Depression Level Classification Using Machine Learning Classifiers Based on Actigraphy Data

JUNG-GU CHOI[1], INHWAN KO[2], AND SANGHOON HAN[1,2]

[1]Yonsei Graduate Program in Cognitive Science, Yonsei University, Seoul 03722, Republic of Korea
[2]Department of Psychology, Yonsei University, Seoul 03722, Republic of Korea

Corresponding author: Sanghoon Han (sanghoon.han@yonsei.ac.kr)

ABSTRACT Estimating and classifying depression status are critical in the clinical and psychological domains to map the course of treatment. Prior researchers used biosignal time-series data to reflect the variation in factors associated with depression. In addition, machine learning algorithms were applied to determine the underlying relationships between depressive symptoms and these factors. In this study, we introduce a classification framework for depression levels using actigraphy data based on machine learning algorithms. Fourteen circadian rhythm features (minimum, amplitude, alpha, beta, acrotime, upmesor, downmesor, mesor, f_pseudo, interdaily stability (IS), intradaily variability (IV), relative amplitude (RA), M10, and L5) extracted from accelerometer-based actigraphy data were used to model depression status with survey variables. Six evaluation metrics (accuracy, precision, recall, F1-score, receiver operating characteristic curve, and area under the curve) were applied to validate the performance of the proposed framework. Among the four candidate classifiers (XGBoost classifier, support vector classifier, multilayer perceptron, and logistic regression), the XGBoost classifier was the best at classifying depression levels. Moreover, we confirmed that the actigraphy data of two days were optimal for feature extraction and classification. The results of this study provide novel insights into the relationship between depression and physical activity in terms of both identification of depression and application of actigraphy data.

INDEX TERMS Accelerometer-based actigraphy data, classification algorithm, circadian rhythm, depression level, machine learning, multi-level classification, physical activity.

I. INTRODUCTION

Depression is commonly recognized as the main factor in psychiatric and psychophysiological disorders [1]. In addition, depressive symptoms affect daily life, leading to feelings of helplessness, anxiety, sleep disturbances, and decreased concentration [2]. To identify factors associated with related disorders, many researchers have focused on both internal and external elements of the patient group. For example, chronic stress affects the onset of major depressive disorder [3], acute stroke is associated with the occurrence of major depression [4], and socio-environmental factors (e.g., family members' health and relationships) affect unipolar depression [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan[ID].

Classifying the depression level is critical in both the clinical and psychological domains. In the clinical domain, continuously monitoring depressive symptoms is important for mental health management. In addition, it is beneficial for all patient groups, including the treatment group and the candidate group expected to receive treatment. In the psychological aspects, it is important that individual factors are considered to confirm a subject's depression level [6]–[9].

Many previous researchers utilized various modalities to identify factors associated with depression, such as electroencephalogram (EEG), brain magnetic resonance imaging (MRI), and self-report physical activity [10]–[13]. However, these methods have several limitations. In the case of EEG and brain MRI, an expensive MRI equipment should be prepared, and data can only be collected in a limited laboratory

environment. In terms of self-report physical activity, many researchers collected intensity or duration of activity by self-report questionnaires from participants. However, self-report measures can be influenced by self-report bias or response bias [14], [15]. To overcome these limitations, we used accelerometer-based actigraphy data (i.e., actigraphy) in our study. This time series data can collect continuous activity data and is relatively inexpensive.

Considering previous studies, biosignal and life-log time series data collected from participants (e.g., Electrocardiogram, accelerometer-based actigraphy) have been widely used to screen for depression [16]–[18]. Heart rate variability (HRV) parameters extracted from ECG signals were used with machine learning (ML) algorithms to diagnose major depressive disorder [19]. In addition, multimodal data (steps, energy expenditure, body movement, sleep time, heart rate (HR), and skin temperature) including several life-log time series data collected from wristband-type wearable devices were analyzed using a ML model to evaluate depression [20].

In related studies using self-report physical activity, researchers focused on comparing the intensity level (e.g., weak, moderate, or vigorous) and activity duration (e.g., once a week or five hours a day) of participants in study groups [21]–[23]. Moreover, actigraphy time-series data collected using wearable devices were used to analyze activity variation [24]–[26]. To deduce the activity patterns from actigraphy, the characteristics of actigraphy time series (e.g., peak, slope, amplitude) were calculated from cosinor modeling as indices [27]–[30]. These indices represent the circadian rhythm cycles of the participants, which reveal major symptoms of depression [31]–[33].

Circadian rhythm patterns from actigraphy data reflect the stability or fluctuation of activity from participants [34], [35]. These patterns can be summarized using parametric and non-parametric metrics [36]–[39]. In previous studies, circadian indices related to activity, sleep disturbances, and daily mood variations were compared between the patient groups and control groups to identify differences in the circadian patterns of depression patients [40]. Disturbances of the circadian cycle indicated that depression patients exhibited differences compared to the healthy group [41], [42]. Moreover, ML algorithms have been used to characterize daily sleep-activity cycles using actigraphy data [43].

Based on previous studies identifying the association between physical activity and depression, we hypothesized that actigraphy data have sufficient potential to classify depression levels when an adequate analysis is conducted. To prove our hypothesis, we constructed an experimental design using ML classification algorithms. In this study, we proposed a classification framework for depression levels (e.g., 'mild,' 'moderate,' or 'severe' depression) through ML algorithms based on actigraphy data. We collected the actigraphy data (Fig. 1), as well as demographic, physical activity, subjective health status, and mental health variables of the same participants. Fourteen circadian rhythm indices based on parametric and non-parametric metrics were

extracted to verify the validity of these data as features. Four non-parametric features were used: interdaily stability (IS), intradaily variability (IV), total activity of the ten most active hours (M10), and total activity of the five least active hours (L5). The nine parametric features used were minimum, alpha, beta, acrotime, amplitude, mesor, upmesor, downmesor, and f_pseudo. The distribution of all features, including the survey variables and extracted circadian features, was verified; in sequence, they were log-transformed and standardized. To prevent multicollinearity between features, we selected only fifteen features based on the lasso and ridge regression models. The selected features were applied to four classification algorithms, XGBoost classifier, support vector classifier (SVC), logistic regression (LR), and multilayer perceptron (MLP). Finally, the performance of each model was evaluated using six evaluation metrics: accuracy, precision, recall, F1-score, receiver operating characteristic (ROC) curve, and area under the curve (AUC).

The objective of this study was to develop a depression level classification framework based on actigraphy data with ML algorithms. The major contributions of this study are as follows:

(1) We proposed an ML-based classification framework for depression levels based on the circadian rhythm characteristics embedded in physical activity. In addition, we evaluated the performance of the model under various conditions, including binary and multiclass classification, on a large dataset. Moreover, we compared the classification performance of commonly used ML classification algorithms: XGBoost classifier, SVC, MLP classifier, and LR;

(2) Advancing from analyzing simple characteristics of physical activity data, we extracted various features about the inherent circadian rhythm of activity from participants. Furthermore, we identified the optimal length of actigraphy data for extraction of circadian rhythm features. In addition, we confirmed that the feature importance of the trained XGBoost model in our framework was in agreement to that of previous studies.

The remainder of the paper is organized as follows: Section II includes a detailed description of the dataset and methodologies used in the study. In Section III, the classification performance of the four ML models in the experiments is reported. In Section IV, we discuss the results and their implementation. Finally, the conclusions and summary of our study are presented in Section V.

II. METHODS
A. OVERVIEW
To prove our hypothesis, we used six steps in our experimental design. First, we extracted demographic, physical activity, mental health, subjective health status variables, and collected actigraphy data from the Korea National Health and Nutrition Examination Survey (KNHANES) dataset. Selected variables and actigraphy data were combined based on the participant ID. Second, we extracted thirteen circadian indices
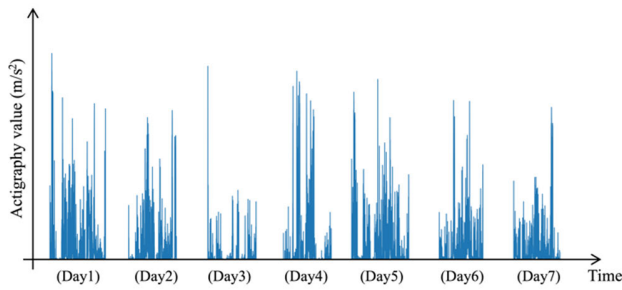
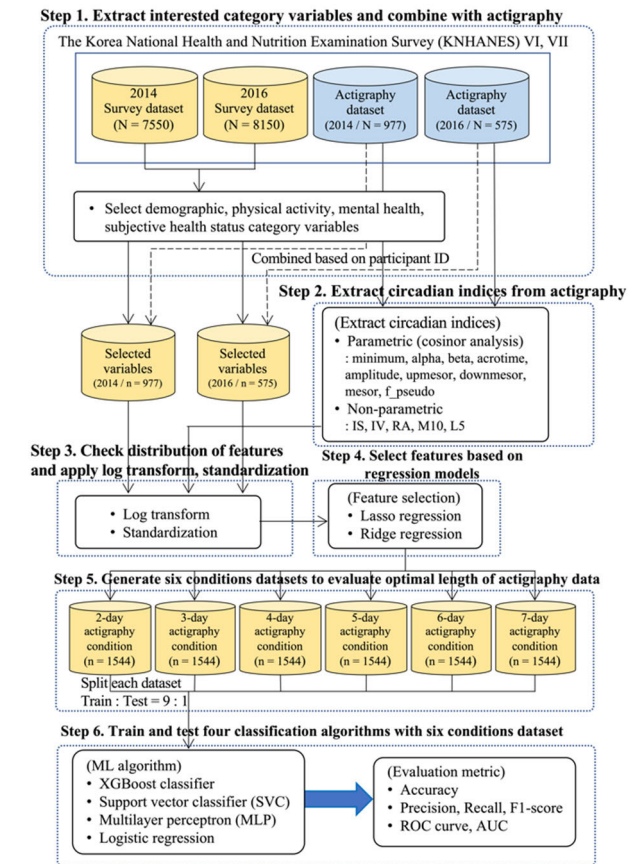**FIGURE 1.** Example of actigraphy data used in this study.

FIGURE 2. Overview of experimental design in this study.

from the actigraphy data using parametric and non-parametric methods. Third, selected variables were log-transformed and standardized after their distribution was checked. Fourth, to select suitable features, all the variables, including selected variables and extracted circadian indices, were filtered using the coefficient with the lasso and ridge regression models. Fifth, we generated six types of datasets to evaluate the optimal length of the actigraphy data used in an ML algorithm. Finally, four classification algorithms were trained and evaluated using the evaluation metrics. The detailed steps are shown in Fig. 2.

B. DATA SOURCES

In our study, we used the open-source KNHANES dataset released by the Korea Disease Control and Prevention

TABLE 1. Abbreviations used in this article.

Abbreviation	Full name
ML	Machine learning
EEG	Electroencephalogram
HR	Heart rate
HRV	Heart rate variability
ECG	Electrocardiogram
MLP	Multilayer perceptron
SVC	Support vector classifier
ROC	Receiver operating characteristic
AUC	Area under the curve
TP	True positive
TN	True negative
FN	False negative
FP	False negative
IS	Interdaily stability
IV	Intradaily variability
M10	Most active 10-h period
L5	Least active 5-h period
RA	Relative amplitude
PHQ-9	Patient health questionnaire-9
BMI	Body mass index
KNHANES	The Korea National Health and Nutrition Examination Survey
KDCA	The Korea Disease Control and Prevention Agency

Agency (KDCA). KNHANES is a longitudinal survey conducted by the KDCA to investigate the health status, health-related awareness and behavior, and nutritional status of people in Korea. The survey started in 1998 and was conducted every three years until 2005; subsequently, it has been conducted annually. The original dataset is available on the KDCA website and the current dataset for 18 years from 1998 to 2019 is publicly available. This dataset consists of nine categories of survey variables, which are listed in Table 2.

