IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 March 2023, accepted 3 May 2023, date of publication 24 May 2023, date of current version 23 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3279720

RESEARCH ARTICLE

Affect Recognition in Muscular Response Signals

MATTHIAS BOEKER[1], PETTER JAKOBSEN[2], MICHAEL A. RIEGLER[1,3], LENA ANTONSEN STABELL[2,4], OLE BERNT FASMER[4], PÅL HALVORSEN[1,3], AND HUGO LEWI HAMMER[1,3]

[1]Department of Holistic Systems, SimulaMet, 0167 Oslo, Norway
[2]NORMENT, Division of Psychiatry, Haukeland University Hospital, 5021 Bergen, Norway
[3]Department of Computer Science, Oslo Metropolitan University, 0167 Oslo, Norway
[4]Department of Clinical Medicine, University of Bergen, 5036 Bergen, Norway

Corresponding author: Matthias Boeker (matthias@simula.no)

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT This study investigated the potential of recognising arousal in motor activity collected by wrist-worn accelerometers. We hypothesise that emotional arousal emerges from the generalised central nervous system which embeds affective states within motor activity. We formulate arousal detection as a statistical problem of separating two sets - motor activity under emotional arousal and motor activity without arousal. We propose a novel test regime based on machine learning assuming that the two sets can be distinguished if a machine learning classifier can separate the sets better than random guessing. To increase the statistical power of the testing regime, the performance of the classifiers is evaluated in a cross-validation framework, and to test if the classifiers perform better than random guessing, a repeated cross-validation corrected t-test is used. The classifiers were evaluated on the basis of accuracy and Matthew's correlation coefficient. The suggested procedures were further compared against a traditional multivariate paired Hotelling's T-squared test. The classifiers achieved an accuracy of about 60%, and according to the proposed t-test were significantly better than random guessing. The suggested test regime demonstrated higher statistical power than Hotelling's T-squared test, and we conclude that we can distinguish between motor activity under emotional arousal and without it.

INDEX TERMS Affect recognition, emotion detection, motor activity, arousal, soccer, machine learning, hypothesis testing, time series analysis.

I. INTRODUCTION

The nervous shaking of a hand, sweaty palms, and tight jaws are indicators of human affection. While we are often able to keep a straight face during emotional trigger scenarios, the body reveals an instant motor expression [1], [2]. The research field of affect recognition focusses on exposing affective states within physiological signals which allows computers to recognise affect, known as affect computing coined by Picard [3]. In this work, we focus on motor activity as a physiological signal, how it embeds affective states, and how we can recognise these states. Affect recognition can generally be distinguished between behavioural recognition and recognition based on physiological signals [4]. Behavioural recognition focusses on facial expression, eye tracking, and interaction with the smartphone or text. Physiological signals cover sensors of various biological reactions, such as heart rate, blood volume pulse, skin responses, respiration, and brain activity. Most of the enumerated signals capture the affective state through instant physiological expression. Emotions originate or result in changes in the physiological signal [2]. These signals can be investigated in unimodal or multimodal analysis [5], [2], [4]. Unimodal analysis takes into account only one indicator of affect expression, like heart rate alone. Consequently, the multimodal analysis combines various indicators. In recent years, studies have shown that multimodal analysis performs better than unimodal analysis [6].

The rapid technological development of wearable sensors and their continuous integration into commercial and private

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

use leads to large collections of physiological data. Affect recognition can be seen as a pattern recognition task that first identifies and then classifies or quantifies the affective state within a signal. Although affective computing originated in the early nineteen-nineties, the task of affect recognition remains challenging. Affect is a psychological construct defined by Russell J. and others, and a weakly defined latent artefact-like affect is difficult to extract from often noisy sensor signals [6], [7].

The ground truth for supervised classification tasks is hard to obtain. Experiments to measure affect range from strict laboratory settings to more real-life scenarios, depending on the research question. However, the experimental results are difficult to transfer to real situations, whereas experiments in real situations lead to more noise and difficulty in determining the ground truth. This partly impedes the generalisability of the methods and models researched [6]. In addition, technical difficulties arise with data collection. The choice of appropriate sensors is concerned with intrusiveness, costs, and noise [6]. Affect computing also combines many research fields such as signal processing, image compression, computer vision, texture modelling, statistical physics, machine learning, and cognitive psychology [5], [6], [8].

This work presents three main contributions. We studied the potential for affect recognition in motor activity collected by wrist-worn accelerometers. We hypothesise that motor activity embeds affective states, like the other physiological signals mentioned above. Our hypothesis is grounded in an evidence-based theory postulating that aroused activation emerges from a primitive and elementary mammalian neural force, as generalised central nervous system (CNS) arousal. This system regulates behavioural reactions to environmental developments, expresses emotional states, and is observable as both autonomic and motoric activation [9], [10]. Furthermore, motor activity has previously been proven to be a beneficial tool for categorising various mood disorders. Pathological mood states, such as depression and mania, could be considered conditions of sustained emotional reactions [11], [12]. The first contribution and the main goal of this work is to demonstrate that the muscular response emerging from the expression of emotional arousal can be measured in motor activity data, which to our knowledge has never been investigated before. Our second contribution is an experiment we conducted to collect data and prove our hypothesis. We collected the motor activity of ten soccer fans during a live television soccer game, all gathered at the same location. The choice of a soccer game as an emotional stimulus allowed us to create a closed environment for the experiment that was still as natural as possible. This experimental design can be easily extended, and other physiological signals can be measured. We defined a scored goal in the soccer game as an intense emotional stimulus to a fan, which is confirmed in other studies [13]. We assumed to recognise the affect within the motor activity if we managed to segregate the intervals of goals and intervals without any noteworthy events.

Our aim is to significantly differentiate between emotionally stimulated intervals of motor activity and control intervals. Accordingly, the third main contribution is a novel test regime based on machine learning. We assume that a machine learning classifier that performs significantly better than random guessing will make a correspondingly significant distinction between two sets. We introduced a repeated lower-bound cross-validated corrected t-test for classification performance metrics. The t-test uses the correction term introduced by Bengio and Grandvalet [14]. We applied different machine learning methods for binary classification. Due to a small sample size of 80 samples, we chose k-fold cross-validation to assess the classification results. The multivariate Hotelling's T-squared test was used as a baseline for our proposed test regime. The code for the data pre-processing, experiments, and statistical testing regime is available on Github [15]. The data is available through the Open Science Framework [16].

In summary, the main contributions are as follows.

- We suggest that the CNS emerges arousal as an expression of emotional states and showed that this arousal is measurable in motor activity.
- We conducted a real-world experiment in a closed environment to measure motor activity in response to emotional stimuli. The motor activity of ten soccer fans watching a soccer game (from their favourite team) live on television is recorded and the data is made publicly available.
- We contribute with a novel statistical test regime based on machine learning to test if two sets of data are outcomes from different statistical distributions.

In the remainder of the paper, we will first discuss concepts of affect recognition and related work, followed by a detailed description of our method and experiments. Finally, we discuss and present the results, followed by the conclusions and future work.

II. CONCEPTS OF AFFECT RECOGNITION

Before moving on to the topic of affect recognition, we introduce the expressions *affect* and *emotion*. Although terminology is often used interchangeably [2], we would like to use a consistent definition for each of the expressions. Russell J. defines *affect* as a neurophysiological state that is a primitive, non-reflective, and non-directed feeling [7], a definition that resembles Pfaff's hypothesis about generalised CNS arousal [9], [10]. Affects can be linked to an object, and thus become directed toward the object. Russell's example is the encounter of a bear with a person. The person feels fear, upset, and discomfort caused by an object, the bear. Once the affect is cognitively processed and directed at an object, it becomes an attributed affect [7]. A compound of attributed affects can build an *emotion* [1].

Generally, the various theoretical models of emotions are divided into categorical or dimensional models. Categorical models state that emotions can be organised into numerous groups, a topic thoroughly discussed in the literature,

throughout the ages. The ancient Roman philosopher Cicero defined four categories of emotions, desire, delight, fear, and distress [17]. Ekman, the pioneer of modern-time emotion studies, argues for six basic emotions: anger, fear, sadness, enjoyment, disgust, and surprise. The different basic emotions are defined according to "the appraisal of a current event which is influenced by one's past history" [18]. A three-dimensional model projects emotions onto a multidimensional hyperplane. The two-dimensional circumplex model suggested by Russell [7] has met great popularity and application within the field of science and technology studies [2], [19], [20], [21], [22]. Especially in the field of machine learning within affect recognition, the model has been used to create classification problems [20], [23], [24], [25]. The two dimensions of the circumplex models are defined as valence and arousal. Valency refers to an affective quality that ranges from displeasure to pleasure. The arousal dimension is the feeling of mobilisation and energy and ranges from unconscious sleep to energetic enthusiasm [7]. Figure 1 illustrates the two-dimensional circumplex model.

