotion in voice signals in a call center," May 2011.

[46] J. Dobson, *Emotions: Can You Trust Them?: The Best-Selling Guide to Understanding and Managing Your Feelings of Anger, Guilt, Self-Awareness and Love*. Gospel Light, 2003.

[47] P. Gloor, A. F. Colladon, G. Giacomelli, T. Saran, and F. Grippa, "The impact of virtual mirroring on customer satisfaction.," *Journal of Business Research*, vol. 75, pp. 67–76, 2017.

**Keith April Araño** is a PhD student at the Department of Management, Economics and Industrial Engineering of Politecnico di Milano in Milan, Italy. She is currently a visiting student at the Center for Collective Intelligence of MIT's Sloan School of Management. Her primary research interest is on the application of various Artificial Intelligence (AI) techniques in the tasks of sentiment analysis and emotion recognition. In her PhD research, she aims to investigate how multimodal data- sensor, text, audio, and visual can be exploited to aid the tasks of sentiment and emotion detection through the application of different AI techniques such as machine learning, deep learning, and natural language processing. She received her Bachelor of Science degree in Industrial Engineering from the University of the Philippines and went on to work as a SAP technical consultant in IBM Solutions Delivery Inc.. She was awarded an Erasmus Mundus scholarship and obtained her master's degree in Industrial Management from the Universidad Politecnica de Madrid, Politecnico di Milano, and Heriot Watt University, where she did her master thesis on estimating personal pollution exposure by combining outdoor pollution and indoor pollution through sensor networks.

**Peter Gloor** is a Research Scientist at the Center for Collective Intelligence at MIT's Sloan School of Management where he leads a project exploring Collaborative Innovation Networks (COIN). He is also Founder and Chief Creative Officer of a software company galaxyadvisors where he puts his academic insights to practical use, helping clients to coolhunt by analyzing social networking patterns on the Internet – spot the next big thing by finding the trendsetters, and to coolfarm – increase organizational happiness, creativity and performance through workforce analytics. In addition, Peter is a Honorary Professor at University of Cologne and a Honorary Professor at Jilin University, Changchun China. He has also taught classes at University of Bamberg, Universidad Cattolica, Santiago de Chile, Aalto University Helsinki, University of Rome Tor Vergata, and University of Applied Sciences Northwestern Switzerland. Previously, Peter was a Partner with Deloitte Consulting, leading its e-Business practice for Europe, a Partner with PricewaterhouseCoopers and the section leader for software engineering at UBS. He was a post-doctoral fellow at the MIT Lab for Computer Science in the Advanced Networking Architecture group, working on hypertext well before the Web emerged and creating the multimedia CD "Animated Algorithms" (published by MIT Press). He obtained a Ph.D in Computer Science from the University of Zurich, and a Master's degree (diploma) in Mathematics also from the University of Zurich. He has published several books and papers in high-rank international journals.

**Carlotta Orsenigo** Carlotta Orsenigo is associate professor of computer science at Politecnico di Milano where she teaches courses in Optimization and Machine Learning. She is co-director of the International Master in Business Analytics and Big Data at MIP-Politecnico di Milano and founder an head of research of Aiblooms, a spin-off of Politecnico di Milano devoted to the development of software based on machine learning and artificial intelligence techniques. Her research interests include the development of novel methods and algorithms for pattern recognition, dimensionality reduction and text analytics, and their application to problems arising in several contexts such as biolife sciences, innovation and entrepreneurial studies.

**Carlo Vercellis** is a full professor of Machine Learning at Politecnico di Milano in Milan, Italy. He currently heads *door*, the Machine Learning and Big Data Analytics research group, and the Observatory on Big Data Analytics & Business Intelligence. He is also the director of the International Master in Business Analytics and Big Data at MIP-Politecnico di Milano. His research interests include machine learning, focusing on deep learning, support vector machines, nonlinear dimensionality reduction, sentiment analysis; application of predictive methods to marketing, social networks, internet of things, finance, biolife science; optimization models, with applications to supply chain and revenue management. His past research also dealth with design and analysis of algorithms for combinatorial optimization, project management, and transportation models. He has authored more than 80 scientific writings, among which 6 books and more than 60 papers in high-rank international journals and books. He has been an associate editor of several journals and has coordinated several national and international research programs funded by EEC, CNR, MIUR and private companies. He developed projects with several companies on machine learning, predictive analytics, big data, sentiment analysis, marketing, risk analysis, supply chain optimization, and quality control. Previously, after his graduation in Mathematics at Università degli Studi di Milano, he has been with the National Research Council (CNR), Bocconi University and Università degli Studi di Milano.