This dataset consists of two sub-datasets in separated csv files. The first sub-dataset consists of health behavior, blood tests, and grip test results covering the first four categories. The second sub-dataset includes the last five categories. All the survey results in the sub-datasets can be combined with the participant ID. A total of 216,815 people participated in this survey from 1998 to 2019. In our study, we selected the datasets for two years only (2014 and 2016), which were the only ones containing actigraphy data [44], [45]. The baseline characteristics of both datasets are presented in Table 3.

The actigraphy data in the KNHANES dataset were collected using an ActiGraph GTX3 wearable device. Acceleration values from the participants' activity were measured in one sample in one minute (sampling rate: 1/60 Hz). All the participants were instructed to wear the device for one week. Although the acceleration values of the three axes

TABLE 2. Categories of variables in KNHANES dataset.

No.	Categories	Types
1	Health behavior	Categorical
2	Blood pressure measurement	Continuous
3	Blood test	Continuous
4	Grip strength test	Continuous
5	Dietary life survey	Categorical
6	Food safety investigation	Categorical
7	Food intake frequency survey	Categorical
8	Food intake survey	Continuous
9	Dietary life evaluation index	Continuous

TABLE 3. Baseline characteristics of KNHANES dataset (2014, 2016).

Characteristic		KNHANES
Age (years), mean (SD)		42.1 (22.99)
No. of participants (N)		15 700
Gender, n (%)	Male	7 050 (44.90%)
	Female	8 650 (55.09%)
Height (cm), mean (SD)		156.1 (19.73)
Weight (kg), mean (SD)		57.1 (18.07)
BMI, mean (SD)		22.6 (4.21)
PHQ-9[a], mean (SD)		2.7 (3.86)

[a]PHQ-9: Depression status survey index

(X, Y, and Z) were collected, only single vertical values (Z axis) were published.

C. DATA PREPROCESSING

1) COLLECT AND COMBINE ACTIGRAPHY AND INTERESTED VARIABLES FROM DATASET

We analyzed whether the actigraphy data were in agreement with survey variables. Among the nine categories included in the survey data, demographics, physical activity, subjective health status, and mental health variables were selected to identify levels of depression. The dimensions of the original dataset in 2014 and 2016 were (7,550, 746), and (8,150, 800) before and (7,550, 73) and (8,150, 71), after the extraction of relevant variables, respectively. For the remaining dataset, we merged the actigraphy dataset and survey dataset based on the participant ID in the actigraphy dataset (977 participants in 2014 and 575 participants in 2016). Following merging, the dimensions of both datasets were unchanged.

2) EXTRACTION OF PARAMETRIC AND NON-PARAMETRIC CIRCADIAN INDICES

To classify the depression status, we extracted circadian rhythm indices from actigraphy time-series data. Both parametric and non-parametric metrics were applied to deduce the various characteristics of the circadian cycles from activity

patterns. To extract the indices, the duration of the actigraphy data collection was varied (two, three, four, five, six, and seven days) to confirm the optimal actigraphy length for index extraction. For each length condition, the data were sliced into windows and data of several length conditions were overlapped. For example, in a two-day length condition, a window with a length of two days was sequentially sliced from actigraphy data (e.g., the first window consists of days 1 and 2 from actigraphy data, the second window consists of days 2 and 3, ... In total, six windows were applied). Consequently, we obtained six index vectors from the two-day actigraphy data. The detailed process for the two-day length condition is depicted in Fig. 3.

The indices were extracted using the same process and also for the other conditions. After extraction, the indices were arranged in column-wise matrices. Each column was denoted as x_i, where $i = 1, \ldots, N$. The rows of the matrix indicate the extracted indices for the six conditions per participant. We termed this matrix the circadian index matrix, which is denoted as follows:

$$C = [(x_i, y_j), i = 1, \ldots, N, j = 1, \ldots, M] \qquad (1)$$

Finally, the corresponding condition labels vector $y = [$**two days**$, \ldots, $**seven days**$]$ were merged with the circadian indices matrix. A detailed description of each circadian rhythm index is provided in the following subsections.

a: NON-PARAMETRIC CIRCADIAN INDICES
[1] INTERDAILY STABILITY (IS)

In this study, the stability of activity over multiple days was calculated by normalizing the number of actigraphy samples at 24-h values. This indicator was calculated using (2) [46]:

$$IS = \frac{N \sum_{h=1}^{P} (\bar{X}_h - \bar{X})^2}{P \sum_{i=1}^{N} (X_i - \bar{X})^2} \qquad (2)$$

where N is the total number of samples, p is the number of samples per day, $\bar{X}$ is the mean value of all samples, $\bar{X}_h$ are the hourly means, and x_i indicates the individual actigraphy samples. Changes in IS can represent a coupling between the rest-activity cycle and decreased IS values indicated higher day-to-day variation in activity patterns [47].

[2] INTRADAILY VARIABILITY (IV)

The IV index is the ratio of the mean squared first derivative of the sample to the total variance from the actigraphy samples as in (3):

$$IV = \frac{N \sum_{i=2}^{N} (X_i - X_{i-1})^2}{(N - 1) \sum_{i=1}^{N} (\bar{X} - X_i)^2} \qquad (3)$$

The elements included in the equation have the same meaning as in (2). This index indicates fragmentation of the rest-activity rhythm [46].

[3] MOST ACTIVE 10-H PERIOD (M10)

The M10 was computed by averaging the ten highest hourly means. This index indicates the activity during the most active period of the day.
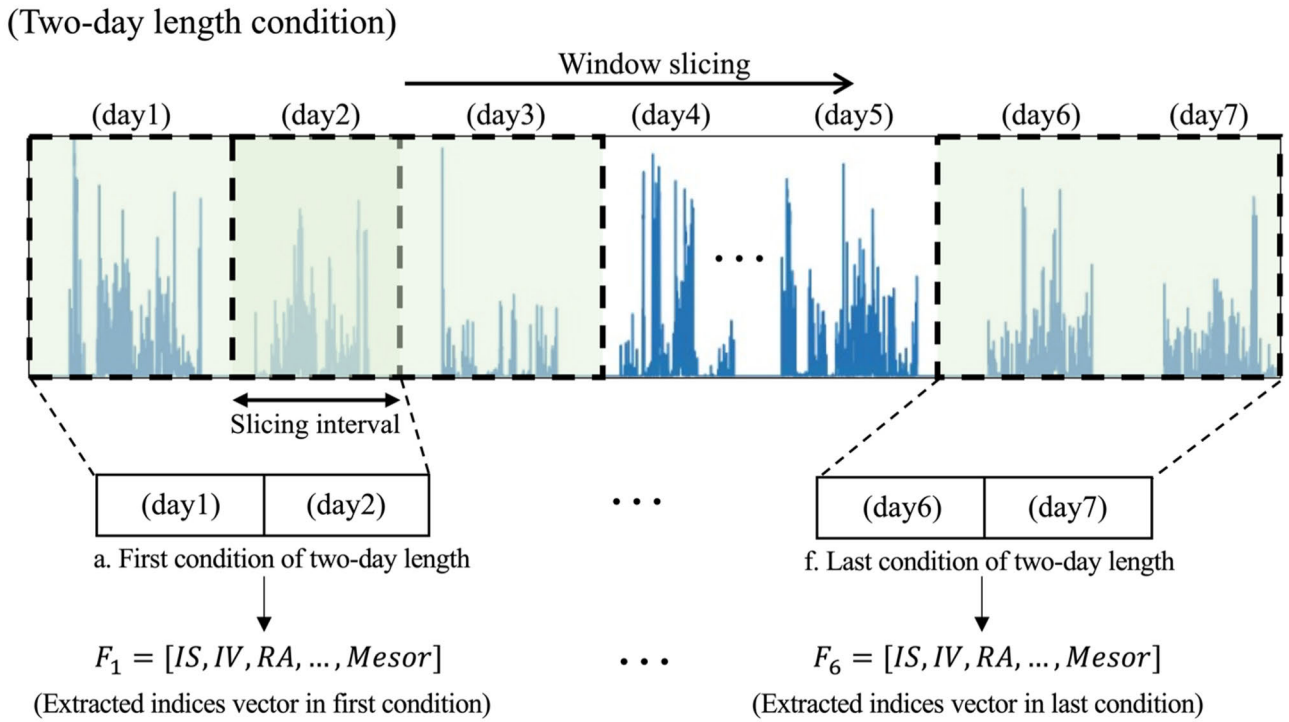
(Two-day length condition)

FIGURE 3. Example of the selection process in indices extraction (two-day length condition).

[4] LEAST ACTIVE 5-H PERIOD (L5)

The L5 represents movement during sleep and nighttime arousals. This value indicates the average value in the five least active hours in the entire actigraphy.

[5] RELATIVE AMPLITUDE (RA)

The RA of the activity cycle in actigraphy can be calculated from M10 and L5 values, as in (4) [48], [49]:

$$RA = \frac{M10 - L5}{M10 + L5} \tag{4}$$

b: PARAMETRIC CIRCADIAN INDICES

To extract parametric indices from the actigraphy data, we used cosinor analysis. The least squares method was used to fit a cosine wave to the actigraphy data [50]. We calculated nine parametric indices: minimum, amplitude, alpha, beta, acrotime, upmesor, downmesor, mesor, and f_pseudo.

[1] MINIMUM

This index is the minimum value of the fitted cosine function with actigraphy data.

[2] AMPLITUDE

This index represents the highest activity value in the activity cycle.

[3] ALPHA

Alpha indexes determine whether the peaks of the curve are wider than the troughs. High alpha values indicate wide troughs and narrow peaks. On the contrary, low alpha values indicate narrow troughs and wide peaks.

TABLE 4. Selected features from feature selection criteria.

Rank	Features (circadian indices and survey variables)	Frequency
1	EQ5D	42
2	BP1	42
3	LQ4_00	42
4	D_2_1	42
5	RA	41
6	HE_BMI	37
7	L5	35
8	mh_stress	34
9	alpha	31
10	DF2_dg	30
11	Sex	30
12	D_1_1	30
13	pa_aerobic	30
14	MESOR	29
15	minimum	24

[4] BETA

This index determines whether the transformed function rises and falls more steeply than the cosine curve. Large values of the beta index indicate that the curves are nearly square waves.
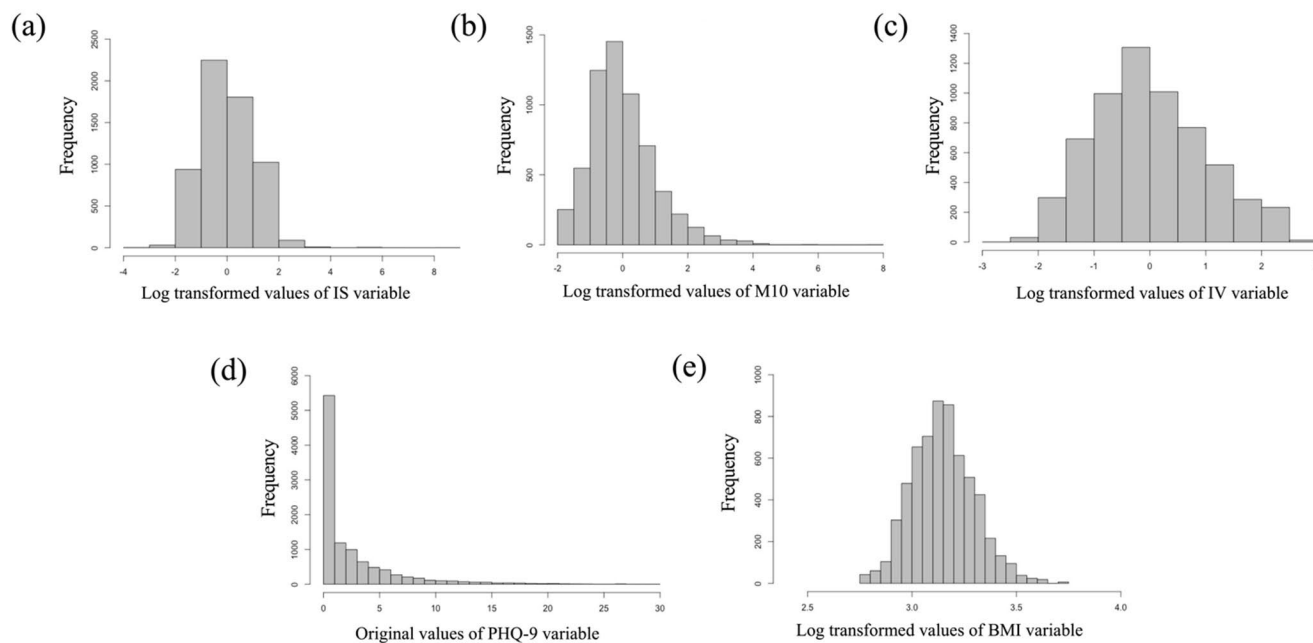
FIGURE 4. Distribution examples of demographic and circadian rhythm variables used in this study. Log-transformed distribution of (a) IS variables, (b) M10 variables, and (c) IV variables. (d) Distribution of original PHQ-9 variables. (e) Log-transformed distribution of BMI variables.