In an elementary biological understanding, emotions and arousal are outcomes of ecological inputs assessed and processed in the brain. Various centres in the cerebral cortex receive visual, auditory, and olfactory signals, as well as taste and skin sensations, and forward these sensory impressions to the limbic brain structure for evaluation [26]. The limbic system's function is to institute emotional state, initiate behaviour, recall learnt experiences, and store new memories [27]. The established emotional state is equal to the level of generalised CNS arousal, and this is necessarily a tremendously responsive system, to enable escaping potential dangers [10]. Within the brain, the CNS communicates through neurons and hormones. Outside the brain, the CNS communicates through the nerves of the peripheral nervous system (PNS), both for the reception of sensory impressions and for the generation of responsive actions. The hypothalamus, a tiny area located deep in the brain, is the central communication centre of the nervous system. Its main objective is to keep bodily functions and rhythms stable and safe. Generalised CNS arousal communicates bodily functions through two branches of the PNS, the autonomic (ANS) and somatic (SNS) nervous systems. ANS regulates involuntary muscles in the body and makes generalised CNS arousal observable in heart rate, blood volume pulse, skin conduction, and respiration. SNS controls voluntary muscles, therefore, CNS arousal is recognisable in motor activity [26]. Consequently, the relationship between generalised CNS arousal observed in SNS (motor activity) and ANS (skin conductance and heart rate) has great potential for multimodal analyses [28].

III. RELATED WORK

Emotion recognition encompasses a wide range of elicitation methods, especially behavioural expressions that are only remotely related to our work. Therefore, we focus on the literature on emotion recognition based on physiological signals. Already in the early nineteen nineties, statistical approaches showed that different emotions can be elicited from physiological signals. Levenson et al. demonstrated that the autonomic nervous system differs for different emotions. The authors applied multivariate analysis of variance (MANOVA) on heart rate, skin conductance, finger pulse transmission time, and finger pulse amplitude data [29]. Pecchinenda and Smith investigated skin conductance during complex problem solving and identified statistically significant differences for mean changes in skin conductance [30]. Scheirer used Hidden Markov models to detect periods of frustration in skin conductivity and blood volume, which performed significantly better than random guessing [31]. Vrana [32] conducted an experiment on negative emotions, where participants were exposed to different images and their EMG, heart rate, and skin conductance level was recorded. The aim was to distinguish different negative emotional contexts. Using an ANOVA model, it was shown that disgust and anger-triggering images could be distinguished according to EMG. Heart rate was significantly higher for disgust, anger, and joy than for pleasant images [32].

With the rise of machine learning, researchers can explore more complex relationships within physiological signals. Especially in research on electroencephalography (EEG), machine learning algorithms have been applied. Mehmood et al. extracted the Hjorth parameters of the EEG [33]. The Hjort parameters are statistical properties of an EEG signal, namely activity, mobility, and complexity [34]. The authors applied a 10-fold cross-validation for support vector machine (SVM) and K-nearest neighbour algorithms (k-NN) and achieved accuracy scores around 50% [33]. The authors spared to report the standard deviation, hence it is not certain if the classifiers actually performed better than random guessing. In particular, because of this case, we introduced a lower-bound statistical test for the classification of performance scores in our study. Zheng et al. aimed to find EEG patterns that were stable between sessions, as well as common across subjects. SVM and Graph regularised extreme learning machine (GEM) were applied to extracted EEG features, like power spectral density. GEM outperformed SVM, and based on feature extraction, they show the importance of different features for different emotions [35]. One observable physiological change that occurs, e.g. in exciting or fearful situations, is the respiration rate. Wu et al. segmented the respiration signal according to five distinct emotions. The respiration signals of the 33 participants were recorded while the participants were exposed to emotionally related video clips. Time series features such as approximate entropy, root mean square, complexity, etc. were extracted from the respiration signal. The authors applied a KNN classifier and a probabilistic neural network to distinguish between various emotions [36]. As mentioned earlier, [6] showed that the multimodal approach performs better than the univariate approach. Still, it is important to investigate the

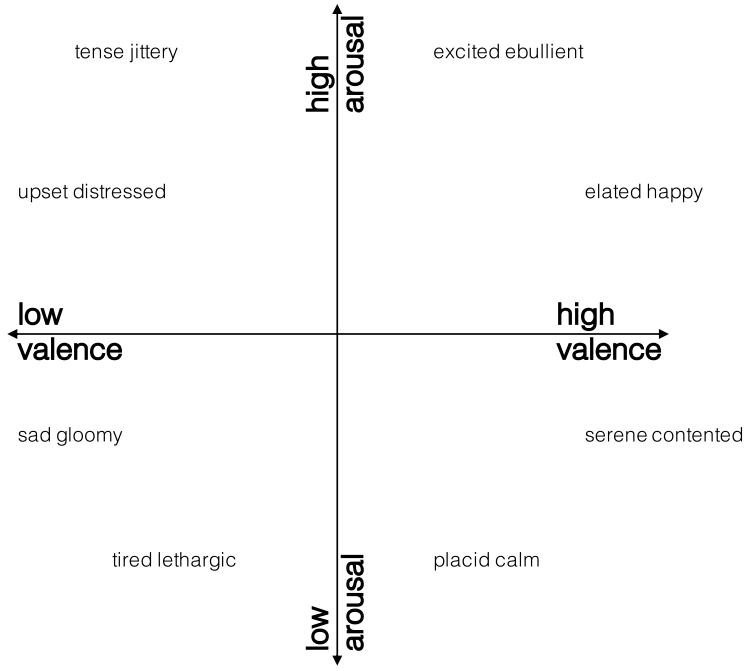


FIGURE 1. The two-dimensional circumplex model by Russell. Emotions can be understood as points in a two-dimensional space spanned by valence and arousal. The illustration is inspired by the original by Russell [7].

potential affective expression within a single physiological signal to include them or exclude them from a multimodal model.

Kim et al. studied the possibility of recognising valence and arousal within EMG, ECG, skin conductivity, and respiration changes. Emotions were triggered by music. The study applied linear discriminant analysis and an emotion-specific multilevel dichotomous classification [37]. Verma et al. compared a multimodal approach based on EEG with 32 channels and EEG combined with various physiological signals. Twelve different emotions were constructed based on their location in the two-dimensional circumplex model. Discrete Wavelet Transformation was used for feature extraction. The classifications were performed by SVM, Multilayer perceptron, KNN, and meta multiclass. When physiological signals were included, accuracy improved only slightly. As the authors reported, their approach achieved an accuracy of 81.4%, and outperformed similar approaches on the same dataset, which were 66.7% and 68.5%, respectively [38].

So far, we have introduced related work that does not incorporate motor activity as a physiological signal to recognise affects. However, motor activity recordings have been applied as a baseline for affect recognition during physical activity [2], [39]. Moreover, motor activity has been applied for stress detection, where accelerometer data from smartphones was used to identify stressful periods. Ciman et al. and Garcia et al. showed that stress could be detected based on smartphone accelerometers [40], [41]. However, to the best of our knowledge, we have not seen any studies that investigated affect recognition in wrist-worn piezoelectric accelerometers.

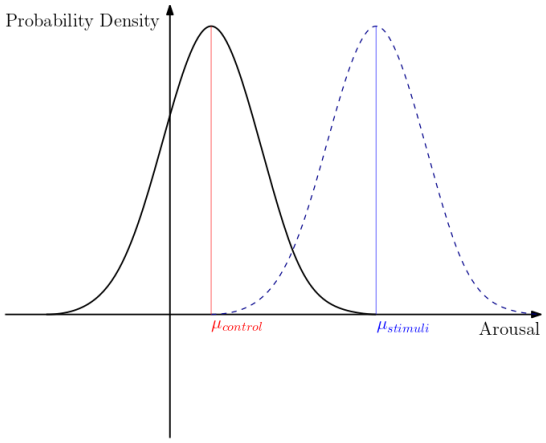
IV. STATISTICAL TESTING BASED ON MACHINE LEARNING

Our hypothesis states that environmental stimuli trigger the build-up of emotions. This built-up emotion is activated by generalised CNS arousal, reflected in motor activity. In the circumplex model, we can map an emotion to the dimension of arousal and valence, and stronger emotions, such as joy and anger, are associated with higher arousal. We assume that general activation does not differentiate by valence, therefore we only consider the arousal dimension. Eventually, our goal is to separate the motor activity of emotionally stimulated situations from the motor activity in situations with the absence of stimuli.

