Manic episodes of bipolar disorder can lead to uncritical behavior and delusional psychosis, often with destructive consequences for those affected and their surroundings. Early detection and intervention of a manic episode are crucial to prevent escalation, hospital admission, and premature death. However, people with bipolar disorder may not recognize that they are experiencing a manic episode and symptoms, such as euphoria and increased productivity can also deter affected individuals from seeking help. This work proposes to perform user-independent, automatic mood-state detection based on actigraphy and electrodermal activity acquired from a wrist-worn device during mania and after recovery (euthymia). This article proposes a new deep learning-based ensemble method leveraging long (20 h) and short (5 min) time intervals to discriminate between the mood states. When tested on 47 bipolar patients, the proposed classification scheme achieves an average accuracy of 91.59% in euthymic/manic mood-state recognition.

Bipolar disorder is a severe mental disorder characterized by intense periodic mood fluctuations, lifelong disability, and a high disease burden that affects more than 1% of the global population.1,2 People with bipolar disorder have a mortality risk twice as high as the general population, due to somatic comorbidities and suicide rates 20–30 times higher than the general population.2 Bipolar disorder is usually divided into two subgroups: 1) bipolar I and 2) bipolar II. Bipolar I is defined by the presence of manic episodes, typically characterized by increased energy, inflated self-esteem, increased need to pursue goal-directed actions, reduced subjective need for sleep, and is often associated with the presence of hallucinations and delusions. The elevated mood defining bipolar II is hypomania, a less severe form of mania, and without hallucinations and delusions. Another difference is that at least one major depressive episode is needed for the diagnosis of bipolar II but not bipolar I. The presence of depressive episodes, which are typically characterized by diminished initiative and energy, as well as disturbed sleep patterns are nevertheless common in bipolar I. The neutral state euthymia can be characterized as the in-between state that neither meets the criteria for depression nor mania/hypomania.

Early help and intervention is an important factor in mitigating the risks associated with mania.3 However, it can be hard for the affected person to realize that they are experiencing an episode.1 Furthermore, even when recognizing that a manic episode is
occurring, the sense of euphoria and increased productivity can be dissuading factors in seeking help.

As no biomarker has yet been approved for the diagnosis of bipolar disorder, current practices in assessing mood episodes focus on subjective observation in conjunction with semistructured clinical rating scales. Consequently, it remains challenging to perform efficient targeted interventions, due to the delicate balance between adequately monitoring the patient and moderating the impact of repeated appointments on the healthcare system and the patient’s life. Changes in mood triggered by an affective disorder are not only associated with changes in behavior but are also reflected in several biological processes, such as in the autonomic nervous system. As a result, much effort has been deployed in characterizing mood states in affective disorders from various biosignals [e.g., electrodermal activity (EDA), actigraphy, electrocardiogram], with the aim of automatically identifying state change without human intervention. To achieve this goal, however, the system used to record the biosignals must be nonintrusive to allow continuous recording without affecting the patient’s daily life. Smartwatches and smart wristbands are especially well suited for such an application, as in addition to being non-intrusive, they are easy to set up, commonly available, and relatively inexpensive. Consequently, this work focuses on the problem of manic–euthymic automatic state recognition using biosignals recorded from a wrist-worn wearable. For state recognition, the sensors considered are a three-axis accelerometer (actigraphy), an EDA, and a photoplethysmography (PPG) [from which the heart rate (HR) can be derived]. Furthermore, this work considers the setting where no data for training are available from the patient that is to be predicted on. This is necessary for the system to be calibration-free and user-independent.

The literature on state recognition in affective disorders primarily focuses on feature engineering, with the goal of characterizing a segment generated from a given modality (e.g., HR, actigraphy, speech) in a discriminative way. While these types of approaches have been shown to be able to discriminate between different states, they often do not explicitly consider the temporality of the characterized segment. Contrarily, time series classification (TSC) algorithms are made specifically to leverage this temporal information. For multivariate TSC, InceptionTime is a method based on convolutional networks, which was shown to achieve state-of-the-art results for real-time multivariate TSC applications. As such, one of this work’s contributions is to divide a multimodal segment into multiple subsegments, from which meaningful features are extracted before applying an InceptionTime-based architecture to perform automatic manic–euthymic state recognition for never-seen-before patients.

The type of information derived from the characterization of biosignals is dependent on the considered timespan (e.g., seconds, minutes, and hours). Consequently, another contribution of this work is to employ an ensemble of networks, which are fed features extracted from both minute-long and hour-long intervals to leverage the information extracted from both horizon lengths. Note that in this work, the term network is used in a machine learning context to refer to deep learning models.