[5] ACROTIME

The acrotime indicates the time of the peak activity from the total activity time.

[6] UPMESOR

The upmesor is the time of the day in which the switch from low to high activity occurs. In the rest-activity rhythm, this value indicates the timing of the variation. Lower values indicate increased activity earlier in the day.

[7] DOWNMESOR

The downmesor is the time of the day in which the switch from high to low activity occurs. It indicates the timing of the change in the rest-activity cycle. Lower values represent a decline in activity.

[8] MESOR

This index, calculated similarly to the MESOR of the cosine model, can be calculated using (5).

$$MESOR = \frac{minimum + amplitude}{2} \qquad (5)$$

However, as it goes through the middle of the peak, it is not equal to the MESOR of the cosine model. Generally, this index represents the mean of the actigraphy data.

[9] F_PSEUDO

Measure the improvement of the fit obtained by nonlinear estimation of the transformed cosine model.

3) REMOVAL OF VARIABLES WITH NO-RESPONSE DATA

In the KNHANES dataset, -8 indicates the 'not applicable' answers of the participants. To reflect the exact response in each variable, we checked the distribution of each variable, including both the survey and circadian indices

TABLE 5. Hyperparameters of ML classifiers.

Model	Features (circadian indices and survey variables)	Value (argument)
XGBoost classifier	Eta	0.3
	Gamma	0
	max_depth	6
	min_child_weight	1
SVC	Kernel	rbf
	Gamma	auto
LR	Penalty	L2
	Solver	newton-cg
MLP	hidden_layer_sizes	10
	Activation	relu
	Solver	adam
	Alpha	0.0001
	learning_rate	0.001

matrices through histograms. We removed variables with more than half inapplicable responses, after which 33 variables remained.

TABLE 6. Classification performance results for classifiers with binary-class conditions.

Actigraphy condition	2 days				3 days				4 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	97.01	77.96	72.52	67.05	95.01	67.89	72.25	67.40	92.65	67.82	71.66	67.63
Precision	96.50	67.26	73.29	66.19	94.95	66.95	72.64	66.50	92.71	66.84	71.62	66.73
Recall	97.10	66.51	69.87	66.07	95.04	66.30	69.72	66.40	93.08	66.25	69.18	66.66
F1-score	96.83	66.68	70.18	66.10	94.93	66.45	70.05	66.42	92.53	66.39	69.51	66.67
AUC	99.00	91.10	76.30	76.00	99.00	90.00	75.00	74.00	99.00	91.00	74.00	73.00
Actigraphy condition	5 days				6 days				7 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	90.05	67.68	71.16	67.23	82.03	67.56	70.35	67.38	68.75	66.25	66.44	62.10
Precision	90.22	66.75	71.38	66.39	82.09	66.69	70.82	66.56	69.16	65.18	65.57	61.68
Recall	90.07	66.07	70.39	66.25	82.07	65.94	67.53	66.35	65.73	64.34	65.32	61.81
F1-score	90.10	66.19	70.62	66.21	82.07	66.08	67.71	66.37	65.55	64.38	65.28	61.48
AUC	96.00	86.00	73.00	71.00	89.00	76.00	67.00	64.00	68.00	60.00	68.00	66.00

AUC, area under the curve; SVC, support vector classifier; MLP, multilayer perceptron; LR, logistic regression.

TABLE 7. Classification performance results for classifiers with three-class conditions.

Actigraphy condition	2 days				3 days				4 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	95.61	84.67	69.65	67.87	93.79	70.14	84.43	68.28	92.89	69.30	84.17	68.09
Precision	95.77	77.99	55.72	53.23	93.88	54.24	82.02	52.49	92.62	54.55	69.45	52.30
Recall	95.97	56.23	64.67	59.03	93.69	63.21	55.41	58.21	91.26	62.05	53.14	58.30
F1-score	95.91	60.75	54.55	51.78	93.52	54.27	59.43	51.61	92.55	53.39	55.57	51.59
AUC	99.00	90.00	72.00	72.00	99.00	90.00	69.00	70.00	98.00	89.00	70.00	71.00
Actigraphy condition	5 days				6 days				7 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	83.86	70.64	83.48	67.76	76.01	70.06	82.94	67.85	81.83	66.93	77.41	67.05
Precision	85.26	55.29	64.38	52.12	78.66	56.09	53.47	51.36	50.43	52.47	48.07	49.80
Recall	83.17	62.35	51.51	57.09	72.42	62.57	50.35	55.25	51.64	51.37	45.97	48.51
F1-score	84.81	54.57	53.15	51.03	75.96	54.43	51.11	50.25	48.51	47.95	46.02	48.23
AUC	95.00	87.00	69.00	70.00	91.00	79.00	65.00	65.00	64.00	62.00	63.00	63.00

AUC, area under the curve; SVC, support vector classifier; MLP, multilayer perceptron; LR, logistic regression.

4) CHECK DISTRIBUTION AND TRANSFORM BY LOG TRANSFORMATION AND STANDARDIZATION

After removing the irrelevant variables with invalid responses, we confirmed the distributions of the remaining variables to improve the evaluation of the ML algorithms. In addition, we applied log transformation and z-score standardization to the arranged dataset to overcome possible unequal and skewed distribution of variables.

In the case of the 'PHQ-9' variable (target variable), which was discrete and not continuous, we could not apply log

transformation. Distributions of variables used in our study, including 'PHQ-9,' are depicted in Fig. 4.

5) FEATURE SELECTION

Features of both the extracted circadian rhythm indices and selected survey variables may have two problems. First, features can have a high correlation between them (multicollinearity or redundant variables in classifying dependent variables). Second, a low correlation can be observed between features and class (irrelevant features for classifying

TABLE 8. Classification performance results for classifiers with four-class conditions.

Actigraphy condition	2 days				3 days				4 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	95.80	58.10	71.23	53.28	92.85	58.70	71.76	53.76	89.80	57.65	70.96	53.43
Precision	94.66	43.75	60.11	39.34	92.64	43.58	61.03	39.85	89.38	42.94	56.39	39.82
Recall	95.74	48.13	43.29	42.51	91.05	47.90	43.84	43.43	82.86	46.58	41.58	42.43
F1-score	95.96	40.08	45.48	38.24	92.95	40.54	46.28	39.03	87.13	39.07	43.04	38.71
AUC	100.0	89.00	68.00	68.00	99.00	91.00	69.00	69.00	97.00	90.00	70.00	68.00
Actigraphy condition	5 days				6 days				7 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	84.72	58.00	70.51	53.43	71.72	57.59	69.84	52.56	62.07	54.99	67.28	52.04
Precision	84.91	43.12	52.86	39.80	71.47	43.68	43.19	39.39	34.89	40.26	35.77	37.31
Recall	84.41	46.78	40.04	43.22	59.19	45.40	38.62	41.44	33.33	38.61	35.29	39.52
F1-score	84.00	40.04	40.52	38.86	63.32	39.57	37.81	37.94	32.95	35.14	33.25	36.27
AUC	97.00	89.00	67.00	69.00	89.00	77.00	68.00	68.00	57.00	57.00	63.00	68.00

AUC, area under the curve; SVC, support vector classifier; MLP, multilayer perceptron; LR, logistic regression

TABLE 9. Classification performance results for classifiers under five-class conditions.

Actigraphy condition	2 days				3 days				4 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	93.82	37.64	55.50	37.17	91.86	38.57	55.14	37.54	88.57	37.58	54.49	38.49
Precision	94.75	32.81	46.11	31.75	91.32	33.98	40.08	32.62	88.58	33.90	43.95	33.98
Recall	94.91	33.99	33.11	32.80	89.01	34.62	33.01	33.86	81.95	34.61	32.48	34.42
F1-score	94.79	30.01	30.35	29.89	85.38	30.46	30.23	30.70	82.07	30.58	29.96	31.88
AUC	99.00	90.00	67.00	67.00	99.00	91.00	67.00	68.00	97.00	90.00	70.00	70.00
Actigraphy condition	5 days				6 days				7 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	83.02	36.98	54.45	38.05	70.90	35.62	53.54	37.05	43.39	33.65	51.87	33.65
Precision	83.91	33.81	38.78	33.64	71.10	33.75	36.60	32.62	26.74	28.96	27.38	30.10
Recall	81.95	34.66	32.72	33.85	71.28	34.34	31.83	32.68	27.33	30.17	29.77	29.14
F1-score	82.95	30.52	30.05	31.49	70.30	53.54	29.26	30.34	26.45	26.20	25.89	26.87
AUC	97.00	88.00	68.00	65.00	90.00	78.00	67.00	70.00	55.00	54.00	62.00	61.00

AUC, area under the curve; SVC, support vector classifier; MLP, multilayer perceptron; LR, logistic regression.

dependent variables). To select adequate features, we applied three-step rank and frequency feature selection methods. Two feature selection criteria, lasso and ridge regression models [51], [52], were applied.

In the feature selection steps, we first fitted the lasso and ridge regression models based on a dataset including both circadian indices and survey variables. The coefficients were confirmed for the individual features and sorted by their magnitude. Second, we selected the top 15 features based on each coefficient. All high-ranking features selected from the regression models were collected. Finally, the collected features were sorted again according to their frequency. After feature sorting, we chose the top-15 features to reflect both rank and frequency from both selection criteria. The features selected in this section are listed in Table 4.

Considering the selected features, 'EQ5D' indicates subjective quality of life index, 'BP1' indicates awareness of usual stress, 'LQ4_00' indicates uncomfortable physical activity, 'D_2_1' indicates uncomfortable experience in the last two weeks, 'HE_BMI' indicates the BMI index values, 'mh_stress' indicates stress awareness, 'DF2_dg' indicates the doctor's diagnosis about depression, 'D_1_1' indicates

TABLE 10. One-way ANOVA test results of evaluation indices.

Evaluation index	F-value	p-value
Accuracy	235.60	3.5e-10
Precision	224.01	5.6e-9
Recall	221.81	1.3e-12
F1-score	245.21	1.3e-9
AUC	308.21	2.5e-11

the subjective health status, and 'pa_aerobic' indicates aerobic physical activity.

6) GENERATION OF DATASET UNDER SIX CONDITIONS TO CONFIRM THE OPTIMAL LENGTH OF ACTIGRAPHY DATA

To confirm the optimal length of the actigraphy monitoring for classification, we constructed six datasets by varying the duration of the actigraphy data. The first dataset contained circadian rhythm indices extracted from the actigraphy data of two days. Similarly, the second to sixth datasets consisted of three, four, five, six, and seven days of actigraphy data, respectively. The numbers of rows in the first and second datasets were 9158, 7681, 6160, 4627, 3086, and 1544, respectively. Additionally, each dataset was split into training and test datasets at a 9:1 ratio.

7) EVALUATION OF CLASSIFICATION PERFORMANCE IN EACH CONDITION

In the final step, we constructed an additional dataset with various class conditions to compare the classification performance at diverse class levels. Classification performance was evaluated by four conditions: binary, three, four, and five classes.