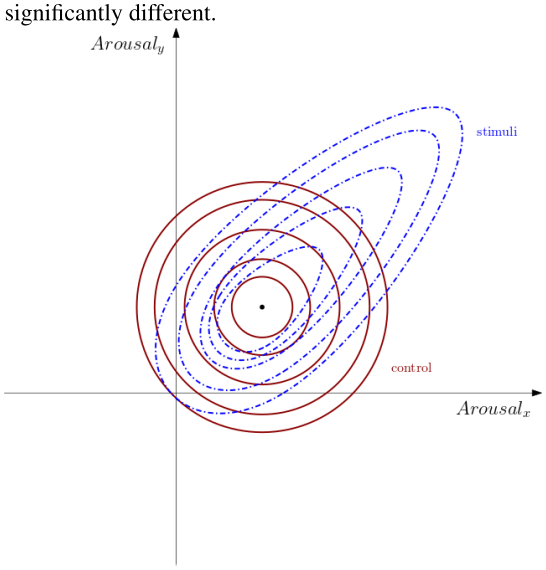
We suggest translating this into a statistical problem with the aim of proving that motor activity is generated by two distinct probability distributions. The probability distribution of motor activity in the presence of emotional stimuli is different from that in the absence of emotional stimuli. The observations of each distribution are paired since the two measurements - motor activity with and without emotional stimuli - come from one participant but under different conditions. Consider paired data in the form of (X_{i0}, X_{i1}). X_{ij} is the observation of a participant i and j refers to the class label $Y_j = j$, where $j \in \{0, 1\}$. The class labels in Y are binary and denote the controlled and the emotionally stimulated motor activity.

A. REPEATED CORRECTED K-FOLD CROSS-VALIDATION T-TEST FOR LOWER BOUNDS

One of the main contributions of this paper is to introduce a hypothesis testing regime based on machine learning.



(a) An illustration of two univariate normal distributions. The illustration shows that the two distributions are significantly different.



(b) An illustration of two bivariate distributions which share the same mean value but reveal different shapes.

FIGURE 2. Two different problems for statistical testing.

The regime can test whether a machine learning classifier can perform better than some given performance level. A prominent example is to test if a classifier performs better than random guessing, confirming that it is information in the data that separates between the two (or more) classes. The machine learning-based hypothesis testing regime does not assume paired data. Therefore, we denote $Z_i = (X_{ij}, Y_j)$. The classification error is derived from a loss function $\mathcal{L}(Z^{n_1}, Z^{n_2})$, where n_1 and n_2 are the training size and test size, respectively. Consequently, Z^{n_1} and Z^{n_2} are the supervised data tuples for training and testing. The loss function can have different functional forms. In the case of a classification problem, a common function is the indicator function $\mathcal{I}[\hat{y} \neq y]$, which derives the accuracy score. The generalised classification error is defined as $\mu = \mathbb{E}[\mathcal{L}(Z_i^{n_1}, Z_l^{n_2})]$ [14].

Cross-validation is a resampling procedure to estimate the generalised classification error $\hat{\mu}$. Cross-validation partitions the data into K subsets. The model is trained on $K-1$ subsets and evaluated on the remaining subset. This procedure can be repeated K times. Cross-validation is an especially useful approach when the data sample size is small since the performance will be tested on every datapoint. However, it requires a high computational cost as a result of multiple training runs. Let μ denote the classification performance of the classifier. An unbiased estimate of the classification error is derived as the average over the classification error of each fold. J. Kim suggests that variability in the estimation of classification errors can be reduced by repeating the complete K-fold cross-validation multiple times [42]. Estimates of the model performance are averaged over the repetitions and the folds, and the resulting procedure is known as repeated K-fold cross-validation. After establishing an understanding of the generalised classification error and cross-validation, we derive the test statistic for a hypothesis test for the lower bounds of the classification error.

Given a hypothesis test that the expected classification performance of a classifier is above some threshold μ_0

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned} \tag{1}$$

Student's t-test is a common approach

$$t = \frac{\mu - \mu_0}{\sigma} \tag{2}$$

However, to be able to use the statistic, the mean, μ, and standard deviation, σ, of the distribution must be estimated. Within the cross-validation framework, estimating the mean classification error is not problematic, and an unbiased cross-validation estimate is given by

$$\hat{\mu}_{RK} = \frac{1}{R \times K} \sum_{r=1}^{R} \sum_{k=1}^{K} \tilde{\mu}_{kr} \tag{3}$$

where R denotes the number of repetitions of the complete K-fold cross-validation. On the other hand, estimating the variance of the classification error estimate has proven to be a challenge, since cross-validation introduces several dependencies [14]. Nadeau and Bengio [43] however, showed that the classification error can be approximated in the sense that the correlation between the classification errors becomes $\frac{n_2}{n_1+n_2}$.

The estimator for the sample variance of the classification errors is formulated in Equation 4, where $(\frac{1}{R\times K} + \frac{n_2}{n_1})$ is called the correction term.

$$Var[\hat{\mu}_{RK}] = \left(\frac{1}{R \times K} + \frac{n_2}{n_1}\right) \tilde{\sigma}^2 \tag{4}$$

We determined the estimates for the mean and variance of the error term μ in Equations 3 and 4. Based on the two estimates, we derive the test statistics for the repeated corrected K-fold cross-validation t-test for lower bounds in Equation 5.

$$t = \frac{\frac{1}{R\times K} \sum_{r=1}^{R} \sum_{k=1}^{K} (\hat{\mu}_{kr} - \mu_0)}{\sqrt{(\frac{1}{R\times K} - \frac{n_2}{n_1})\tilde{\sigma}^2}} \tag{5}$$

The test statistic follows a Student's t-distribution with $df = K \times R - 1$ degrees of freedom. The classification error can be derived from any measure of classification performance. Repeating the K-fold cross-validation increases the number of samples. The distribution of the samples approaches a normal distribution according to the central limit theorem as the number of samples increases.

The test statistic in Equation 5 is similar to the t-test introduced by Nadeau and Bengio, but they used it for the comparison of two machine learning models [43]. To the best of our knowledge, we have not seen any research using the test statistic suggested in Equation 5.

B. HOTELLING's T-SQUARED

We consider Hotelling's T-squared test as a baseline comparison to our machine learning test regime. The paired Hotelling's T-squared test is a classic statistical method to confirm that the difference observed between two multivariate Gaussian distributions is significant, and not only caused by coincidence. It does so by analysing whether the difference between the two mean values $\mu_{control}$ and μ_{goal} is not zero, as visualised in Figure 2a. The paired Hotelling's T-squared test tests whether the sample mean of the differences between two paired sets is significantly different from zero. The test assumes independent, identical, and normally distributed data. According to the null hypothesis, the mean difference between sets is zero [44]:

$$\begin{aligned} H_0 &: \delta = 0 \\ H_1 &: \delta \neq 0 \end{aligned} \tag{6}$$

The test statistic is defined as

$$T^2 = n(\bar{\mathbf{D}} - \delta)' \mathbf{S}_d^{-1} (\bar{\mathbf{D}} - \delta) \tag{7}$$

where the population mean is $\bar{\mathbf{D}} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{D}_j$ and $\mathbf{D}_j = \mathbf{X}_{j1} - \mathbf{X}_{j2}$, where $\mathbf{X}_{j1}$ and $\mathbf{X}_{j2}$ refer to a paired data point in the two classes $Y_j = 1$ and $Y_j = 2$, respectively. The sample covariance is given by $\mathbf{S}_d = \frac{1}{n-1} \sum_{j=1}^{n} (\mathbf{D}_j - \bar{\mathbf{D}})'(\mathbf{D}_j - \bar{\mathbf{D}})$. The test statistic follows a Fisher distribution with p and $n-p$ degrees of freedom [44].

Figure 2 illustrates two potential scenarios of different probability distributions of motor activity. Figure 2a illustrates the simple case of two univariate normal distributions with different expectation values. Figure 2b, on the other hand, shows two probabilities that share the expectation but differ in shape. In the latter case, more complex relationships must be considered than the observed mean of a distribution. Instead of the paired Hotelling's T-squared that works with the mean vector of multivariate distributions, we propose a test regime based on machine learning classifiers that could identify the differences in the description beyond the mean vector, as in Figure 2b. Machine learning classifiers are known for their ability to learn complex relationships within the data [45], [46]. This might result in more powerful tests compared to traditional statistical tests. Moreover, a large number of features can be useful to separate the two sets. This can potentially result in overfitting of the multivariate Gaussian distributions for the two emotional states because of the high dimensionality of the mean vectors. In the machine learning test regime, the problem of overfitting is addressed by dividing the data into disjoint training and test sets. The application of the paired Hotelling's T-squared as a baseline test will demonstrate the necessity of the machine learning test regime in more complex scenarios than the one illustrated in Figure 2a.