DATA ACQUISITION AND PREPROCESSING

Participants and Data Acquisition
As a first step in the goal of automatically detecting manic episodes, this work focuses on a dataset that was recorded in a two-phased clinical study of bipolar disorder. All participants were diagnosed according to the International Classification of Diseases (ICD)-10. A group of 58 participants was included and of these 28 were recorded when hospitalized due to an ongoing manic episode (ICD-10 diagnosis F31.1 (current episode manic without psychotic symptoms) and F31.2 (current episode manic with psychotic symptoms)). The clinical psychiatrists residing at the two closed affective wards at Haukeland University Hospital suggested potential candidates after assessing their ability to consent.

A group of 30 euthymic patients, not overlapping with the manic group, were also included in this study. These participants were additional participants from the first part of the study (when discharged from the hospital) or recruited from the hospitals’ outpatient clinic or the local advocacy group for bipolar disorder.

Inclusion criteria for both phases of the study were Norwegian-speaking individuals between 18 and 70 years diagnosed with bipolar disorder, able to comply with instructions, and having an IQ above 70. Exclusion criteria were previous head trauma needing hospital treatment, having an organic brain disorder, substance dependence (excluding nicotine), or being in a withdrawal state. The study protocol was approved by The Norwegian Regional Medical Research Ethics Committee West (2017/937). Written informed consent was obtained from all participants and no financial compensation/treatment perks were provided. All patients (except two in the euthymic group) were taking various combinations of prescribed medications.
The patients' mood states were established at inclusion and at regularly repeated clinical assessments using the Young Mania Rating Scale (YMRS).\(^1\) YMRS rates the severity of mania based on clinical observations and the patients' subjective description of their state. The total score spans from 0 to 60, and a YMRS score below 10 is considered as being in remission, or in an euthymic state.\(^6\) The participants were also assessed with the Montgomery Asberg Depression Rating Scale (MADRS),\(^1\) a commonly used scale for measuring the presence and severity of ongoing depression. MADRS scores span between 0 and 60, and scores below 10 are defined as the absence of depression.\(^9\) For the euthymic participants, the bipolar diagnosis was validated using the Mini-International Neuropsychiatric Interview (MINI) version 6.0.0.\(^10\)

Table 1 presents the demographic characteristics for both groups.

The data used in this work were recorded with the Empatica E4 wristband worn on the dominant wrist.
for 24 h. The device provides a three-axis accelerometer, an EDA sensor, a skin temperature sensor, and a PPG.

Postrecording Exclusion
This study aimed to limit the impact of the recording process on the participants’ behavior. Therefore, besides being asked to wear the smart wristband, participants continued their treatment unhindered by the research protocol. Consequently, depending on when the next day’s assessment took place, the total recording period varied between participants and could span less than 24 h. In addition, some participants removed their wristband during recording, sometimes multiple times and for multiple hours. Therefore, manual segmentation based on skin temperature and accelerometer was performed to identify and remove the data recorded when the wristband was off. Because of these two factors aggregating, three participants (all manic) did not reach the minimum amount of data defined within this study (> 20 h) and were not considered when reporting results.

Dataset Segmentation
Acquiring data in a clinical context is a laborious process, often making the creation of large datasets impractical. Furthermore, as the samples are not independent and identically distributed, special care has to be taken to avoid data leakage (i.e., the information contained within the test set indirectly being used during training). Consequently, within this work, data are compartmentalized such that samples from the same individual will only be considered within the same set (i.e., train/validation/test set). Furthermore, a subset of the recorded dataset was reserved for data exploration, architecture building, and hyperparameter optimization. This subset, dubbed the exploration dataset, is comprised of three manic and five euthymic randomly selected participants. An additional two manic participants come from two of the three previously excluded individuals (as > 18 h of recording was available for both). This was done to minimize the number of participants that had to be taken out and to leverage otherwise discarded data.

The dataset containing the remaining 47 participants (22 manic and 25 euthymic) will be referred to as the main dataset. Due to the limited amount of participants, leave-one-out cross-validation is employed for evaluating the different methods considered. In other words, to evaluate a classifier, 47 independent rounds of training will be performed where the held-out test set will correspond to a different individual each time. Furthermore, the exploration dataset is concatenated with the main dataset’s training set to increase the amount of training data, which can facilitate better generalization. Due to the stochastic nature of the considered algorithms, all results are reported as an average of 20 runs.

DATA PROCESSING
The following section details the data processing employed for each modality and presents the different feature sets considered. Note that skin temperature can be influenced by external factors (e.g., ambient temperature), which can lead to data leakage (e.g., higher room temperature on average for a given group). As this factor was not controlled for, skin temperature’s contribution in distinguishing the mood state was not investigated.

Processing of the Different Modalities
Data processing of the biosignal was facilitated by the NeuroKit2\textsuperscript{11} library in Python.

Electrodermal Activity
The EDA employed in the wristband has a sampling frequency of 4 Hz and a range between 0.01 and 100 $\mu$Siemens.