Four classification algorithms (XGBoost, SVC, MLP, and LR) were compared in a total of 24 conditions (six actigraphy length conditions × four conditions for classification labels). To check the relevance as input features for depression level classification, we confirmed the list of features sorted by feature importance from trained ML models.

Due to the imbalance in the number of subjects belonging to each class label, weights were applied to complement the algorithm training. We conducted a random search to determine the optimal hyperparameters of the four ML classifiers, as listed in Table 5. In addition, 10-fold cross validations were applied to prevent overfitting of classification algorithms.

D. CLASSIFICATION ALGORITHMS

In this study, we utilized four classification algorithms to model the relationship between the selected features and the level of depression. The first classification model was the XGBoost classifier, which is based on an ensemble of several decision tree models, according to (6) [53]. This
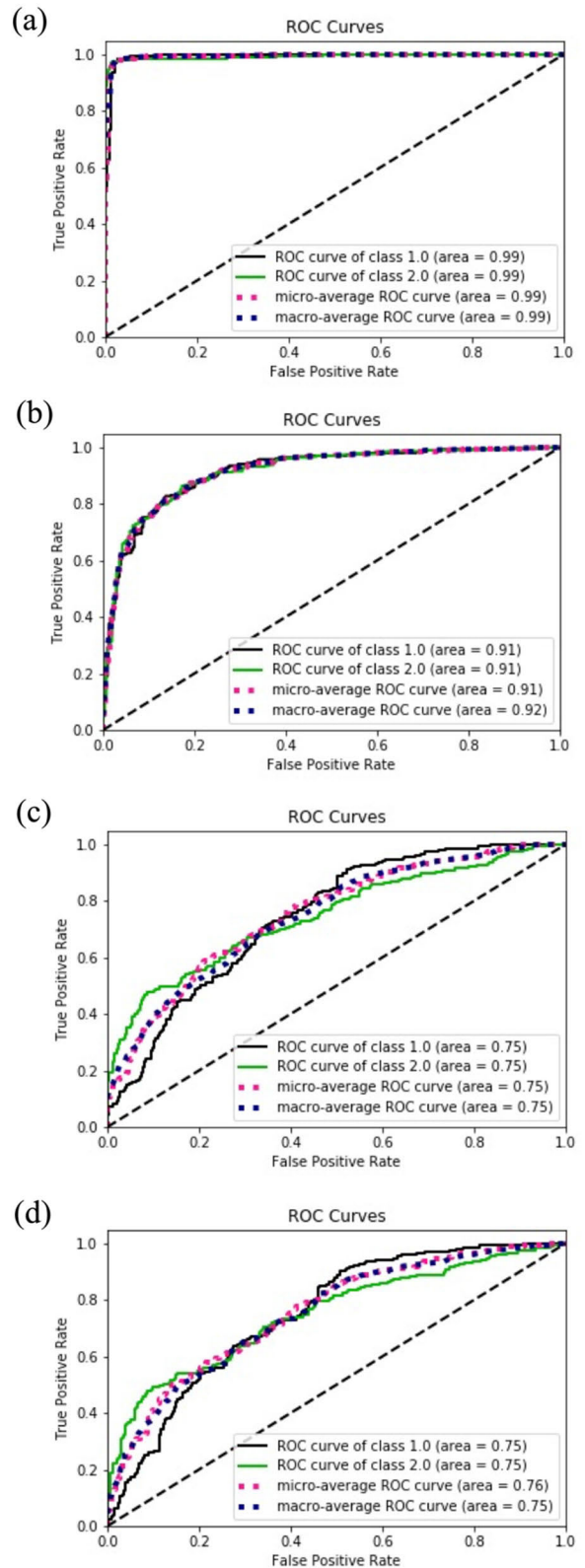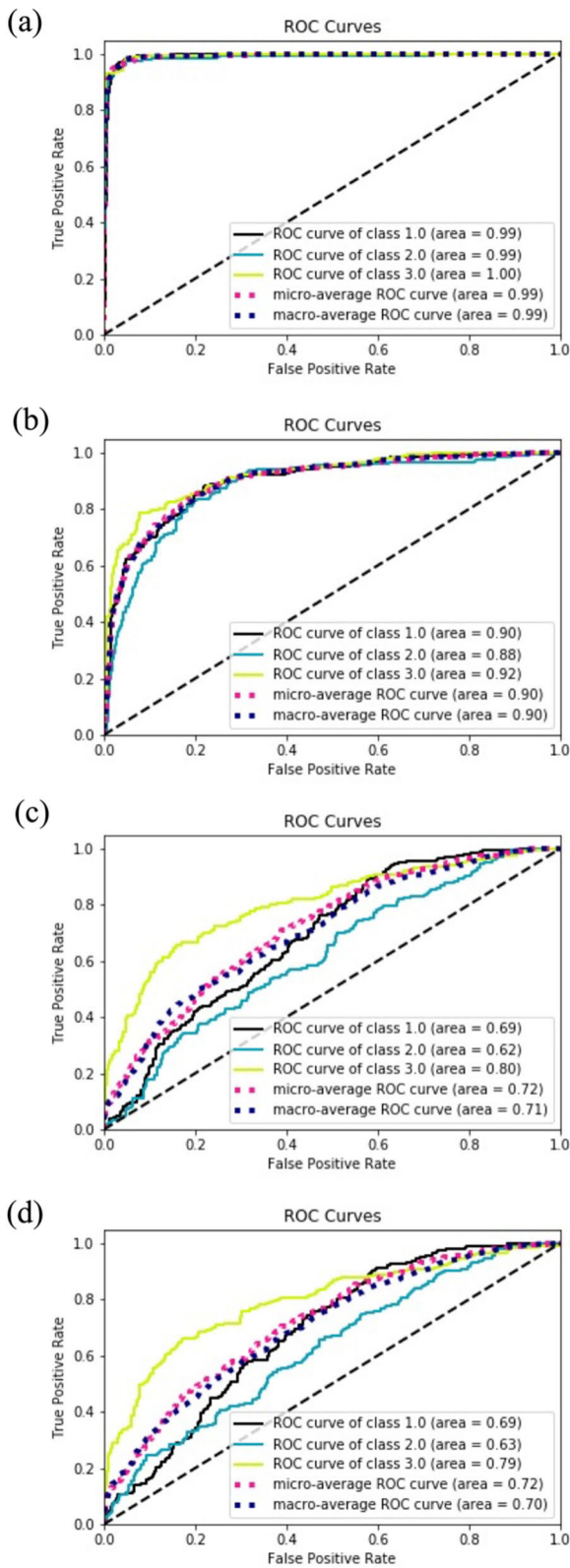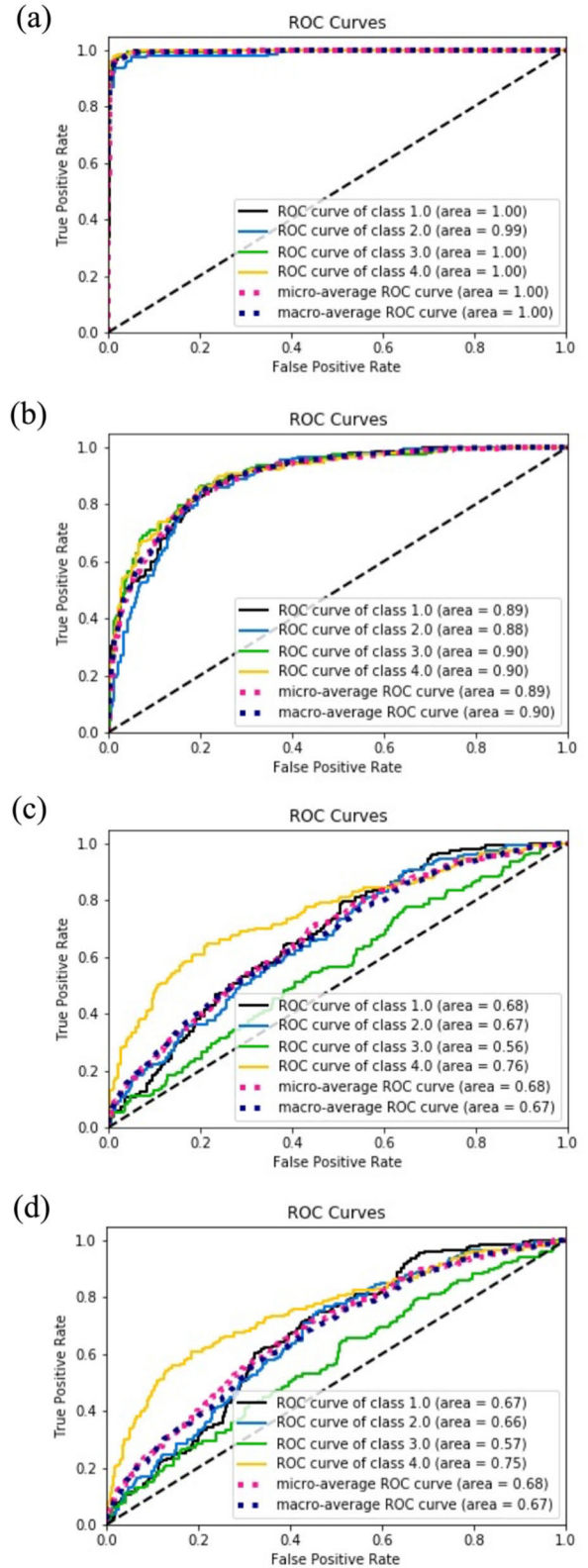
FIGURE 5. ROC curves of classifiers from actigraphy data of two days under binary condition. (a) XGBoost classifiers, (b) SVC, (c) MLP classifier, (d) LR.
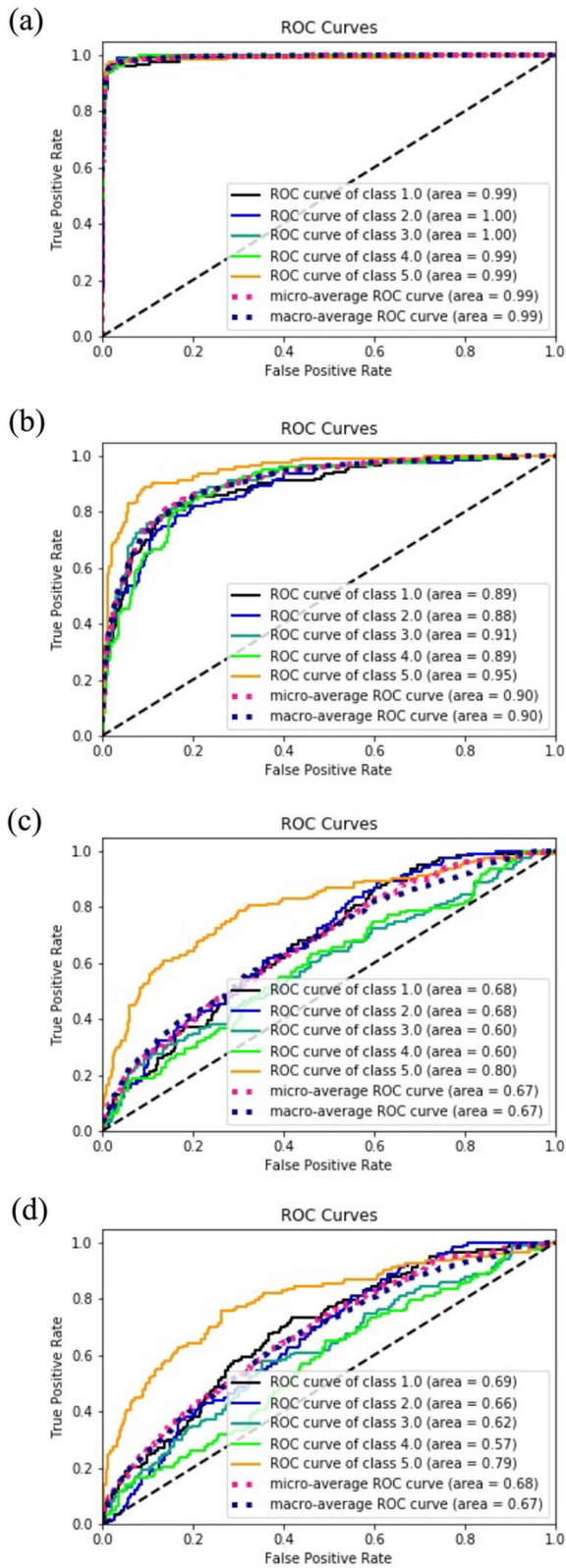
model trains a given dataset with n samples and m features $D = \{(x_i, y_i)|x_i \in R^m, y_i \in R\}$. In our cases, we used

FIGURE 6. ROC curves of classifiers from actigraphy data of two days under three-class conditions. (a) XGBoost classifiers, (b) SVC, (c) MLP classifier, (d) LR.