V. DATA PREPROCESSING AND FEATURE EXTRACTION

This section addresses the data preprocessing steps that were necessary to transform raw motor activity into valuable features for our hypothesis test and machine learning classifiers. In addition, we discuss the performance evaluation of the classifiers. Finally, we describe how the machine learning experiments were conducted.

A. SIGNAL PROCESSING

The activity was recorded with a GENEActiv wristband, which contains a 3-dimensional microelectromechanical accelerometer module, which measures the acceleration in equivalent gravitational forces (g), with a sampling frequency of 100 hertz [47], [48]. According to the Nyquist-Shannon sampling theorem, frequencies up to 50 hertz can be reconstructed without aliasing [49]. Human activity frequency varies from 0 to 20 hertz. However, most of the activities are located below 10 hertz [50]. Consequently, high-quality human activity preferably records up to 40 hertz. The raw signals captured by the GENEActiv wristband are combined into one activity signal. The three signals are combined according to Equation 8.

$$activity = \sqrt{acc_x^2 + acc_y^2 + acc_z^2} \tag{8}$$

In the second step, we apply a Butterworth filter of order 5. The Butterworth filter of order 5 is commonly applied for signal processing of wrist-worn accelerometers [51]. The signal is filtered to a bandwidth of 10 hertz to 49 hertz. We expected muscular responses to affect higher bandwidth, thus filtering out basic human activity up to 10 hertz.

B. AUGMENTED DICKEY-FULLER TEST

The application of the proposed paired tests assumes that the pairs are identically and independently sampled [44]. However, we sampled eight intervals from each participant's activity time series and couple them into four pairs. This might violate the assumption of identical independent samples due to temporal dependencies. These dependencies could take the form of a time-dependent mean or a time-dependent covariance. A time series that has a constant mean and a constant variance over time is called covariance-stationary [52]. The property of stationarity would confirm that the mean and covariance of the four pairs are not subject to time dependence. The Augmented Dickey-Fuller test is designed to show that a time series is stationary. The presence of a unit

root in the underlying linear stochastic process causes non-stationarity. For example, an autoregressive process of order one in Equation 9.

$$\triangle X_t = (\rho - 1)X_{t-1} + \epsilon_t \tag{9}$$

X_t is an observation of a time series at time t, ρ is a coefficient of the autoregressive process. The process has a unit root when $\rho = 1$. A null hypothesis can be formulated accordingly, where $\delta = \rho - 1$ [53].

$$\begin{aligned} H_0 &: \delta = 0 \\ H_1 &: \delta < 0 \end{aligned} \tag{10}$$

For a more detailed insight into the augmented Dickey-Fuller test, we refer to the book of Hamilton [52] or Shumway et al. [53].

C. FEATURES AND FEATURE SELECTION

We considered nine time series features that are well recognised in the literature for affect recognition [54], [55], [56]. The spectral centroid characterises the frequency spectrum and is understood as the median of the spectrum. Furthermore, the four first moments of a probability distribution, namely mean, standard deviation, skewness, and kurtosis, are applied. In addition, we include features that display the variability of a time series. The root-mean-square deviation and the median absolute deviation describe the deviation of a time series from its centre. The number of peaks is calculated as the local maximum within a certain window size. Complexity is the root-lagged square difference and is defined as

$$complexity = \sqrt{\sum_{i=1}^{n-1}(x_i - x_{i-1})^2} \tag{11}$$

It cannot be specified when exactly a muscular expression occurs as a response to environmental stimuli. The response to stimuli is highly individual for each participant. Therefore, we further split the one-minute intervals into 10 subintervals. The aforementioned features are extracted for each subinterval. We assume that the most informative subinterval is selected during feature selection and that this subinterval consequently contains the stimuli. During feature selection, 90 features are evaluated in total. Of the 90 features, we selected the three best performing ones. We selected the features based on their information gain. The mutual information between the multivariate feature vector and the labels is determined. Let X be a multivariate random vector and Y be a univariate random vector. The formula for mutual information is given in Equation 12.

$$I(X, Y) = \int_X \int_Y p(x, y) log \frac{p(x, y)}{p(x)p(y)} dxdy \tag{12}$$

Mutual information between two random vectors can essentially be described as the ratio of the joint distribution $p(X, Y)$ and the product of their marginal distributions $p(X)$ and $p(Y)$. Mutual information is a generalisation of the correlation coefficient, which accounts for the linear dependency [57]. A feature is selected if its inclusion increases the information gain measured by mutual information.

VI. EXPERIMENTS

We measured the motor activity of devoted football enthusiasts watching a live football match on television between their favourite teams. Goal scoring is considered a trigger point for generalised CNS arousal. Of course, a goal can evoke different emotions, anger, or joy, depending on the team that the enthusiast is supporting. However, since we only considered an increase in aroused energy, we assumed that aroused anger or joy will look relatively similar in motor activity data. We analysed the motor activity of each participant during a goal and compared it to a control sequence. All sequences were in intervals of one minute.

In Section V-C, we demonstrate a feature selection method that is suitable for our problem. We used this feature selection to identify the best feature set for the suggested machine learning test regime and for Hotelling's T-squared. The feature selection and classification were evaluated in repeated cross-validation with $K = 10$ due to a small sample size of 80 intervals. Cross-validation was conducted on a per-participant basis to avoid splitting the paired data. The cross-validation was repeated $R = 1, 5, 10$ and 20 times. We evaluated machine learning models based on the Matthews correlation coefficient (MCC) and accuracy. Seven different classification methods were used to show that the two sets, emotionally stimulated motor activity and control, are significantly different. The variety of chosen classifiers includes linear methods, ensemble methods, and a non-parametric method. Namely, we compared logistic regression, naive Bayes, support vector machine (SVM), K-nearest neighbours (k-NN), Adaptive Boosting (AdaBoost), linear discriminant analysis (LDA), and Random Forest (RF). This variety of methods allowed us to explore simple and complex relationships in the data. The Appendix addresses the models in more detail. Logistic regression is applied with a L1 penalty. SVM uses a radial basis function kernel and a regularisation term of 1. k-NN is trained on 10 nearest neighbours. LDA is solved with the eigendecomposition. The other methods are trained on the default settings.

We aimed to show that the classification methods are better than random guessing. For binary classification, accuracy scores of approximately 50% are regarded as random guessing [37]. Therefore, we have shown that our classification accuracy is significantly higher than 50% and therefore the classification performance is hence better than random guessing.

Cohen's kappa is a frequent performance measure in affect recognition [58], [59], [60], [61]. However, Cohen's kappa reveals some inconsistent behaviour in certain cases [62]. Hence, we chose MCC over Cohen's kappa. An MCC that is well above zero indicates a performance that is better than random guessing [63]. A more detailed explanation of the three performance metrics and a brief discussion on the choice between MCC and Cohen's kappa are attached in the

Appendix. In addition to the experiment described above, we examined how robust the classification performance is for other control intervals. We moved the initial control interval window back and forth by one minute in increments of one second. This led to 119 different intervals around the one initially drawn for the experiment. This allowed us to understand how robust the results are for a larger control sample distribution. For robustness investigation, a 10-fold cross-validation is performed with 20 repetitions. The results are presented as a boxplot figure in Section VII-C.

The experiments were carried out using Python and R [15] software programmes. A MacBook Pro 2020 with M1 chip and 16 GB RAM was used for the execution.

VII. RESULTS AND DISCUSSION

In this section, we present and discuss the results of this work. The section is organised as follows. First, we present the data collection and the data obtained from it in Section VII-A. The data collection is followed by the results of the augmented Dickey-Fuller test in Section VII-B. Afterwards, we present the general classification performance of the machine learning models. Additionally, the results of the robustness of the classification are illustrated in Section VII-C. Finally, Section VII-D demonstrates the results of the proposed machine learning-based test regime and compares them to the baseline Hotelling's T-squared test.