During processing, a low-pass Butterworth filter of order 4 at 1.5 Hz is applied to better capture both the tonic and phasic components of the signal.\textsuperscript{12} From the cleaned signal, a high-pass Butterworth filter of order 2 at 0.05 Hz is applied to extract the phasic component of the signal.\textsuperscript{11,12} Skin conductance response (SCR) peaks are then identified by extracting the local maxima of the filtered signal, rejecting peaks with an amplitude below 10% of the standard deviation from the mean of the amplitude, as implemented in Makowski et al.’s work.\textsuperscript{11}

Photoplethysmograph and HR
The wristband’s PPG employs green and red light-emitting diode. The E4 uses a black box algorithm to fuse the information retrieved from the green and red exposures to limit the impact of motion artifacts. The black box algorithm’s output is what is made available at a sampling rate of 64 Hz. Within this work, a bandpass Butterworth filter of order 3 was applied between 0.5 and 8 Hz to the signal. The systolic peaks were then extracted from the filtered signal based on the method described in Elgendi et al.’s work\textsuperscript{13} and implemented in Makowski et al.’s work.\textsuperscript{11} The distances between these peaks are referred to as NN to
emphasize the fact that abnormal beats have been removed.\textsuperscript{7}

The HR is also made available by the E4 at a sampling rate of 1 Hz and represented the average HR values computed in a span of 10 s.

**Actigraphy**

The three-axis accelerometer integrated in to the E4 has a range of ±2 g and is cadenced at 32 Hz. For each participant, each data point was processed as follows:

\[
\sqrt{x^2 + y^2 + z^2} - 1g
\]

where \(x, y,\) and \(z\) represent the recorded value for their associated axis and \(1g\) represents the gravitational constant.

**FEATURE EXTRACTION**

**EDA Feature Set**

Two features were extracted from the EDA modality. First, the autocorrelation with a lag of 4 was computed from the filtered low-pass EDA signal, as suggested in van Halem et al.’s work.\textsuperscript{14} The second feature was extracted by taking the mean amplitude of the SCR peaks.

**Heart Rate Variability (HRV) Feature Set**

The sample entropy (SampEn) was extracted to measure the level of predictability in successive NN intervals.\textsuperscript{7} The standard deviation of the NN intervals (SDNN)\textsuperscript{7} was also calculated. Note that popular features, such as the root mean square of successive differences (RMSSD)\textsuperscript{7} and low-frequency/high-frequency ratio\textsuperscript{7} were not considered as they are particularly noisy when computed from a PPG signal.\textsuperscript{15} Consequently, the feature set extracted for the HRV is as follows:

\[
[\text{SDNN, SampEn}].
\]

**Actigraphy and HR Feature Sets**

Multiple feature sets were considered for the characterization of both the processed actigraphy and HR.

**Bipolar Complexity Variability (BCV) Features Set**

The BCV feature set is derived from Jakobsen et al.’s work,\textsuperscript{16} and is defined as follows:

\[
\left[\frac{\sigma}{\mu} \cdot \frac{\text{RMSSD}}{\text{SD}}, \text{SampEn}\right]
\]

where \(\mu\) and \(\sigma\) correspond to the mean and standard deviation of the signal, while RMSSD corresponds to the root mean square of successive difference.

**Temporal-Spatial Descriptors (TSD)**

The initial features proposed by Khushaba et al.\textsuperscript{17} are considered as a feature set and referred to as TSD. TSD consists of the root squared zero, second and fourth moments as well as the sparseness, irregularity factor, coefficient of variation, and the Teager–Kaiser energy operator.

In addition, a new feature set proposed in this work is the combination of TSD with BCV, which will be referred to as the TSD-BCV feature set.

**MOOD-STATE CLASSIFICATION METHODS**

Two types of intervals from which to compute the different feature sets are considered: long (20 h) and short (5 min). The following section provides a thorough description of the classifiers used for both intervals and their combinations.

**Long-Interval Classification**

Sequences lasting 20 h were selected in this work as a balancing act between including as many of the participants as possible for evaluation (as their recording needed to be at least that long) and is as close to a full day cycle as possible. The previously presented feature sets are thus computed directly from these long intervals for each participant. When considering multiple modalities simultaneously, features from each sensor are concatenated together into a single vector. As a form of data augmentation, a sliding window with an overlap of 19.5 h is applied to generate the examples from each participant. This data augmentation procedure resulted in an average of \(\approx 13\) examples per participant.