FIGURE 7. ROC curves of classifiers from actigraphy data of two days under four-class conditions. (a) XGBoost classifiers, (b) SVC, (c) MLP classifier, (d) LR.

(a)

TABLE 11. Comparison performance of classifiers for depression status in previous studies.

Authors	Used data	Extracted features	Classifier (class)	Performance
Hosseinifard *et al.* 64]	EEG	Four EEG bands and DFA, Higuchi fractal, correlation dimension, and Lyapunov exponent	LR (binary)	Accuracy: 90.0%
Acharya *et al.* [71]	EEG	Fractal dimension, sample entropy, Hurst's exponents, higher order spectra	SVM (binary)	Accuracy: 97.0% Recall: 97.0% Precision: 97.9%
Rubin-Falcone *et al.* [72]	sMRI	Gray matter volume	SVM (binary)	Accuracy: 75.0%
Zhong *et al.* [67]	rs-fMRI	Functional connectivity	SVM (binary)	Accuracy: 91.9%
Roh *et al.* [73]	ECG	Eight HRV features	SVM (binary)	Accuracy: 71.0%
Kim *et al.* [74]	Electrod-ermal activity	Four EDA features	Decision tree (binary)	Accuracy: 74.0% Recall: 74.0% Precision: 71.0%
Proposed Classifier	Actigra-phy	Fourteen selected features (circadian indices and survey variables)	XGBoost classifier (binary)	Accuracy: 97.88% Recall: 97.88% Precision: 97.89% F1-score: 97.88% AUC: 99.00%

LR, logistic regression; SVM, support vector machine; sMRI, structural MRI; rs-fMRI, resting-state functional MRI; HRV, heart rate variability

(b)

(c)

algorithms with regularized objectives constitute the basis for the model.

$$L(\phi) = \sum_i l(y_i, y_i') + \sum_k \Omega(f_k) \qquad (6)$$

$$\text{where} \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda \, ||\omega||^2 \qquad (7)$$

$$y_i' = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \qquad (8)$$

To optimize the algorithms with a dataset, we minimize the regularized objective function in (6), where y_i' indicates the predicted value from the tree model and each f_k corresponds to individual trees. Function l is a differentiable convex loss function that compares the difference between the predicted y_i' and target y_i. In the second term, function Ω is

(d)

FIGURE 8. ROC curves of classifiers from actigraphy data of two days under five-class conditions. (a) XGBoost classifiers, (b) SVC, (c) MLP classifier, (d) LR.

a dataset with *n* rows and *m* features (including circadian indices, survey variables, and class labels). Gradient boosting

TABLE 12. Feature ranking by importance from XGBoost classifier.

Rank	Features	Rank	Features
1	HE_BMI	9	pa_aerobic
2	MESOR	10	sex
3	minimum	11	RA
4	alpha	12	D_2_1
5	EQ5D	13	DF2_dg
6	D_1_1	14	LQ4_00
7	BP1	15	mh_stress
8	L5		

the penalization term for the complexity of models. To avoid overfitting the partial dataset, an additional regularization term smoothens the last learned weight. In this study, we set y_i as class labels to which depression levels are assigned (e.g., 'mild,' 'moderate,' 'severe' depression).

The second classification algorithm applied was the SVC with nonlinear kernels [54]. This algorithm classifies the feature space using hyperplanes that are separated by class labels. In previous studies, researchers used linear kernels to classify binary-class conditions for stress [55]. In contrast, we used a nonlinear kernel (radial basis function kernel) to evaluate the classification performance with more diverse class levels. In addition, to avoid overfitting when nonlinear kernels are used, we developed and tested the model performance using completely participant-separated datasets.

The third classification algorithm was an LR classifier. To estimate the coefficient of the regression model, a maximum likelihood estimation method was applied. Consequently, the classifier yields a likelihood value $L(x)$, where $0 \leq L(x) \leq 1$. This value indicates the association between class labels and input vectors. A likelihood value higher than 0.5, which is the assigned threshold, indicates that the condition was classified as severe depression levels in binary cases. For this classifier, we considered the basic form of the LR model with our features and depression classes as follows:

$$F(z) = E(\frac{Y}{x}) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \quad (9)$$

$$\text{where} \quad z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (10)$$

where Y represents the depression level as a class. We considered Y as a specified value of either 'mild,' 'moderate,' or 'severe' in the three classes. In summary, the LR model suggested probability values to categorize each class under various conditions.

The final classification algorithm used in this study was an MLP classifier (i.e., an artificial neural network model). It consists of multiple layers of, at least, three layers of nodes (input, hidden, and output layers). Each node calculates the output vectors through the activation function g with weight

and bias vectors. The detailed calculation is as follows:

$$o_k = g(w_k^{(L)\top} h^{(L)} + b_k^{(L)}) \quad (11)$$

$$\text{where} \quad h_k^{(l)} = f^{(l)}(\sum_{j=1}^{m_{l-1}} w_{kj}^{(l-1)} h_j^{(l-1)} + b_k^{(l-1)}) \quad (12)$$

E. EVALUATION METRICS

We compared the classification performance of the ML classifiers based on six evaluation metrics. To evaluate the classification results of the algorithms using other indicators rather than only the accuracy, we calculated the true positive (TP), true negative (TN), false negative (FN), and false positive (FP) values from the confusion matrix. The correctly classified samples were calculated using the TP and TN values. In contrast, incorrectly classified samples were indicated by FN and FP. Based on the four basic values from the confusion matrix, we obtained four additional indicators: precision, recall, F1-score, and accuracy, calculated using (13–16), respectively. Furthermore, we confirmed the true positive rate (TPR) and false positive rate (FPR), using (17) and (18), respectively, to draw the ROC curve. In addition, we evaluated the performance based on AUC values using an ROC curve.

To validate the classification performances of each classifier, we applied a one-way analysis of variance (ANOVA) test considering the evaluation indices values from classification algorithms.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (13)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (14)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (15)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (16)$$

$$TruePositiveRate = \frac{TP}{TP + FN} \quad (17)$$

$$FalsePositiveRate = \frac{FP}{FP + TN} \quad (18)$$

F. TOOLS

All codes for ML classifiers and data preprocessing were written using Python (version 3.7.1; scikit-learn, version 2.4.1) and R (version 4.0.3).

III. RESULTS

The performance of ML algorithms at classifying depression levels is shown in Tables 6–9. Specifically, we examined the classification performance and optimal length of the actigraphy data extracted for circadian rhythm indices. First, in terms of the classification performance, the XGBoost classifier outperformed the other algorithms based on all the evaluation metrics. In addition, to identify the classification performance in terms of various label conditions, we compared the values of evaluation metrics under four conditions (binary, three, four, and five classes). The evaluation
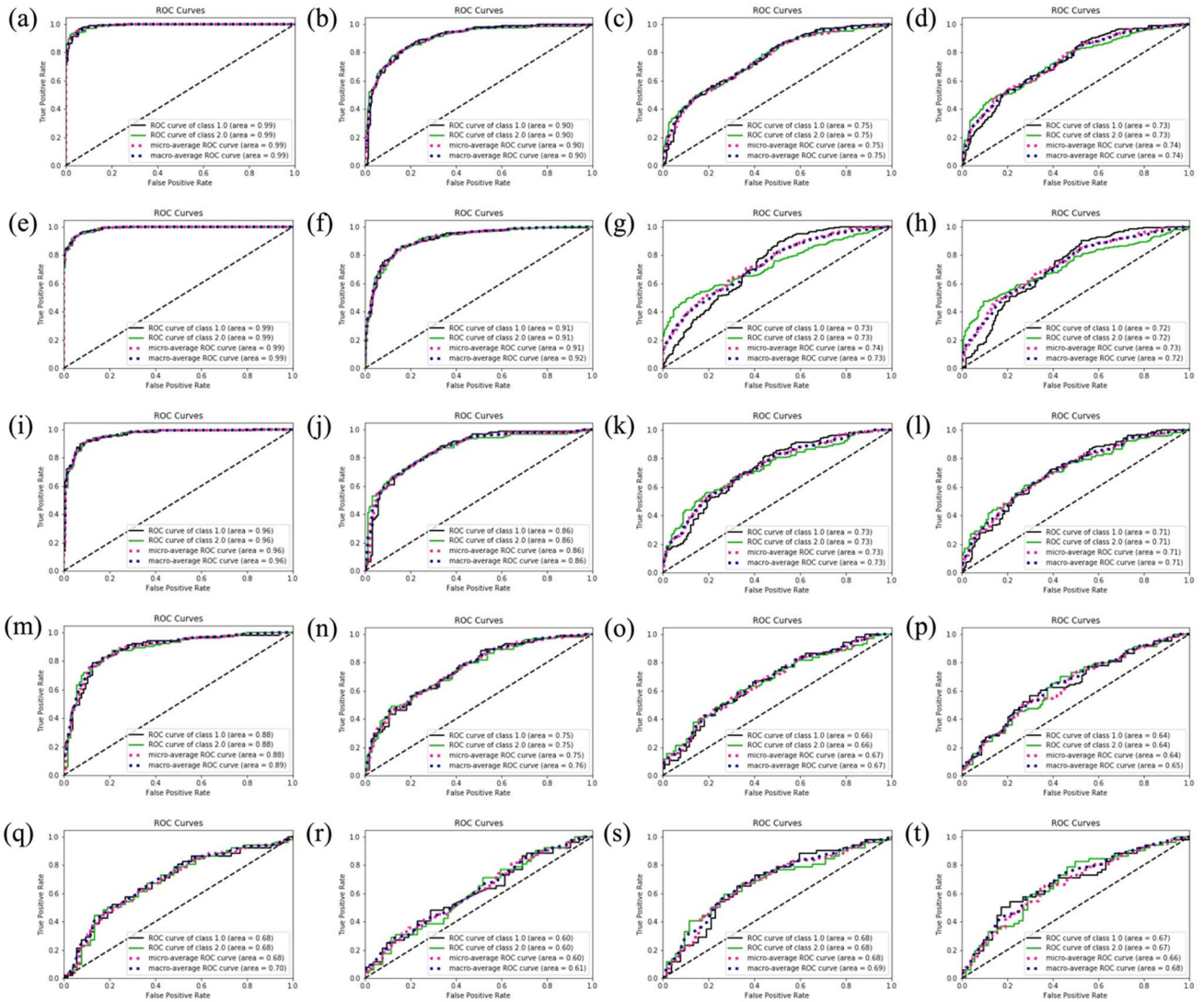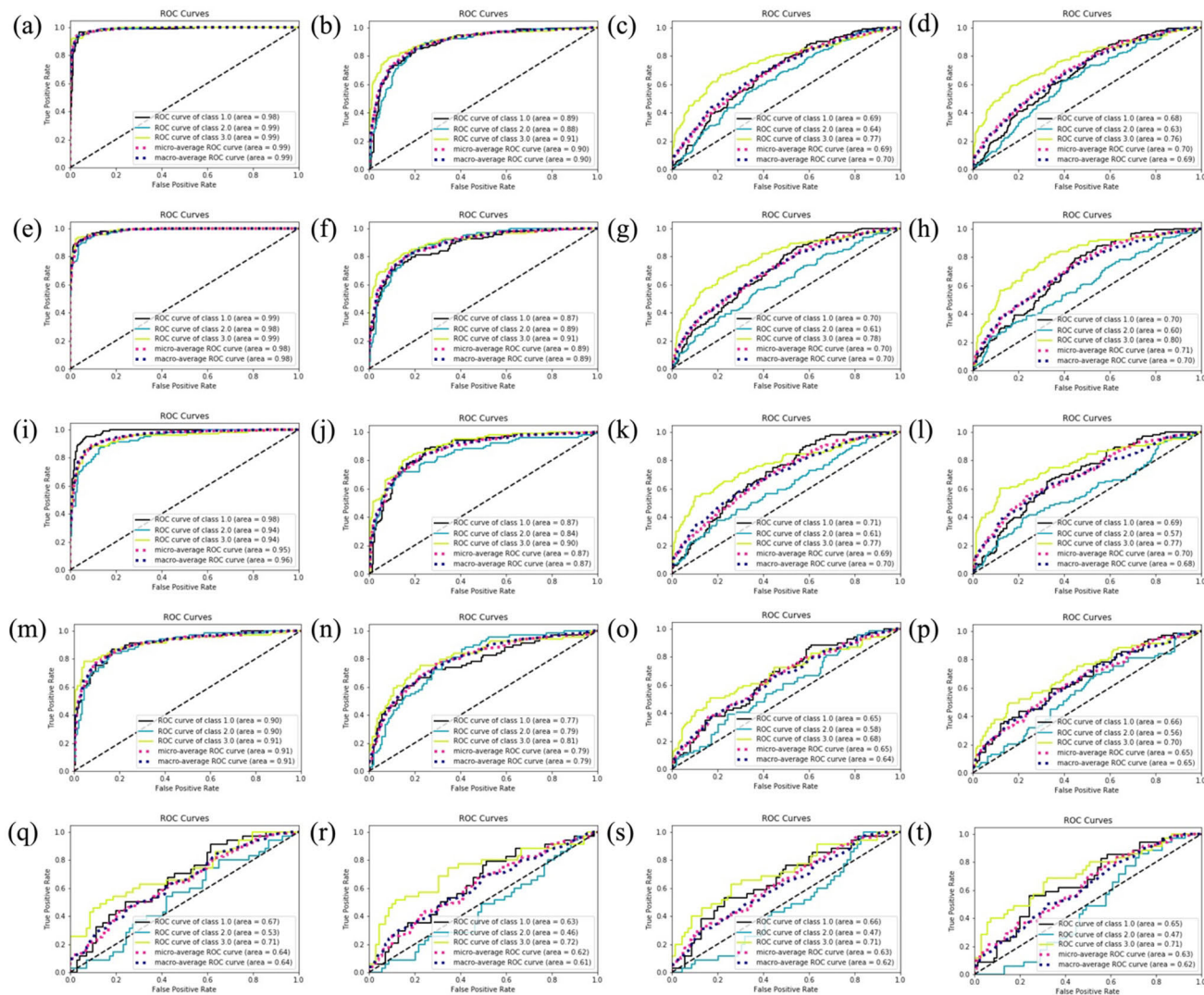
FIGURE 9. ROC curves of classifiers under binary-class conditions. (a), (b), (c), (d) Three-day length actigraphy, (e), (f), (g), (h) four-day length actigraphy, (i), (j), (k), (l) five-day length actigraphy, (m), (n), (o), (p) six-day length actigraphy, (q), (r), (s), (t) seven-day length actigraphy, (a), (e), (i), (m), (q) XGBoost classifier, (b), (f), (j), (n), (r) SVC, (c), (g), (k), (o), (s) MLP, (d), (h), (l), (p), (t) LR.