A. PARTICIPANTS AND DATA COLLECTION

Eleven soccer supporters gathered to watch a live broadcast of the Premier League match between Liverpool and Manchester United (4 - 0) on 19^{th} of March 2022, all equipped with a wrist-worn accelerometer of type GENEActiv [48]. Goals occurred in minutes 5, 22, 68, and 85 of the game. We chose four control intervals in minutes 3, 8, 27, and 36. The intervals were of one minute duration consisting of 6000 sample points. The time points were identified according to a game summary [64]. Before and during the game, no drugs and nothing stronger in terms of alcohol than low-alcohol beer were accepted for consumption. For one participant, the actigraph did not collect any data and the participant was excluded from the analysis. Consequently, we analysed 10 football supporters, operationalised as 80 time intervals with 40 emotionally stimulated intervals and 40 control intervals. All participants reported being fans of their team since childhood (before 12 years of age). One supporter reported being a lifetime fan of a deviant football team but reported a strong dislike for one of the competing teams. Consequently, our sample contained four people who wished for Liverpool's defeat and six people who opposed Manchester United. Nine participants were men and one was female, with a mean age of 24 years (range 18 – 50). In addition, participants were asked to rate three questions from the self-rated seven-item identification with their team scale [65]. Responses to each item were on a 1 to 8 Likert scale, and higher scores indicated a more passionate fan identification. We utilised the following three items; how important to you is it that your football team wins, the result was 7.3 (1.1) (mean (standard derivation)); How strongly do you see yourself as a fan of your football team, which resulted in 7.1 (1.0); How much do you dislike the other football team, resulted in 5.8 (2.3). The questionnaire used is attached in the Appendix.

TABLE 1. Descriptive values of the dataset.

Dataset description	Totals
Number of participants	10
Number of goals	4
Length of motor activity	654000
Length of interval	6000

TABLE 2. The results of the augmented Dickey Fuller test. The test for covariance-stationarity was performed for the activity time series of each participant. The table reports the test statistic and the p-values for each participant. A p-value below the significance level of $p = 5\%$ indicates covariance-stationarity.

Participant	Test-Statistic	p-Value
1	-41.50	0.01
2	-43.18	0.01
3	-46.49	0.01
4	-35.08	0.01
5	-34.29	0.01
6	-40.15	0.01
7	-43.55	0.01
8	-41.77	0.01
9	-46.09	0.01
10	-39.63	0.01

Table 1 summarises the datasets with the most important information about size and shape. The environmental setting is considered a closed real-life scenario. Because, although unquestionably a closed environment, the scenery still resembles a natural real-life situation. All participants were aware of the purpose of this experiment and consented to participate by attending the event and by wearing the accelerometer. No personally sensitive information was collected, all data is fully anonymised following the GDPR guidelines, and all procedures were in accordance with the recommendations of the data protection agent at Oslo Metropolitan University.

B. RESULTS OF THE AUGMENTED DICKEY-FULLER TEST

We used the augmented Dickey-Fuller test to evaluate whether the activity time series of the participants are covariance stationary. The property of covariance stationarity ensures a constant mean and covariance of the time series. A constant mean and covariance are desirable in our analysis because the sampled intervals from the activity time series of the participants need to be identically and independently. By showing that the mean and covariance are constant, we ensure that the distribution of the sample intervals is not time-dependent.

The results of the augmented Dickey-Fuller test are summarised in Table 2. The results show that the test is significant for each time series. This means that each time series

TABLE 3. Matthew's correlation coefficient scored by each model. The score is reported for different repetitions of the cross-validation. We cannot conclude that a certain number of repetitions improve the results.

Repetitions	Log. Regression	SVM	AdaBoost	k-NN	Naive Bayes	LDA	RF
5	0.238	0.249	0.232	0.263	0.181	0.285	0.213
10	0.253	0.259	0.246	0.253	0.184	0.305	0.216
20	0.238	0.255	0.240	0.245	0.180	0.289	0.214

TABLE 4. Accuracy of the models for different repetitions of the cross-validation. The ensemble models AdaBoost and Random Forest were outperformed. The other classifiers scoreed similar scores around 60%.

Repetitions	Log. Regression	SVM	AdaBoost	k-NN	Naive Bayes	LDA	RF
5	0.605	0.612	0.615	0.607	0.567	0.627	0.597
10	0.611	0.612	0.612	0.608	0.566	0.625	0.597
20	0.603	0.612	0.611	0.621	0.567	0.632	0.597

TABLE 5. The table reports the p-values of the corrected repeated 10-fold cross-validation one-sample t-test. The p-values are calculated for different numbers of repetitions. We observe that the p-values decrease with an increasing number of repetitions. The p-values below 5% are highlighted.

Repetitions	Log. Regression	SVM	AdaBoost	k-NN	Naive Bayes	LDA	RF
5	0.002	0.001	0.058	0.018	0.008	0.001	0.043
10	0.003	0.001	0.044	0.013	0.017	0.001	0.038
20	0.001	0.001	0.039	0.008	0.011	0.001	0.023

is stationary. We could conclude that the assumptions of the t-tests are not violated.

C. CLASSIFICATION RESULTS

Before we present the results of our proposed test regime, we evaluate the general performance of the machine learning models in terms of their accuracy and MCC. Table 3 contains the MCC scored for each machine learning model. We further analyse the value for a different number of cross-validation repetitions. The MCC values achieved by the proposed machine learning models range from 0.171, scored by naive Bayes to 0.318 scored by linear discriminant analysis. We observed that a larger number of repetitions did not greatly affect the MCC.

Table 4 shows the accuracy scores for all models and the different number of repetitions. We also notice that the accuracy did not improve with an increasing number of repetitions. Models that performed best according to MCC also performed so for accuracy. Linear discriminant analysis (LDA) scored the best accuracy result of 63.7%. All models, except naive Bayes and random forest, consequently scored above 60% accuracy.

Additionally, we analysed how robust the binary classification performed for different control intervals drawn from the motor activity time series. Figure 3 illustrates the distribution of accuracy at 119 different control intervals. We represent the distribution of accuracy for each model as boxplots. The figure allows us to understand the variance and the scale of the accuracy score. The median accuracies of logistic regression, support vector machine, naive Bayes, and linear discriminant analysis were above 55%. AdaBoost, k-NN, and Random Forest achieved a median accuracy of less than 55%. These three models also had the highest variance when comparing the range of the boxplots. We set the whiskers of the boxplot to the 5^{th} percentile and the 95^{th} percentile. The lower whiskers of all models were less accurate 50%. The results showed that for all models except AdaBoost the lower bounds of the accuracy, the boxplots are confidently above the 50% level.

D. RESULTS OF THE REPEATED CORRECTED K-FOLD CROSS-VALIDATION T-TEST FOR LOWER BOUNDS

Lastly, we present the results of the repeated corrected 10-fold cross-validation t-test for lower bounds and compare them with the results of the paired Hotelling's T-squared test. Table 5 shows the p-values of the t-test. We see that most of the p-values were very low and below the significant level 5%. AdaBoost is the only method that did not achieve significance for 5 repetitions. The paired Hotelling's T-squared is 7.581 and scores a p-value of 0.0836. The result indicates a difference between the distributions but is usually considered to be statistically non-significant. Both hypothesis tests seem to detect a difference. However, unlike our proposed test regime, the paired Hotelling's T-squared failed to achieve a significant result. We could argue that our proposed test regime reached a higher statistical power. As discussed in Section IV-B, the difference between emotionally stimulated motor activity and motor control activity could be captured by higher order differences in the distributions than the mean vector.

The p-values in Table 5 seem to decrease with increasing number of repetitions. We discussed repeated cross-validation in Section IV-A and concluded that the repeated procedure will result in reduced noise from the performance metric estimates. Therefore, we assume that the p-values converge and do not systematically decrease. Logistic regression, AdaBoost, k-NN, and Random Forest did not reach a significant level with a single cross-validation. However, all models achieved significance with increasing repetitions. We conclude that the seven proposed machine learning models were

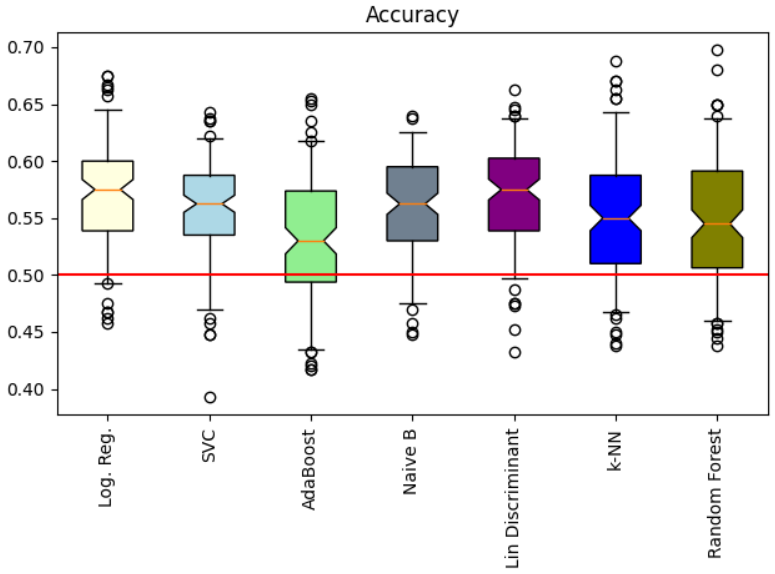


FIGURE 3. Illustration of the robustness of classifiers analysed. The classification is applied to 119 shifted control intervals to show the robustness of the classification performance. The boxplots show the distribution of the accuracy scores. The orange line shows the median and the box covers the first and the third quartile. The outer lines illustrate the 5^{th} and 95^{th} percentile. The dots show outliers in the distribution.

able to discriminate between intervals of motor activity with emotional stimuli and intervals without emotional stimuli for the given dataset.