For each fold in the leave-one-out cross-validation scheme, each feature is scaled between -1 and 1 using min–max scaling. Note that the minimum and maximum values are obtained from the training set and the min–max normalization is performed on both the training and test sets. The following eight classifiers are then considered for mood-state classification: K-nearest neighbors (KNN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), decision tree (DT), random forest (RF), AdaBoost, and support vector machine (SVM) both with a linear and radial basis function (RBF) kernel. Class weights are balanced to account for the under/overrepresentation of a given class. Hyperparameter selection is performed using a random search with 50 candidates. The validation set employed for the random search is extracted from the current training set fold by randomly selecting two manic and two euthymic participants. The hyperparameters considered for each classifier are...
presented in Appendix A. The classifiers’ implementation comes from scikit-learn (v0.24.1) in Python.\textsuperscript{16}

**Short-Interval Classification**

Instead of characterizing the signal by extracting features over long intervals, this classification approach proposes considering much shorter intervals (five minutes) as sub-windows of the full example from which to extract the features. For each fold in the leave-one-out cross-validation scheme, each feature is then scaled between -1 and 1 using min–max scaling, as previously described. An example is then created by aggregating consecutive subwindows to form a $F_T \times W$ matrix, where $F_T$ represents the number of input features and $W$ is the number of sub-windows forming the example. The idea is then to perform feature learning via an InceptionTime network to discriminate between the different mood states. Note that due to the structure of the network’s architecture employed, it is possible to train with examples of varying lengths (i.e., number of subwindows). As such, the examples created vary in length between 20 and 24 h using increments of 40 min. In addition, examples were created with a sliding window using increments of 25 min. This data augmentation procedure yields an average training set containing \textasciitilde 4000 examples. From the exploration dataset, it was found that the best combination of sensors was obtained by combining EDA and actigraphy data (with the TSD-BCV feature set), each example has a shape varying between 11 $\times$ 240 and $11 \times 288$ (Feature $\times$ Time).

Figure 1(a) details the proposed network’s architecture, which is referred to as the short network. Ranger-\textsuperscript{Lars}\textsuperscript{19} is employed for the network’s optimization with a batch size of 128. The learning rate (lr=0.0037) was selected from the exploration dataset by random search using a uniform random distribution on a logarithmic scale between $10^{-6}$ and 1 with 50 candidates (each candidate was evaluated five times). Minibatches are built using a bucket approach, where sequences of the same length are grouped together. Early stopping, with patience of 20 epochs is applied by using 10% of the participants in the training set as a validation set (randomly selected). In addition, learning rate annealing, with a factor of 5 and patience of 10 was also used.

**Long–Short Interval Classification**

Features extracted from biosignals spanning different time intervals represent different characteristics of human behavior.\textsuperscript{4,7} Therefore, this work proposes leveraging features extracted from both short (five min) and long (20 h) periods. To do so, first, a short network is trained as described in the previous section. After training, the network’s weights are frozen and a second network is created, which is shown in Figure 1(b). This network takes the concatenation of the long-interval features and the learned features from the short network (directly after the global average pooling) as input and will be referred to thereafter as the short–long network. The short–long network’s architecture was built using the exploration dataset and the training procedure is as described in the previous section. Note, however, that this time, the interval length is static (20 h).

**Ensemble Method**

As mentioned in Fawaz et al.’s work,\textsuperscript{6} InceptionTime networks can exhibit high variance in terms of performance between training and therefore can benefit from an ensemble method approach. Consequently, this work also considers an ensemble of five networks for mood-state classification for both the short and the short–long networks. The predicted state will thus be the average prediction over the five networks’ output. These methods will be referred to as the short ensemble networks and the short–long ensemble networks, respectively.

It should be noted that ensemble approaches substantially augment both the training and inference time of the model. However, in the current context, mood states evolve over a period of orders of magnitude higher than the latency added by considering ensemble methods (which take less than a second for inference). Consequently, the considered ensemble approaches do not reduce the practical application of the proposed method within this work’s context.

The methods to extract the feature sets and networks implementation are available at GitHub.\textsuperscript{8}

**Experiments and Results**

In this article, accuracy represents the per-participant mean percentage of correctly classified classes averaged over all participants (i.e., each participant’s contribution to the accuracy score is weighted equally regardless of the number of examples provided by said participant). Note that, given the slight class imbalance on a per-participant basis of the considered dataset, a classifier systematically predicting the most common class would achieve an accuracy of 53.19%.

**Long Interval**

Figure 2(a) presents a comparison of the accuracy for mood-state recognition from the different modalities available on the E4 (and combinations of these...
modalities). For the sake of concision, only the best performing classifier and feature set for each sensor (and their combination) are reported (extended results are provided in Appendix B).

Short and Short–Long Interval

Figure 2(b) shows a bar graph comparing the best performing long-interval method (actigraphy with TSD-BCV using the linear-SVM classifier) against the short network, short–long network, and their ensemble variants.