metric values of the XGBoost classifier were 97.42% (accuracy), 97.51% (precision), 97.19% (recall), 97.40% (F1-score), and 99.00% (AUC) in the binary-class condition. In the three-class condition, accuracy (95.21%), precision (95.01%), recall (95.21%), F1-score (95.87%), and AUC (99.00%) were obtained. In the four-class condition, accuracy (95.92%), precision (95.02%), recall (95.46%), F1-score (95.82%), and AUC (99.00%) were found. In the five-class condition, as the last condition, accuracy (94.81%), precision (94.21%), recall (94.90%), F1-score (94.94%), and AUC (99.00%) were examined using the XGBoost classifier. The ROC curves for each class condition with two days of actigraphy data are shown in Fig. 5–8. The remaining ROC curves under other conditions are listed in Appendix A.

Second, the performance of each classification algorithm was compared based on the length of the actigraphy data from

which the circadian indices were extracted. The maximum evaluation metric values of the classifiers were obtained for the actigraphy data for two days. Additionally, we investigated whether the number of rows differed for each of the datasets obtained for the five durations of actigraphy monitoring. The number of rows gradually decreased as the length of the actigraphy data increased. To prevent the dataset size from affecting the performance, we controlled the size of the dataset and conducted an additional experiment. In this experiment, 1000 rows were sampled through stratified random sampling for all the datasets to reduce the bias for each class label. Subsequently, the same experimental process was applied to each dataset. We confirmed the same tendency in additional experiments and concluded that circadian indices extracted from two-day actigraphy data were sufficient to classify depression levels. The detailed results are presented in Appendix B.

FIGURE 10. ROC curves of classifiers under three-class conditions. (a), (b), (c), (d) Three-day length actigraphy, (e), (f), (g), (h) four-day length actigraphy, (i), (j), (k), (l) five-day length actigraphy, (m), (n), (o), (p) six-day length actigraphy, (q), (r), (s), (t) seven-day length actigraphy, (a), (e), (i), (m), (q) XGBoost classifier, (b), (f), (j), (n), (r) SVC, (c), (g), (k), (o), (s) MLP, (d), (h), (l), (p), (t) LR.

Finally, we validated the classification performance using a one-way ANOVA test. The null hypothesis established was that the average performance of the four algorithms was the same. We verified that the test results of evaluation indices (accuracy, precision, recall, F1-score, AUC) were statistically significant. As a result, statistical significance of performance was confirmed, and the null hypothesis was rejected. Detailed one-way ANOVA test results are shown in Table 10.

IV. DISCUSSION

In this study, we attempted to classify depression levels using actigraphy data based on ML algorithms. Survey variables and circadian rhythm indices extracted from actigraphy data were collected from the KNHANES dataset. To obtain reasonable evidence for depression status identification with physical activity, we found several studies

related to clinical and technical aspects. First, considering the relationship between depression and physical activity, Wu *et al.* [56] established that physical inactivity in patients with Parkinson's disease caused depression and degeneration of motor skills through a comprehensive review of relevant studies. Teixeira *et al.* [57] proved that physical activity was associated with depression and anxiety in elderly groups. Moreover, Roshanaei-Moghaddam *et al.* [58] verified that decreased levels of physical exercise or sedentary lifestyle were a significant risk factor of depression. Ku *et al.* [59] tracked elderly groups for 11-year periods. They identified that physical activity engagement was associated with lower risk of depressive symptoms. Based on these previous studies, we determined that physical activity including aerobic exercises can work as a main factor to depression.
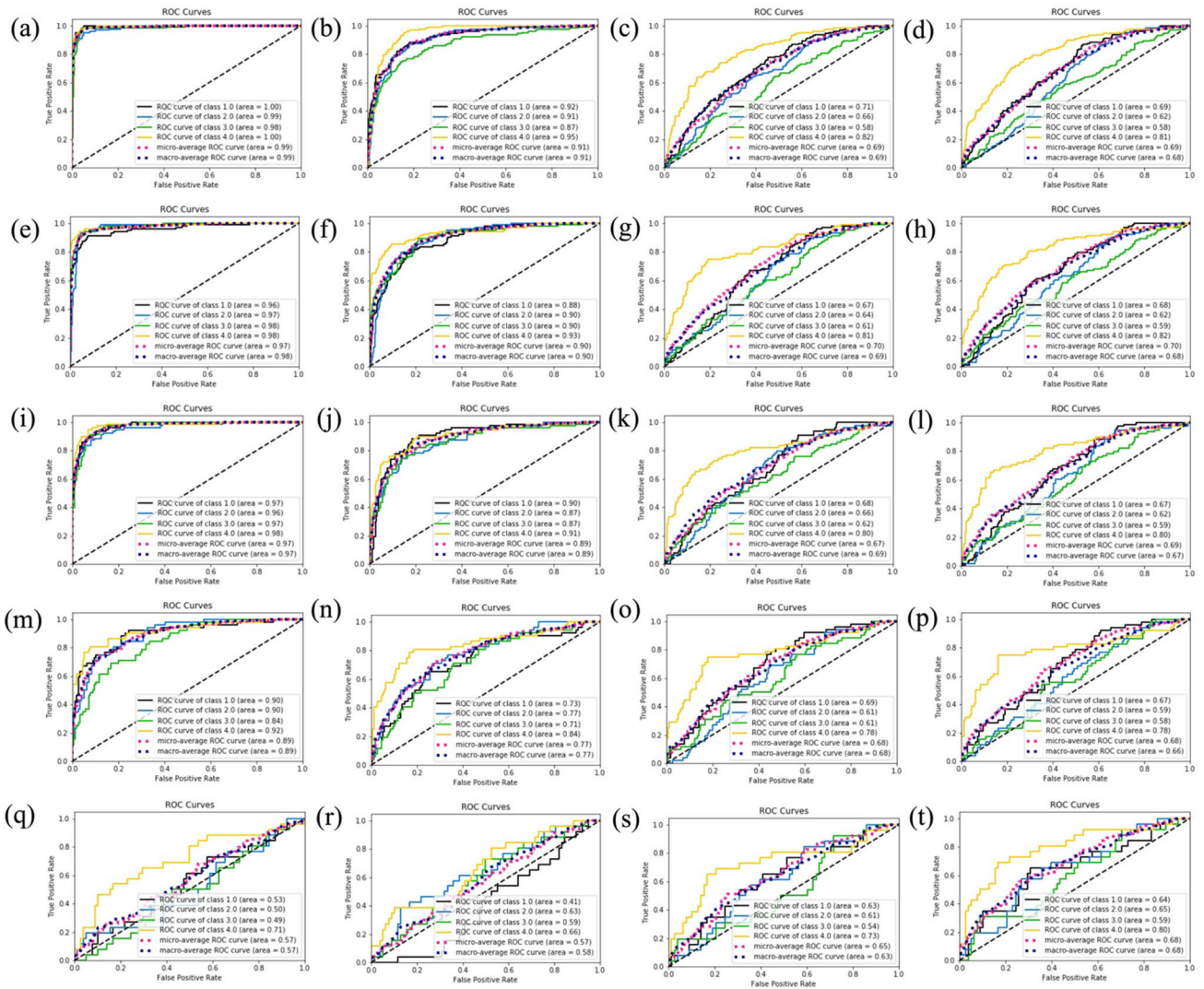
FIGURE 11. ROC curves of classifiers under four-class conditions. (a), (b), (c), (d) Three-day length actigraphy, (e), (f), (g), (h) four-day length actigraphy, (i), (j), (k), (l) five-day length actigraphy, (m), (n), (o), (p) six-day length actigraphy, (q), (r), (s), (t) seven-day length actigraphy, (a), (e), (i), (m), (q) XGBoost classifier, (b), (f), (j), (n), (r) SVC, (c), (g), (k), (o), (s) MLP, (d), (h), (l), (p), (t) LR.

Second, related to analyses with ML algorithms, Albahli *et al.* [60] suggested a thoracic disease identification framework through deep neural network models. Albahli *et al.* [61] showed that the detection performance of a convolutional neural network in X-ray images was superior to that of other models. In addition, Chekroud *et al.* [62] built ML models to find predictive factors for determining the responsiveness to antidepressant treatment in patients with depression. Furthermore, Bhakta and Arkaprabha [63] compared five ML algorithms to predict depression in the elderly population. Based on these studies, we concluded that ML algorithms can be used to detect or identify diseases. Therefore, our topic about classification of the level of depression using ML algorithms was well-founded.