VIII. CONCLUSION

This work investigated the potential to discriminate altered arousal in time series of motor activity. Emotional stimuli cause generalised arousal of the CNS. The excitation of the CNS is transmitted to human motor activity through the SNS. We conducted an experiment from which we obtained motor activity data with assumed emotional stimuli and control data. Our hypothesis states that we recognise arousal in motor activity if the two sets of motor activity data can be distinguished. We have suggested a machine learning framework to statistically test whether the two distributions can be separated. The suggested approach was compared with the traditional Hotelling T-squared test. Seven different binary machine learning classifiers were used in the suggested machine learning approach, namely logistic regression, SVM, k-NN, AdaBoost, and linear discriminant analysis achieve accuracy scores greater than 50% (p-value $<$ 0.01), and MCC scores greater than 0% (p-value $<$ 0.01) in our experiment, that is, better than random guessing. The results also prove robust when the control intervals are shifted. Furthermore, the Hotelling T-squared test was unable to significantly separate the two groups (p-value $=$ 0.08), indicating that the suggested machine learning framework results in tests more powerful than the traditional statistical approach.

This work contributes to research on affect recognition of the potential of recognising arousal within motor activity. We substantiate our hypothesis with the theoretical idea of expressed generalised CNS arousal in motor activity through the SNS. To the best of our knowledge, the potential of motor activity to be a physiological signal to recognise affect has not yet been analysed. The results of this study need further replication and validation on several datasets. However, we set a high critical bar before drawing conclusions from our results by introducing hypothesis testing for lower bounds on binary classifier performance metrics. The introduced test is a modification to the already known corrected repeated k-fold cross-validation test. We argue that the proposed hypothesis test can be generally applied in cases where a certain lower bound needs to be established and documents higher power than the traditional statistical approach. Given the limited number of participants, it is important to report uncertainties and understand the lower bounds of the performance metrics.

In future work, the potential to recognise arousal within motor activity should be further evaluated. The data collection can be repeated for different physiological signals. Our proposed feature selection can be applied to various physiological signals. For each signal, the most informative subinterval for the recognition of arousal could be identified. Consequently, for each signal, the agreement over the selected subinterval could be examined. Moreover, improvements in recognition results could be measured when motor activity is used in a multi-model approach. Explanatory techniques, such as Shapely values, can be used to understand the contribution of motor activity to the arousal detection task. The corrected repeated k-fold cross-validation test allowed us to use machine learning methods to test for a significant

difference between two distributions. This can be extended to the multiple-testing approach.

APPENDIX A
PERFORMANCE METRICS

A. ACCURACY

Accuracy is a well-known classification and is among the most popular once [66]. It measures the overall performance of the classification. Accuracy is the ratio of correctly predicted cases and all classes. The formula can be derived directly from the confusion matrix as follows,

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{13}$$

Accuracy can be understood as the probability of drawing an instance at random and correctly predicting it [66]. We introduce an alternative definition of accuracy. We denote correctly classified instances with 1 and incorrectly classified instances with 0. We derive the following indicator function. The accuracy can then be defined as the mean sample of the indicator function. C is the set of correctly specified instances.

$$accuracy = \frac{1}{n}\sum_{x=1}^{n} \mathbb{1}(x)$$
$$\mathbb{1}(x) = \begin{cases} 1, & x \in C \\ 0, & x \notin C. \end{cases} \tag{14}$$

B. COHEN's KAPPA

As a second performance metric, we introduce Cohen's kappa. Its concept is related to the Matthew correlation coefficient. The predicted and true labels are interpreted as random variables and their agreement is calculated. Cohen's work was originally developed to calculate the agreement between two expert decisions [67]. The formula of Cohen's kappa coefficient is given by

$$K = \frac{P_o - P_e}{1 - P_e} \tag{15}$$

P_o is the observed agreement of the true labels and the predictions. Effectively, that is, the accuracy of the model. P_e is the expected agreement. This expected agreement corresponds to the accuracy obtained by chance [66]. We can interpret a Cohen's kappa of 0 as no agreement and values up to 0.2 as slight agreement. A Cohen's kappa that lies between 0.21 and 0.4 is considered a fair agreement. The following intervals [0.41, 0.6], [0.61, 0.8], [0.81, 1] are considered to be in moderate, significant, and perfect agreement, respectively [68].

C. MATTHEW's CORRELATION COEFFICIENT

MCC shows the correlation between the actual values and the predicted ones. A perfect correlation is denoted by one, while random guessing is denoted by 0. MCC is especially advantageous when data is imbalanced, since it takes into account every entry in the confusion matrix [66]. Its formula is recorded in Equation 16.

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{16}$$

D. DISCUSSION ON COHEN's KAPPA AND MATTHEW's CORRELATION

The Cohen's kappa specifies the agreement between predictions and true labels by taking into account the agreement by chance [67]. Consequently, higher Cohen's kappa values indicate that the models perform better than random guessing. But Cohen's kappa suffers from paradoxical behaviour in certain scenarios [62]. We refer to the work of [69] for a detailed discourse on the paradox behaviour of Cohen's kappa. In the case of binary classification, MCC is derived as the discretised Pearson's correlation coefficient of the prediction vector and the true evaluation vector. Reference [70] shows, how the two metrics are equal for symmetric confusion metrics but that MCC disagrees with the paradox behaviour of Cohen's kappa.

APPENDIX B
MACHINE LEARNING CLASSIFIERS

A. LOGISTIC REGRESSION

Logistic regression is the adaptation of a linear regression model to a binary dependent variable. For linear models to work as we know them, the dependent variable must be continuous. The odds p of the binary variable can be transformed using a logistic function. The transformation is described in Equation 17

$$logit(p) = ln(\frac{p}{1-p}) = \beta_0 + \beta X \tag{17}$$

β are the parameters of the linear regression model and X is the design matrix. The relationship between odds p and X forms an s-shaped function, described as the logistic curve [44]. Logistic regression is a well-known and recognised method for binary classification. The fitted coefficients are interpretable and the model fit can be evaluated. However, logistic regression can easily overfit with an increasing number of variables included. Due to the three best-selected features, this is not our concern.

B. NAIVE BAYES CLASSIFIER

The naive Bayes classifier is based on the Bayes theorem. Naive Bayes calculates the conditional probabilities for features given a certain class. Following the Bayes theorem, the posterior distribution is formulated accordingly in Equation 18.

$$p(y|x_1, \ldots, x_n) = p(y)\prod_{i=1}^{n} p(x_i|y) \tag{18}$$

For binary classification, a Bernoulli distribution is used. The naive Bayes classifier assumes that the features are conditionally independent. This assumption is very strong and often gets violated. Therefore, the classifier is called naive [57]. Reference [71] shows that the classifier still performs nearly optimal [71]. Due to its simplicity, the classifier is quite immune to overfitting [57].

C. SUPPORT VECTOR MACHINE

A support vector machine (SVM) classifier separates m dimensional feature vectors with a $m-1$ dimensional hyperplane [72]. The SVM finds a hyperplane that maximises the margin. The margin is the smallest distance between datasets of different classes. The maximisation problem corresponds to [73]. The SVM comes with an interesting property. The maximisation problem is equal to a convex optimisation problem. Therefore, a local optimum is also a global one [73]. SVM is of special interest in our analysis. Bommae and Oertzen suggest the SVM as a comparison test. SVM detects the difference in sets when the classification accuracy is greater than 50% and thus better than random guessing [74].

D. RANDOM FOREST

The random forest is an ensemble learning technique that aggregates a number of decision trees. Decision trees are trained on randomly selected subsets of data [57]. In the case of classification, the majority vote of the trees is the final predictive output of the random forest. The correlation between the errors of single trees can be high in practise, which could lead to a high variance of the final prediction error [73].

E. ADAPTIVE BOOSTING

Adaptive Boosting is also called AdaBoost. Boosting is a form of ensemble learning. The idea behind it is the sequential application of weighted weak classifiers. A majority vote in the manner of a sign function is based on the outcome. The weight is updated with each iteration step. The sequential procedure allows the algorithm to focus on misclassified observations [75].