Following García et al.’s work, a two-step statistical procedure using Friedman followed by Finner’s posthoc test was applied. First, Friedman’s test ranks the algorithms against each other. Then, Finner’s posthoc test is applied (n=47), using the best-ranked method as the control method. Finner’s null hypothesis is that the mean of the results of the control method against the other methods is equal (compared in pairs). The null hypothesis is rejected when \( p < 0.05 \). Overall, the short–long ensemble networks obtained the highest average accuracy at 91.59%±22.02% and was the best ranked. Furthermore, the difference between the short–long ensemble networks and the actigraphy long interval, short and short–long network was statistically significant (\( p = 0.01754 \), \( p < 0.00001 \), and \( p = 0.00001 \), respectively). No statistically significant difference was found between the short ensemble networks and the short–long ensemble networks. Appendix B details these results in a table format.
DISCUSSION

Bipolar disorder is a heterogeneous diagnosis. Consequently, although there are certain common diagnostic criteria, the disorder can manifest widely differently across humans resulting in large behavioral variations during a manic episode. This behavioral variability makes the task of mood-state recognition inherently challenging. Thus, the capability of automatically detecting mood states in people with bipolar disorders in an objective and nonintrusive way would vastly improve patient outcomes. This article proposes leveraging wrist-worn sensors in an effort to meet the challenge. From a clinical perspective, contrastingly to the current cross-sectional mood assessment methods, such an approach could reduce the resource burden and provide evaluations over longer time periods, thereby providing a more comprehensive view of the patients’ mood state.

For the long interval, 336 model combinations were tested (14 possible sensor combinations × 3 feature sets × 8 classifiers). Consequently, one should expect that some form of indirect overfitting took place. This was, however, necessary to get an overall and meaningful picture of the interaction of the different modalities with each other. In addition, these experiments enable this work to provide a competitive comparison basis of more traditional approaches against the proposed short and short–long networks and their ensemble variants.

The long-interval approach was not able to effectively leverage the EDA and HRV features when considered alone. In contrast, when combining the features extracted from the actigraphy and HRV, an average accuracy of 81.54% over 47 participants can be achieved. Similarly, from the results obtained on the exploration dataset, the combination of actigraphy+EDA was essential in achieving the best performance. When testing a version of the short network using only the actigraphy data, the performance degraded to around 79% compared to ~83% with the proposed EDA+actigraphy scheme. These results
The quality of the split was measured either by the log of the total number of features fed as input to the KNN: 1, 3, 5, 11, and 21. The metric distances considered were the Manhattan distance, the Euclidean distance, and the Minkowski distance of the third degree. DT: The quality of the split was measured either by the Gini impurity or its entropy. The maximum number of features considered were both the square root and the \( \log_2 \) of the total number of features fed to the decision tree. The tree could either have a maximum depth of 1, 2, 3, 5, and 10, or an infinite maximum depth. Finally, the minimum sample split was taken from a uniform distribution between 0 and 1. RF: The range of the number of trees considered was 10, 50, 100, 500, or 1000. The other considered hyperparameters were the same as for the DT classifier.

**CONCLUSION**

This article explores bipolar manic–euthymic state recognition using data collected from wrist-worn sensors. A new feature set for this task was proposed in the TSD-BCV, which borrows from both the affective disorder state recognition and the myoelectric-based hand gesture recognition literature. Leveraging actigraphy and HRV data in conjunction with the TSD-BCV, an L-SVM classifier was able to achieve an average accuracy of 81.54%\(\pm\)32.39% over 47 participants (22 manic and 25 euthymic). Furthermore, a new ensemble method comprised of short–long networks was able to achieve an average accuracy of 91.59%\(\pm\)22.02% on the same dataset by leveraging actigraphy and EDA data. Thus, showcasing the advantage of a multisensor approach for bipolar state recognition. As current diagnostic practices can be inaccurate and require expert involvement, our results in automatically predicting mood state in an unknown patient based on wristband data represent a meaningful step in the development of an instrument to facilitate early detection and intervention of manic episodes.

**APPENDIX A**

The hyperparameters considered for each classifier for the long-interval case were as follows.

- **KNN**: The number of possible neighbors considered were 1, 3, 5, 11, and 21. The metric distances considered were the Manhattan distance, the Euclidean distance, and the Minkowski distance of the third degree.
- **DT**: The quality of the split was measured either by the Gini impurity or its entropy. The maximum number of features considered were both the square root and the \( \log_2 \) of the total number of features fed to the decision tree. The tree could either have a maximum depth of 1, 2, 3, 5, and 10, or an infinite maximum depth. Finally, the minimum sample split was taken from a uniform distribution between 0 and 1.
- **RF**: The range of the number of trees considered was 10, 50, 100, 500, or 1000. The other considered hyperparameters were the same as for the DT classifier.
## TABLE 2. Average accuracy over the 47 participants of the best classifier combination for every feature set, sensors, and their combinations. For each participant, the average accuracy over 20 runs is given.