To reflect variations in specific factors, time-series data collected from study participants directly (EEG or ECG recorded by electrodes attached to the skin) or indirectly (actigraphy data measured using wearable devices) were utilized with structured data. For example, Hosseinifard *et al.* [64] used electrical activities of the brain to evaluate depression. EEG signals of depression patients were utilized to extract feature vectors. Both linear features (e.g., power values of four EEG bands from power analysis) and nonlinear features (e.g., detrended fluctuation analysis (DFA), Higuchi fraction, and Lyapunov exponent) were applied to ML classifiers. Three classifiers, linear discriminate analysis, LR, and k-nearest neighbors, were compared to identify depression patients in a study of patient and control groups. LR classifiers yielded a classification accuracy of 83.3% with a correlation dimension. In addition, the LR classifiers showed 90% accuracy with all nonlinear features. The authors indicated that the model performance

TABLE 13. Classification performance results for classifiers with binary-class conditions.

Actigraphy condition	2 days				3 days				4 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	91.55	66.39	72.72	67.20	90.11	65.89	71.51	66.47	86.47	65.45	71.25	66.18
Precision	91.56	67.26	73.00	67.56	90.12	67.00	72.31	66.94	86.50	66.29	72.12	66.87
Recall	91.55	66.39	72.72	67.20	90.11	65.89	71.51	66.47	86.47	65.45	71.25	66.16
F1-score	91.55	65.96	72.64	67.04	90.11	65.32	71.25	66.23	86.47	65.00	70.91	65.83
AUC	98.00	74.00	79.00	75.00	97.00	71.00	76.00	73.00	95.00	74.00	80.00	71.00
Actigraphy condition	**5 days**				**6 days**				**7 days**			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	84.56	64.30	72.02	66.88	75.96	59.13	59.61	58.65	64.42	62.50	68.26	61.53
Precision	84.58	65.59	74.67	67.48	76.00	60.02	59.79	58.86	64.55	63.33	68.33	61.98
Recall	84.56	64.26	71.97	66.85	75.96	59.13	59.61	58.65	64.42	62.50	68.26	61.53
F1-score	84.56	63.51	71.22	66.57	75.95	58.20	59.43	58.40	64.34	61.90	68.24	61.18
AUC	93.00	71.00	77.00	71.00	81.00	68.00	72.00	63.00	67.00	68.00	72.00	68.00

AUC, area under the curve; SVC, support vector classifier; MLP, multilayer perceptron; LR, logistic regression.

TABLE 14. Classification performance results for classifiers with three-class conditions.

Actigraphy condition	2 days				3days				4 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	91.87	49.43	52.03	48.29	90.31	49.61	52.90	50.38	82.36	46.61	53.38	47.82
Precision	92.01	49.26	52.56	46.93	90.64	49.22	53.11	49.98	82.42	45.83	54.09	46.53
Recall	91.87	49.43	52.03	48.29	90.31	49.61	52.90	50.38	82.36	46.61	53.38	47.82
F1-score	91.88	48.73	52.25	46.79	90.39	49.23	52.92	49.32	82.38	45.56	53.00	46.83
AUC	98.00	71.00	72.00	71.00	98.00	70.00	73.00	70.00	95.00	68.00	72.00	70.00
Actigraphy condition	**5 days**				**6 days**				**7 days**			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	76.20	52.41	52.73	50.80	65.21	45.41	46.86	44.92	49.03	45.19	48.07	45.19
Precision	76.35	52.16	52.90	49.91	66.07	45.40	47.40	45.07	49.25	44.00	46.53	42.54
Recall	76.21	52.46	52.77	50.84	65.21	45.41	46.86	44.92	49.02	45.35	48.17	45.26
F1-score	76.20	50.89	52.14	49.38	65.26	45.03	46.82	44.77	49.07	42.94	46.39	42.97
AUC	91.00	70.00	74.00	70.00	81.00	64.00	67.00	64.00	67.00	67.00	66.00	66.00

AUC, area under the curve; SVC, support vector classifier; MLP, multilayer perceptron; LR, logistic regression.

was significantly better when a combination of linear and nonlinear features was used, compared to the case when only linear features were used.

Mohammadi *et al.* [65] proposed a fuzzy function-based ML classifier trained by three nonlinear features (fuzzy entropy, Katz fractal dimension, and fuzzy fractal dimension) to distinguish depression levels. To reflect variation of brain activities, the researchers collected EEG signals from depression patient groups, based on which all the features were calculated. To evaluate the classification performance in combination with each feature, three nonlinear features were randomly combined into groups with one, two, and three feature groups. The proposed algorithms (fuzzy function-based algorithms) were compared with SVM classifiers with 90.0% accuracy. Among the classifiers, those trained using all features (three features) showed the best performance under all conditions.

To classify the stress status of study participants, Rizwan *et al.* [66] proposed classification algorithms using SVM with features extracted from ECG signals. Three features (QT interval, RR interval, and ECG-derived respiration) were applied to the classifiers. In addition, two

TABLE 15. Classification performance results for classifiers with four-class conditions.

Actigraphy condition	2 days				3 days				4 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	91.23	40.09	44.48	38.14	88.17	42.83	44.96	39.14	79.95	39.13	42.99	38.40
Precision	91.40	40.28	45.42	36.85	88.22	40.47	44.57	37.12	80.04	37.72	41.95	35.22
Recall	91.23	40.09	44.48	38.14	88.17	42.83	44.96	39.14	79.95	39.19	43.05	38.50
F1-score	91.27	37.69	43.93	36.88	88.13	39.76	43.57	37.37	79.91	37.12	42.31	35.83
AUC	98.00	66.00	72.00	67.00	98.00	70.00	74.00	68.00	95.00	69.00	72.00	69.00
Actigraphy condition	5 days				6 days				7 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	74.92	42.12	45.01	42.21	66.34	37.01	38.94	36.05	40.38	39.42	43.26	44.23
Precision	75.07	40.68	44.09	39.60	66.30	33.21	38.32	33.11	39.77	38.81	43.35	45.39
Recall	74.91	42.12	45.07	42.14	66.34	37.01	38.94	36.05	40.38	39.42	43.26	44.23
F1-score	74.96	40.45	43.04	39.99	66.28	32.77	37.88	33.77	39.71	34.63	41.75	42.23
AUC	93.00	70.00	73.00	70.00	86.00	66.00	71.00	67.00	63.00	67.00	70.00	70.00

AUC, area under the curve; SVC, support vector classifier; MLP, multilayer perceptron; LR, logistic regression.

TABLE 16. Classification performance results for classifiers with five-class conditions.

Actigraphy condition	2 days				3 days				4 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	88.94	32.35	36.42	31.53	84.49	36.04	34.88	31.97	77.53	37.68	44.92	35.50
Precision	89.27	29.06	39.19	29.57	84.85	34.82	34.10	28.56	77.81	36.71	44.73	34.04
Recall	88.94	32.35	36.42	31.54	84.50	36.09	34.93	32.03	77.52	37.70	44.94	35.53
F1-score	88.98	28.03	36.83	29.26	84.57	32.54	33.56	29.13	77.59	34.80	43.43	34.01
AUC	98.00	66.00	70.00	67.00	98.00	68.00	70.00	66.00	95.00	69.00	75.00	69.00
Actigraphy condition	5 days				6 days				7 days			
Classifier	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR	XGBoost	SVC	MLP	LR
Accuracy	69.45	32.79	39.87	32.15	58.93	31.88	42.99	36.71	20.19	30.76	32.69	28.84
Precision	69.77	30.37	40.49	30.22	59.13	31.79	42.88	36.89	20.75	32.60	33.44	26.59
Recall	69.45	32.83	39.88	32.22	59.02	32.03	43.11	36.89	20.23	30.76	32.81	28.95
F1-score	69.51	29.56	39.18	29.85	59.06	30.17	42.46	35.19	20.21	28.17	32.16	26.97
AUC	93.00	67.00	73.00	66.00	85.00	68.00	74.00	71.00	56.00	65.00	63.00	63.00

AUC, area under the curve; SVC, support vector classifier; MLP, multilayer perceptron; LR, logistic regression.

kernels (Gaussian and cubic) and three model types (linear, quadratic, and cubic) in SVM algorithms were compared to find high-performance algorithms for stress status. The classification algorithms yielded their best performance when all the three features were used, compared to the cases in which only one or two were used. In addition, models with Gaussian kernels exhibited promising accuracy (linear SVM: 98.6%, quadratic SVM: 98.6%, and cubic SVM: 98.6%) compared to that of cubic kernel SVM models (linear SVM: 97.2%, quadratic SVM: 97.1%, and cubic SVM: 97.2%).

Zhong *et al.* [67] used whole-brain resting-state functional MRI data (rs-fMRI) from both depression and healthy groups to identify major depressive disorders. From collected rs-fMRI data, brain activity time series data of 116 brain regions were extracted to construct a functional connectivity network. Functional connectivity represented by Pearson correlation matrix and correlation coefficient vectors in several matrices were applied to SVM classifiers as input features. To select high discriminate features for classification, the Kendall Tau rank method was applied. SVM classifiers with linear kernel function showed the best classification accuracy (91.9%) in
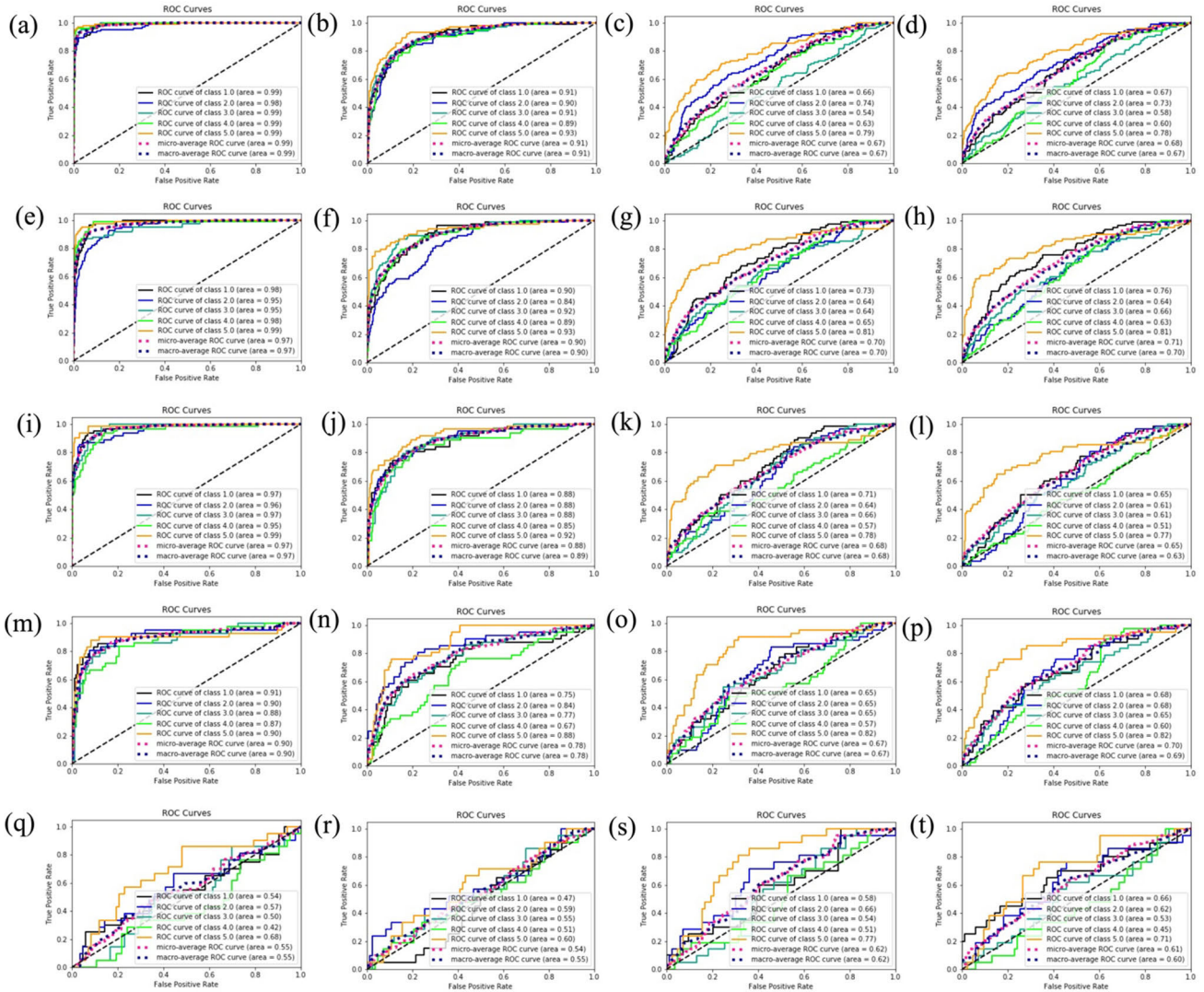
FIGURE 12. ROC curves of classifiers under five-class conditions. (a), (b), (c), (d) Three-day length actigraphy, (e), (f), (g), (h) four-day length actigraphy, (i), (j), (k), (l) five-day length actigraphy, (m), (n), (o), (p) six-day length actigraphy, (q), (r), (s), (t) seven-day length actigraphy, (a), (e), (i), (m), (q) XGBoost classifier, (b), (f), (j), (n), (r) SVC, (c), (g), (k), (o), (s) MLP, (d), (h), (l), (p), (t) LR.

experimental conditions. In addition, only six features were confirmed as efficient from a total 116 features.