F. K-NEAREST NEIGHBOURS

K-nearest neighbours (k-NN) is a non-parametric classification method.An observation is assigned to a specific class based on the classes of neighbouring observations. The assignment of the class is made by majority vote. The algorithm relies on the assumption that observations of the same class are closer together in the feature space. Neighbours are determined by a distance measure, most frequently the Euclidean distance is applied [57].

G. LINEAR DISCRIMINANT ANALYSIS

LDA serves the purpose of separating two populations. The aim is to find a linear function of various variables that maximises the difference ratio between two populations [44]. The ratio of the difference is defined as the between-population variance to the within-population variance. The ratio is called the Fisher criterion [73]. The equation for the ratio is as follows.

$$J(w) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} \tag{19}$$

where s_i is the variance of the population and m_i is the mean value of the population. In the simple case, LDA assumes that populations are normally distributed, with equal covariance matrices [44]. These assumptions restrict the application of the method, especially for problems with very few sample points. With increasing sample size, the central limit theorem holds.

APPENDIX C
QUESTIONNAIRE

TABLE 6. Emotions expressed in motor activity when watching football matches.

NAME: _____ DATE: _____ (a) GENDER: □ Male (0) □ Female (1) GENEActiv ID: _____

STUDY ID: ____ (b) AGE: _ _

(c) YOUR FOOTBALLTEAM: _____ (d) OTHER FOOTBALLTEAM: _____

(e) Since what age have you been a fan of YOUR football team?	□ 1. Childhood (0-12)	□ 2. Adolescence (13-17)	□ 3. Adult (18+)						
(e) How important to YOU is it that YOUR football team wins?	Not Important	1	2	3	4	5	6	7	8 Very Important
(g) How strongly do YOU see YOURSELF as a fan of YOUR football team?	Not at all a Fan	1	2	3	4	5	6	7	8 Very much a Fan
(h) How much do YOU dislike the OTHER football team?	Do not Dislike	1	2	3	4	5	6	7	8 Dislike Very Much

DECLARATION OF INTEREST

All authors declare that they have no conflicts of interest.

REFERENCES

[1] B. Liu, "Many facets of sentiment analysis," in *A Practical Guide to Sentiment Analysis*. Berlin, Germany: Springer, 2017, pp. 11–39.

[2] P. Schmidt, A. Reiss, R. Durichen, and K. V. Laerhoven, "Wearable-based affect recognition—A review," *Sensors*, vol. 19, no. 19, p. 4079, Sep. 2019.

[3] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.

[4] T. Wang and H. Zhang, "Using wearable devices for emotion recognition in mobile human-computer interaction: A review," in *Proc. Int. Conf. Human-Comput. Interact.* Cham, Switzerland: Springer, 2022, pp. 205–227.

[5] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017.

[6] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, pp. 1–36, Apr. 2015.

[7] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychol. Rev.*, vol. 110, no. 1, pp. 145–172, 2003.

[8] R. W. Picard, "Affective computing: From laughter to IEEE," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 11–17, Jan. 2010.

[9] D. P. Calderon, M. Kilinc, A. Maritan, J. R. Banavar, and D. Pfaff, "Generalized CNS arousal: An elementary force within the vertebrate nervous system," *Neurosci. Biobehavioral Rev.*, vol. 68, pp. 167–176, Sep. 2016.

[10] D. Pfaff, *How Brain Arousal Mechanisms Work: Paths Toward Consciousness*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[11] P. Jakobsen, A. Stautland, M. A. Riegler, U. Cote-Allard, Z. Sepasdar, T. Nordgreen, J. Torresen, O. B. Fasmer, and K. J. Oedegaard, "Complexity and variability analyses of motor activity distinguish mood states in bipolar disorder," *PLoS ONE*, vol. 17, no. 1, Jan. 2022, Art. no. e0262232.

[12] P. Jakobsen, E. Garcia-Ceja, M. Riegler, L. A. Stabell, T. Nordgreen, J. Torresen, O. B. Fasmer, and K. J. Oedegaard, "Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls," *PLoS ONE*, vol. 15, no. 8, Aug. 2020, Art. no. e0231995.

[13] J. McLean, D. Brennan, D. Wyper, B. Condon, D. Hadley, and J. Cavanagh, "Localisation of regions of intense pleasure response evoked by soccer goals," *Psychiatry Res., Neuroimaging*, vol. 171, no. 1, pp. 33–43, Jan. 2009.

[14] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of K-fold cross-validation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16, 2003, pp. 1–8.

[15] M. Boeker, "Emotion extraction," Dept. Holistic Syst., SimulaMet, Oslo, Norway, 2023.

[16] M. Boeker, "OSF: Affect recognition in muscular response signals," Dept. Holistic Syst., SimulaMet, Oslo, Norway, 2023.

[17] M. R. Graver, "Cicero on the emotions: Tusculan disputations 3 and 4," Tech. Rep., 2002.

[18] P. Ekman, "An argument for basic emotions," *Cognition Emotion*, vol. 6, nos. 3–4, pp. 169–200, May 1992.

[19] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, Jul. 2020.

[20] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Mar. 2012.

[21] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Sep. 2022.

[22] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.

[23] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2106–2112.

[24] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. 5th Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2015, pp. 73–80.

[25] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 445–450.

[26] E. R. Kandel, J. D. Koester, S. H. Mack, and S. A. Siegelbaum, *Principles of Neural Science*, vol. 6. New York, NY, USA: McGraw-Hill, 2021.

[27] E. T. Rolls, "Limbic systems for emotion and for memory, but no single limbic system," *Cortex*, vol. 62, pp. 119–157, Jan. 2015.

[28] U. Coté-Allard, P. Jakobsen, A. Stautland, T. Nordgreen, O. B. Fasmer, K. J. Oedegaard, and J. Tørresen, "Long-short ensemble network for bipolar manic-euthymic state recognition based on wrist-worn sensors," *IEEE Pervasive Comput.*, vol. 21, no. 2, pp. 20–31, Apr. 2022.

[29] R. W. Levenson, P. Ekman, K. Heider, and W. V. Friesen, "Emotion and autonomic nervous system activity in the Minangkabau of West Sumatra," *J. Personality Social Psychol.*, vol. 62, no. 6, pp. 972–988, 1992.

[30] A. Pecchinenda, "The affective significance of skin conductance activity during a difficult problem-solving task," *Cognition Emotion*, vol. 10, no. 5, pp. 481–504, Sep. 1996.

[31] J. Scheirer, R. Fernandez, J. Klein, and R. W. Picard, "Frustrating the user on purpose: A step toward building an affective computer," *Interacting Comput.*, vol. 14, no. 2, pp. 93–118, Feb. 2002.

[32] S. R. Vrana, "The psychophysiology of disgust: Differentiating negative emotional contexts with facial EMG," *Psychophysiology*, vol. 30, no. 3, pp. 279–286, May 1993.

[33] R. M. Mehmood and H. J. Lee, "Emotion classification of EEG brain signal using SVM and KNN," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jun. 2015, pp. 1–5.

[34] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalogr. Clin. Neurophysiol.*, vol. 29, no. 3, pp. 306–310, Sep. 1970.

[35] W. Zheng, J. Zhu, and B. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 417–429, Jul. 2019.

[36] C.-K. Wu, P.-C. Chung, and C.-J. Wang, "Representative segment-based emotion analysis and classification with automatic respiration signal segmentation," *IEEE Trans. Affect. Comput.*, vol. 3, no. 4, pp. 482–495, 4th Quart., 2012.

[37] J. Kim and E. Andre, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, Dec. 2008.

[38] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage*, vol. 102, pp. 162–172, Nov. 2014.

[39] M. Gjoreski, M. Lustrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *J. Biomed. Informat.*, vol. 73, pp. 159–170, Sep. 2017.

[40] M. Ciman and K. Wac, "Individuals stress assessment using human-smartphone interaction analysis," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 51–65, Jan. 2018.

[41] E. Garcia-Ceja, V. Osmani, and O. Mayora, "Automatic stress detection in working environments from smartphones accelerometer data: A first step," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 4, pp. 1053–1060, Jul. 2016.

[42] J.-H. Kim, "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap," *Comput. Statist. Data Anal.*, vol. 53, no. 11, pp. 3735–3745, Sep. 2009.

[43] C. Nadeau and Y. Bengio, "Inference for the generalization error," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 1999, pp. 1–7.

[44] R. A. Johnson and L. Simar, *Applied multivariate statistical analysis*, vol. 405. 1992.

[45] A. K. Jain, R. P. W. Duin, and J. C. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[46] D. Bhamare and P. Suryawanshi, "Review on reliable pattern recognition with machine learning techniques," *Fuzzy Inf. Eng.*, vol. 10, no. 3, pp. 362–377, Jul. 2018.