| Modalities | Feature set | Best classifier | Accuracy  | SD      | Friedman’s rank | H0 (p-value) |
|------------|-------------|-----------------|-----------|---------|-----------------|--------------|
| Actigraphy | BCV         | L-SVM           | 78.02%    | 35.73%  | 2.06            | 1            |
|            | TSD         | L-SVM           | 79.23%    | 36.45%  | 2.02            | 1            |
|            | TSD-BCV     | L-SVM           | 79.80%    | 33.71%  | 1.92            | -            |
| EDA        | BCV         | AdaBoost        | 54.13%    | 46.57%  | 2.02            | 1            |
|            | TSD         | RBF-SVM         | 57.16%    | 45.76%  | 2.05            | 1            |
|            | TSD-BCV     | AdaBoost        | 56.81%    | 46.29%  | 1.93            | -            |
| HRV        | BCV         | KNN             | 57.15%    | 38.48%  | 2.02            | N/A          |
|            | TSD         | AdaBoost        | 58.45%    | 47.27%  | 1.94            |              |
|            | TSD-BCV     | AdaBoost        | 57.94%    | 47.12%  | 2.04            | N/A          |
| HR         | BCV         | LDA             | 66.00%    | 43.14%  | 1.94            |              |
|            | TSD         | L-SVM           | 66.38%    | 45.28%  | 2.13            | 1            |
|            | TSD-BCV     | LDA             | 66.00%    | 43.14%  | 1.94            | -            |
| Act. + EDA | BCV         | L-SVM           | 73.86%    | 36.94%  | 2.13            | 1            |
|            | TSD         | L-SVM           | 76.73%    | 38.18%  | 1.96            | 1            |
|            | TSD-BCV     | L-SVM           | 76.06%    | 36.98%  | 1.92            | -            |
| Act. + HRV | BCV         | L-SVM           | 81.41%    | 32.20%  | 1.95            | -            |
|            | TSD         | QDA             | 75.69%    | 36.60%  | 1.97            | 1            |
|            | TSD-BCV     | L-SVM           | 81.54%    | 32.39%  | 2.09            | 1            |
| Act. + HR  | BCV         | L-SVM           | 74.72%    | 38.55%  | 2.03            | 1            |
|            | TSD         | L-SVM           | 76.49%    | 38.71%  | 1.99            | 1            |
|            | TSD-BCV     | LDA             | 74.56%    | 39.74%  | 1.98            | -            |
| EDA + HRV  | BCV         | AdaBoost        | 55.80%    | 46.58%  | 1.99            | 1            |
|            | TSD         | AdaBoost        | 55.22%    | 46.59%  | 1.90            | -            |
|            | TSD-BCV     | KNN             | 54.60%    | 43.40%  | 2.11            | 1            |
| EDA + HR   | BCV         | LDA             | 66.54%    | 44.59%  | 1.96            | -            |
|            | TSD         | KNN             | 66.91%    | 41.22%  | 2.09            | 1            |
|            | TSD-BCV     | LDA             | 66.54%    | 44.59%  | 1.96            | -            |
| HRV + HR   | BCV         | RBF-SVM         | 63.02%    | 44.53%  | 2.00            | 1            |
|            | TSD         | L-SVM           | 62.08%    | 45.86%  | 1.92            | -            |
|            | TSD-BCV     | KNN             | 63.69%    | 41.99%  | 2.08            | 1            |
| Act. + EDA + HRV | BCV | L-SVM | 77.45% | 34.14% | 2.00 | 1 |
|            | TSD         | QDA             | 74.32%    | 36.04%  | 1.90            | -            |
|            | TSD-BCV     | L-SVM           | 74.66%    | 36.15%  | 2.10            | 1            |
| Act. + EDA + HR | BCV | RBF-SVM | 73.92% | 40.96% | 1.93 | - |
|            | TSD         | L-SVM           | 71.32%    | 41.77%  | 2.00            | 1            |
|            | TSD-BCV     | AdaBoost        | 73.14%    | 39.84%  | 2.07            | 1            |
| EDA + HRV + HR | BCV | QDA | 59.98% | 45.95% | 1.98 | 1 |
|            | TSD         | KNN             | 60.96%    | 43.17%  | 1.94            | -            |
|            | TSD-BCV     | RBF-SVM         | 60.83%    | 46.78%  | 2.09            | 1            |
| Act. + EDA + HRV + HR | BCV | L-SVM | 73.03% | 35.76% | 1.89 | - |
|            | TSD         | L-SVM           | 69.91%    | 41.97%  | 2.09            | 1            |
|            | TSD-BCV     | L-SVM           | 71.76%    | 40.46%  | 2.02            | 1            |

Notes: Two-step statistical procedure using Friedman’s rank test followed by Finner’s posthoc test using the best ranked method as comparison basis. Null hypothesis rejected when $H0=0$ (p < 0.05). The row containing the best performing (according to the Friedman’s rank) feature set is bolded for each modality.
**TABLE 3.** Comparison of the best found classification scheme for the long interval with the short and short-long networks and their ensemble variants. For each participant, the average accuracy over 20 runs is given.