[47] C. A. Jenkins, L. C. F. Tiley, I. Lay, J. A. Hartmann, J. K. M. Chan, and C. L. Nicholas, "Comparing GENEActiv against Actiwatch-2 over seven nights using a common sleep scoring algorithm and device-specific wake thresholds," *Behav. Sleep Med.*, vol. 20, no. 4, pp. 369–379, Jul. 2022.

[48] *Geneactiv: Product description*, A-Insights, Amsterdam, The Netherlands, 2022.

[49] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[50] R. Khusainov, D. Azzi, I. Achumba, and S. Bersch, "Real-time human ambulation, activity, and physiological monitoring: Taxonomy of issues, techniques, applications, challenges and limitations," *Sensors*, vol. 13, no. 10, pp. 12852–12902, Sep. 2013.

[51] B. H. T. Lindert and E. J. W. Van Someren, "Sleep estimates using micro-electromechanical systems (MEMS)," *Sleep*, vol. 36, no. 5, pp. 781–789, May 2013.

[52] J. D. Hamilton, *Time Series Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 2020.

[53] R. H. Shumway, D. S. Stoffer, and D. S. Stoffer, *Time Series Analysis and Its Applications*, vol. 3. Berlin, Germany: Springer, 2000.

[54] G. Valenza, A. Lanata, and E. P. Scilingo, "The role of nonlinear dynamics in affective valence and arousal recognition," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 237–249, Apr. 2012.

[55] W. Wen, G. Liu, N. Cheng, J. Wei, P. Shangguan, and W. Huang, "Emotion recognition based on multi-variant correlation of physiological signals," *IEEE Trans. Affect. Comput.*, vol. 5, no. 2, pp. 126–140, Apr. 2014.

[56] Z. Lan, O. Sourina, L. Wang, and Y. Liu, "Real-time EEG-based emotion monitoring using stable features," *Vis. Comput.*, vol. 32, no. 3, pp. 347–358, Mar. 2016.

[57] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[58] A. Raheel, S. M. Anwar, and M. Majid, "Emotion recognition in response to traditional and tactile enhanced multimedia using electroencephalography," *Multimedia Tools Appl.*, vol. 78, no. 10, pp. 13971–13985, May 2019.

[59] A. Al-Nafjan, A. Al-Wabil, A. AlMudhi, and M. Hosny, "Measuring and monitoring emotional changes in children who stutter," *Comput. Biol. Med.*, vol. 102, pp. 138–150, Nov. 2018.

[60] G. Chanel, S. Avry, G. Molinari, M. Bétrancourt, and T. Pun, "Multiple users emotion recognition: Improving performance by joint modeling of affective reactions," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 92–97.

[61] M. P. Arthanarisamy Ramaswamy and S. Palaniswamy, "Subject independent emotion recognition using EEG and physiological signals—A comparative study," *Appl. Comput. Informat.*, Sep. 2022.

[62] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, Dec. 2020.

[63] D. Chicco and G. Jurman, "An invitation to greater use of Matthews correlation coefficient (MCC) in robotics and artificial intelligence," *Frontiers Robot. AI*, p. 78, Mar. 2022.

[64] M. Gamlem, "VG: KAMP report Liverpool 4–0 Manchester United," vg.no, vglive.no, Oslo, Norway, 2023.

[65] D. L. Wann and N. R. Branscombe, "Sports fans: Measuring degree of identification with their team," *Int. J. Sport Psychol.*, Jan. 1993.

[66] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," 2020, *arXiv:2008.05756*.

[67] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[68] M. L. McHugh, "Interrater reliability: The Kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.

[69] L. Flight and S. A. Julious, "The disagreeable behaviour of the Kappa statistic," *Pharmaceutical Statist.*, vol. 14, no. 1, pp. 74–78, Jan. 2015.

[70] R. Delgado and X.-A. Tibau, "Why Cohen's Kappa should be avoided as performance measure in classification," *PLoS ONE*, vol. 14, no. 9, Sep. 2019, Art. no. e0222916.

[71] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Mach. Learn.*, vol. 29, nos. 2–3, pp. 103–130, 1997.

[72] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006.

[73] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4. Berlin, Germany: Springer, 2006.

[74] B. Kim and T. V. Oertzen, "Classifiers as a model-free group comparison test," *Behav. Res. Methods*, vol. 50, no. 1, pp. 416–426, Feb. 2018.

[75] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2. Berlin, Germany: Springer, 2009.

MATTHIAS BOEKER received the degree in industrial engineering from the Karlsruhe Institute of Technology, Germany, in 2020. He is currently pursuing the Ph.D. degree with SimulaMet and Oslo Metropolitan University, Oslo. His research interests include time series analysis in sports and health, statistical modeling, and machine learning. Most of his research works focus on physiological signals like motor activity and their applications.

PETTER JAKOBSEN received the Candidatus Magisterii degree in law/social science from the University of Bergen, Norway, the bachelor's degree in general nursing from the Haraldsplass Deaconess University College, Bergen, Norway, and the master's degree in evidence-based practice from the Bergen University College, Norway. He is currently pursuing the Ph.D. degree with the Haukeland University Hospital, with a focus on identifying predictors of relapse in bipolar disorder and analyzing time series of motor activity collected from both hospitalized patients and outpatients with nonlinear models and machine learning. He has extensive experience in acute clinical mental health, as a research nurse, and in bartending. He was the Norwegian Coordinator of an international multicentre study—"the Pharmacogenomics of Mood Stabilizer Response in Bipolar Disorder (PGBD)" study. He has been involved in the digitization of both the research and the clinical workflow at the Haukeland University Hospital and the development of automated quality registers. Since 2017, he has been partly working as an Adviser with the Division of Psychiatry and is about to finalize his Ph.D. thesis. The Research Council of Norway funded his Ph.D. grant.

MICHAEL A. RIEGLER received the Ph.D. degree from the Department of Informatics, University of Oslo, Oslo, Norway, in 2017. He is currently the Chief Research Scientist with SimulaMet, Oslo, and a Professor with Oslo Metropolitan University, Oslo. His research interests include a wide array of topics, such as machine learning, video and image analysis and understanding, image processing, image retrieval, crowdsourcing, social computing, and biological and medical applications of machine learning.

LENA ANTONSEN STABELL was born in 1977. She received the bachelor's degree in general nursing from the Bergen University College, Norway, in 2001, and the master's degree in evidence-based practice in 2012. She is currently pursuing the Ph.D. degree with the Psychiatric Research Department, Helse-Bergen, and NORMENT Centre of Excellence. She specialized in mental health and counseling in 2004 and 2005, respectively. She is working in inpatient and outpatient clinics at Bergen, and in mental health nursing homes in Bournemouth, U.K. Since 2018, she has been teaching a course in mental health publishing at the University of Bergen. She is currently a Database Coordinator with the Bergen Psychosis Research Group (BPRG). In 2019, she received the Ph.D. funding from Helse-Vest to investigate how clinical insight into illness fluctuates with symptom load and antipsychotic medication. She is currently involved in research on various treatment components for schizophrenia. Her research interests include statistics and knowledge of mental health led to cooperation with the Research Group on Machine Learning, SimulaMet, Oslo, Norway.

OLE BERNT FASMER is currently a Psychiatrist and a Professor Emeritus in adult psychiatry with the Department of Clinical Medicine, University of Bergen, Norway. His research interests include the biological aspects of bipolar disorder and ADHD, with an emphasis on somatic disorders and mathematical analyses of time series, including motor activity and heart rate.

PÅL HALVORSEN received the Ph.D. degree from the University of Oslo, Norway, in 2001. He is currently the Chief Research Scientist with SimulaMet, Norway, a Professor with Oslo Metropolitan University, Norway, and a Professor II with the University of Oslo. At SimulaMet, he is currently leading the Department of Holistic Systems Research, which investigates the challenges of complete end-to-end pipelines with a particular focus on sports and medical applications. His current research interests include several areas in distributed (multimedia) systems and content analysis from a performance and efficiency point of view. More information can be found at http://home.simula.no/paalh.

HUGO LEWI HAMMER received the M.Sc. and Ph.D. degrees from the Norwegian University of Science and Technology, in 2003 and 2008, respectively. He worked as a Researcher with the Norwegian Computer Center, Oslo, Norway. He is currently a Professor with the Department of Computer Science, Oslo Metropolitan University, Oslo, an Adjunct Chief Research Scientist with the Simula Metropolitan Center, Oslo, and a Senior Scientist at the Oslo University Hospital. His research interests include computer-intensive statistical methods, uncertainty quantification, knowledge discovery, and transparent machine-learning methods.