|                      | Best overall long interval (actigraphy + HRV, L-SVM, TSD-BCV) | Short network | Short-long network | Short ensemble network | Short-long ensemble networks |
|----------------------|---------------------------------------------------------------|---------------|--------------------|-----------------------|-----------------------------|
| Accuracy             | 81.54%                                                        | 82.80%        | 84.89%             | 87.45%                | 91.59%                      |
| Standard deviation   | 31.53%                                                        | 20.92%        | 18.69%             | 27.44%                | 22.02%                      |
| Friedman’s rank       | 2.93                                                          | 3.83          | 3.63               | 2.50                  | 2.12                        |
| H0 (p-value)         | 0 (0.01754)                                                   | 0             | 0 (<0.00001)       | 1                     | –                           |
| Cohen’s Dz           | 0.26                                                          | 0.71          | 0.64               | 0.27                  | –                           |

Note: Two-step statistical procedure using Friedman’s rank test followed by Finner’s posthoc test using the best ranked method as comparison basis. Null hypothesis rejected when H0=0 (p < 0.05). The column with the best performing method is bolded.

**REFERENCES**

1. I. Grande, M. Berk, B. Birmaher, and E. Vieta, "Bipolar disorder," Lancet, vol. 387, no. 10027, pp. 1561–1572, 2016, doi: 10.1016/S0140-6736(15)00241-X.
2. A. F. Carvalho, J. Firth, and E. Vieta, "Bipolar disorder," New Eng. J. Med., vol. 383, no. 1, pp. 58–66, 2020, doi: 10.1056/NEJMra1906193.
3. M. Bauer et al., "Areas of uncertainties and unmet needs in bipolar disorders: Clinical and research perspectives," Lancet Psychiatry, vol. 5, no. 11, pp. 930–939, 2018, doi: 10.1016/S2215-0366(18)30253-0.
4. P. Jakobsen et al., "Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls," PLoS One, vol. 15, no. 8, 2020, Art. no. e0231995, doi: 10.1371/journal.pone.0231995.
5. D. Basset, "A literature review of heart rate variability in depressive and bipolar disorders," Australian New Zealand J. Psychiatry, vol. 50, no. 6, pp. 511–519, 2016, doi: 10.1177/0004867415622689.
6. H. I. Fawaz et al., "InceptionTime: Finding AlexNet for time series classification," Data Mining Knowl. Discov., vol. 34, no. 6, pp. 1936–1962, 2020, doi: 10.1007/s10618-020-00710-y.
7. F. Shaffer and J. Ginsberg, "An overview of heart rate variability metrics and norms," Front. Public Health, vol. 5, 2017, Art. no. 258, doi: 10.3389/fpubh.2017.00258.
8. G. S. Malhi, B. Ivanovski, D. Hadzi-Pavlovic, P. B. Mitchell, E. Vieta, and P. Sachdev, "Neuropsychological deficits and functional impairment in bipolar depression, hypomania and euthymia," Bipolar Disord., vol. 9, no. 1/2, pp. 114–125, 2007, doi: 10.1111/j.1399-5618.2007.00324.x.
9. C. Hawley et al. "Defining remission by cut off score on the Madrs: Selecting the optimal value,” J. Affect. Disord., vol. 72, no. 2, pp. 177–184, 2002, doi: 10.1016/S0165-0327(01)00451-7.

**MENTAL STATE, MOOD, AND EMOTION**

*AdaBoost:* The number of estimators was one of 1, 10, 50, 100, and 200. The learning rate was drawn from a logarithm uniform distribution between $10^{-3}$ and $10^{0}$.

* SVM: For both the linear and RBF kernel, the soft margin tolerance (C) was chosen between $10^{-4}$ and $10^{3}$ on a logarithm uniform distribution. In addition, for the RBF kernel, the $\gamma$ hyperparameter was also selected on a logarithm uniform distribution between $10^{-4}$ and $10^{3}$.

**APPENDIX B**

**Results Long Interval**

The best classifier obtained for every combination of sensors and feature set considered for the long interval is given in Table 2.

**Comparison Short, Long, and Short–Long Intervals**

Table 3 shows a comparison between the best performing combination of classifier/feature set/sensors for the long interval and the short, short-long, and their ensemble variants.

**ACKNOWLEDGMENTS**

This work was partially supported by the Research Council of Norway as a part of the INTROMAT project under Grant 259293.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by The Norwegian Regional Medical Research Ethics Committee West under Application no. 2017/937 and performed in line with the declaration of Helsinki.
10. D. V. Sheehan et al., "The Mini-International Neuropsychiatric Interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10," J. Clin. Psychiatry, vol. 59, no. 20, pp. 22–33, 1998.

11. D. Makowski et al., "NeuroKit2: A Python toolbox for neurophysiological signal processing," Behavior Res. Methods, vol. 53, no. 4, pp. 1689–1696, Feb. 2021, doi: 10.3758/s13428-020-01516-y.

12. R. Zangróniz, A. Martínez-Rodrigo, J. M. Pastor, M. T. López, and A. Fernández-Caballero, "Electrodermal activity sensor for classification of calm/distress condition," Sensors, vol. 17, no. 10, 2017, Art. no. 2324, doi: 10.3390/s17102324.

13. M. Elgendy, I. Norton, M. Brearley, D. Abbott, and D. Schuurmans, "Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions," PloS One, vol. 8, no. 10, 2013, Art. no. e76585, doi: 10.1371/journal.pone.0076585.

14. S. van Halem, E. Van Roekel, L. Kroencke, N. Kuper, and J. Denissen, "Moments that matter? On the complexity of using triggers based on skin conductance to sample arousing events within an experience sampling framework," Eur. J. Pers., vol. 34, no. 5, pp. 794–807, 2020, doi: 10.1002/per.2252.

15. C. H. Antink, Y. Mai, M. Peltokangas, S. Leonhardt, N. Oksala, and A. Vehkaoja, "Accuracy of heart rate variability estimated with reflective wrist-PPG in elderly vascular patients," Sci. Rep., vol. 11, no. 1, 2021, Art. no. 8123, doi: 10.1038/s41598-021-87489-0.

16. P. Jakobsen et al., "Complexity and variability analyses of motor activity distinguish mood states in bipolar disorder," PLoS One, vol. 17, 2021, Art. no. e0262232, doi: 10.1371/journal.pone.0262232.

17. R. N. Khushaba, A. H. Al-Timemy, A. Al-Ani, and A. Al-Jumaily, "A framework of temporal-spatial descriptors-based feature extraction for improved myoelectric pattern recognition," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 10, pp. 1821–1831, Oct. 2017, doi: 10.1109/TNSRE.2017.2687520.

18. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

19. M. Grankin, "Optimizers against Imagettenet/Imagewoof dataset," 2019. [Online]. Available: https://github.com/mgrankin/over9000

20. S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," Inf. Sci., vol. 180, no. 10, pp. 2044–2064, 2010, doi: 10.1016/j.ins.2009.12.010.

**MENTAL STATE, MOOD, AND EMOTION**

**ULYSSE CÔTÉ-ALLARD** is currently a researcher with the University of Oslo, Oslo, Norway. His research interests include rehabilitation engineering, biosignal-based control, and human–robot interaction. Côté-Allard received the Ph.D. degree in electrical engineering from Université Laval, Québec City, QC, Canada, in 2020. He is the corresponding author of this article. Contact him at ulysseca@ifi.uio.no.

**PETTER JAKOBSEN** is currently a Ph.D. candidate with the University of Bergen, Bergen, Norway. His research interests include e-health and the characterization of bipolar disorder from actigraphy. Jakobsen received the master’s degree in applied science from Bergen University, Bergen, Norway, in 2015. Contact him at petter.jakobsen@helse-bergen.no.

**ANDREA STAUTLAND** is currently a medical doctor and also working toward the Ph.D. degree in medicine/clinical medicine with the University of Bergen, Bergen, Norway. Since 2017, she has been researching bipolar disorder. Her research interests include biomarkers and personalized medicine. Contact her at andrea.stautland@uib.no.

**TINE NORDGREEN** is currently an associate professor with the University of Bergen, Bergen, Norway, and also the leader of the INTROMAT Project. Her research interests include developing evidence-based Internet-based treatment within mental health. Nordgreen received the Ph.D. degree in clinical psychology from the University of Bergen, Bergen, Norway, in 2011. Contact her at tine.nordgreen@uib.no.

**OLE BERNT FASMER** is currently a professor with the University of Bergen, Bergen, Norway. His research interests include mathematical analyses of time series and the biological aspects of bipolar disorder, and ADHD. Fasmer received the M.D. and Ph.D. degrees from the University of Bergen, Bergen, Norway, in 1977 and 1988, respectively, both in clinical medicine. Contact him at Ole.Fasmer@uib.no.

**KETIL JOACHIM ØDEGAARD** is currently a professor with the University of Bergen, Bergen, Norway, and also leads the Affective Research Group, Bergen, Norway. His research interests include affective disorders and genetics. Ødegaard received the M.D. and Ph.D. degrees from the University of Bergen, Bergen, Norway, in 1994 and 2006, respectively, both in clinical medicine. Contact him at Ketil.Odegaard@uib.no.

**JIM TØRRESEN** is currently a professor with the University of Oslo, Oslo, Norway. His research interests include machine learning, bioinspired computing, and robotics. Tørresen received the Ph.D. degree in computer architecture and design from the Norwegian University of Science and Technology, Trondheim, Norway, in 1996. He is a senior member of the IEEE. Contact him at jimtoer@ifi.uio.no.