To enable comparison with previous studies, we constructed an experimental design of our research composed of similar steps (feature extraction from time series data, feature selection, and classification through ML algorithms). Different from time series data widely used in previous works (EEG, ECG, and time series from rs-fMRI), we attempted to use the variation of physical activity to investigate a possible relationship with depression levels.

To reflect the variation in physical activity, diverse methods can be used to collect physical activity data from participants. Physical activity data obtained by self-report questionnaires are widely utilized to measure the averaged physical activity. De Mello *et al.* [68] used a self-report physi-

cal activity questionnaire to assess the physical activity of depression patients. They surveyed various types of physical activity (e.g., weak, moderate, or vigorous) and regularity of activity (e.g., the number of activities in a week). Additionally, to monitor a subject's physical activity over a long period, Sabia *et al.* [69] collected physical activity questionnaires from elderly groups with dementia for 28 years. Furthermore, detailed physical activity patterns can be collected as time-series data by accelerometer or pedometer.

Harris *et al.* [70] compared self-report and time-series physical activity data to validate each metric. The researchers suggested that each data has an advantage in accordance with research methods and topics. Because of densely continuous activity values (acceleration or step values), time-series data collected by wearable devices can offer detailed intensity of
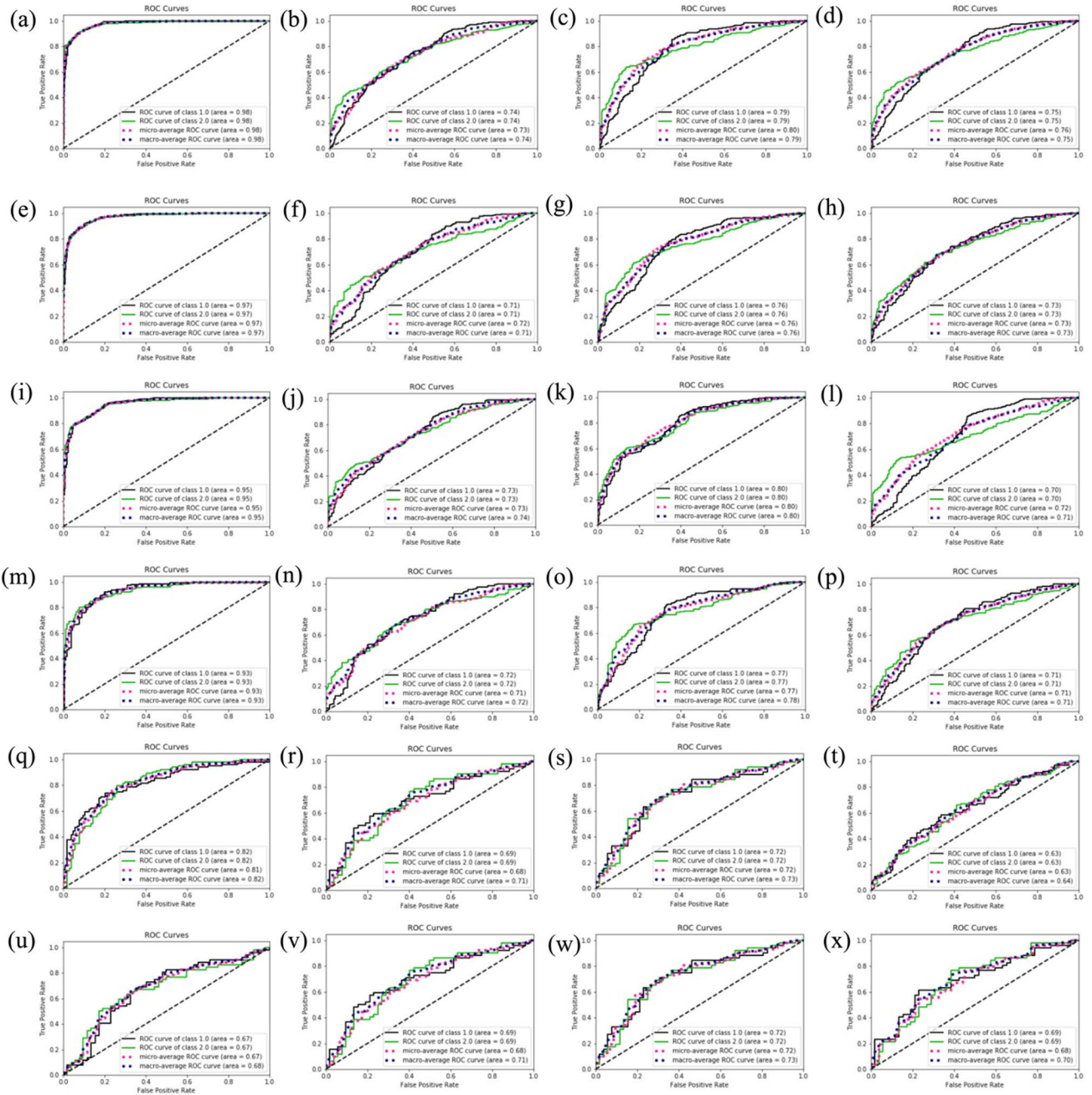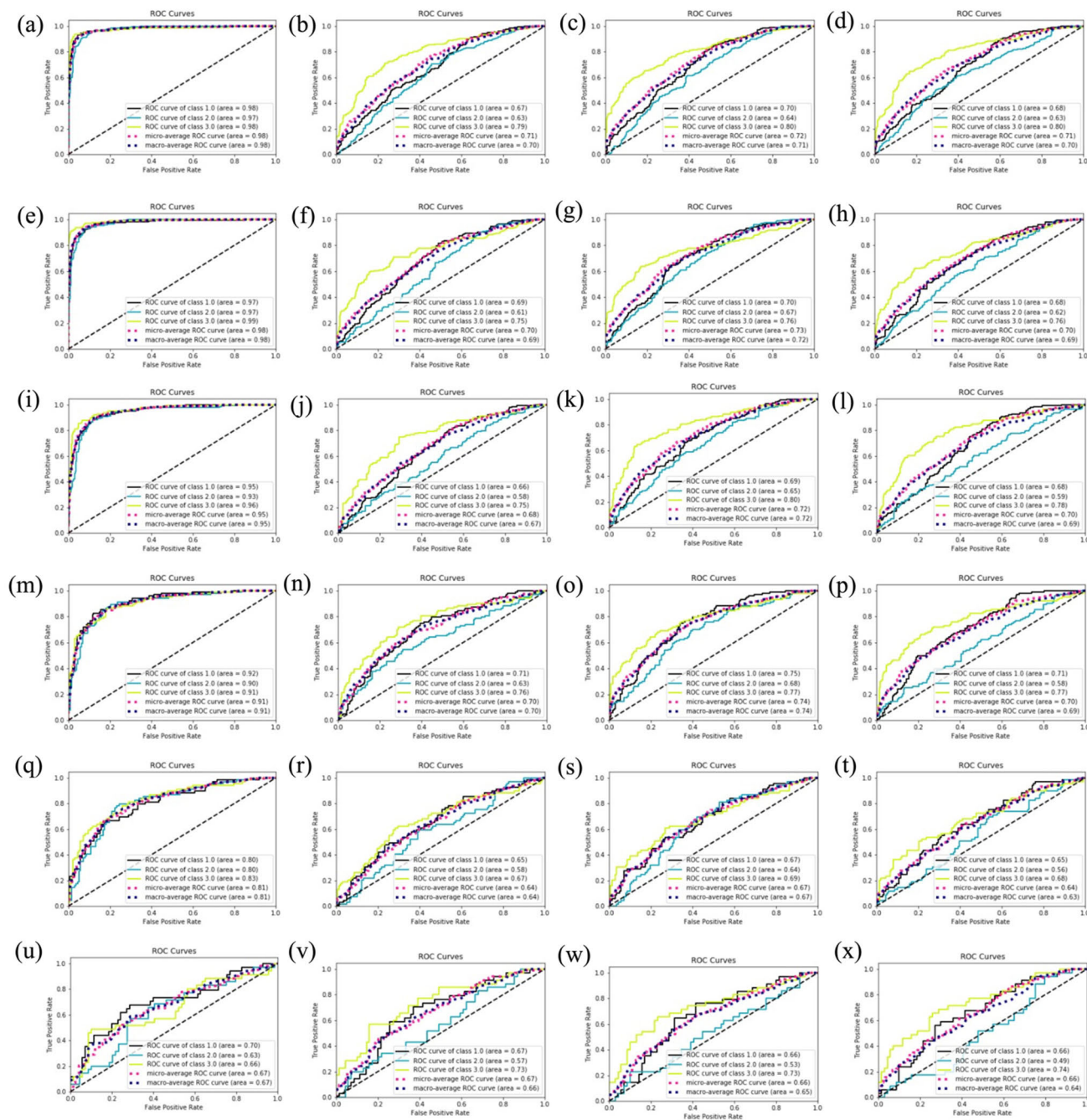
FIGURE 13. ROC curves of classifiers under binary-class conditions. (a), (b), (c), (d) Two-day length actigraphy, (e), (f), (g), (h) three-day length actigraphy, (i), (j), (k), (l) four-day length actigraphy, (m), (n), (o), (p) five-day length actigraphy, (q), (r), (s), (t) six-day length actigraphy, (u), (v), (w), (x) seven-day length actigraphy, (a), (e), (i), (m), (q), (u) XGBoost classifier, (b), (f), (j), (n), (r), (v) SVC, (c), (g), (k), (o), (s), (w) MLP, (d), (h), (l), (p), (t), (x) LR.

activity on an hourly or daily basis. In the case of self-report questionnaire data, the authors suggested that self-report is more convenient for long-term follow-up studies and more useful for evaluating activity type or in combination with other structured datasets.

In our study, we attempted to confirm the relationship between physical activity and depression status. Furthermore, characteristics of the circadian cycle embedded in physical activity were focused to identify an inherent relation

with depression. To extract circadian indices from physical activity, actigraphy data, which are time-series data, were used. Moreover, continuous activity values composed of actigraphy data were more favorable for establishing detailed patterns. Therefore, we used accelerometer-based actigraphy data to calculate circadian rhythm indices. Furthermore, we determined which features were more effective for classification among the features extracted from actigraphy data.
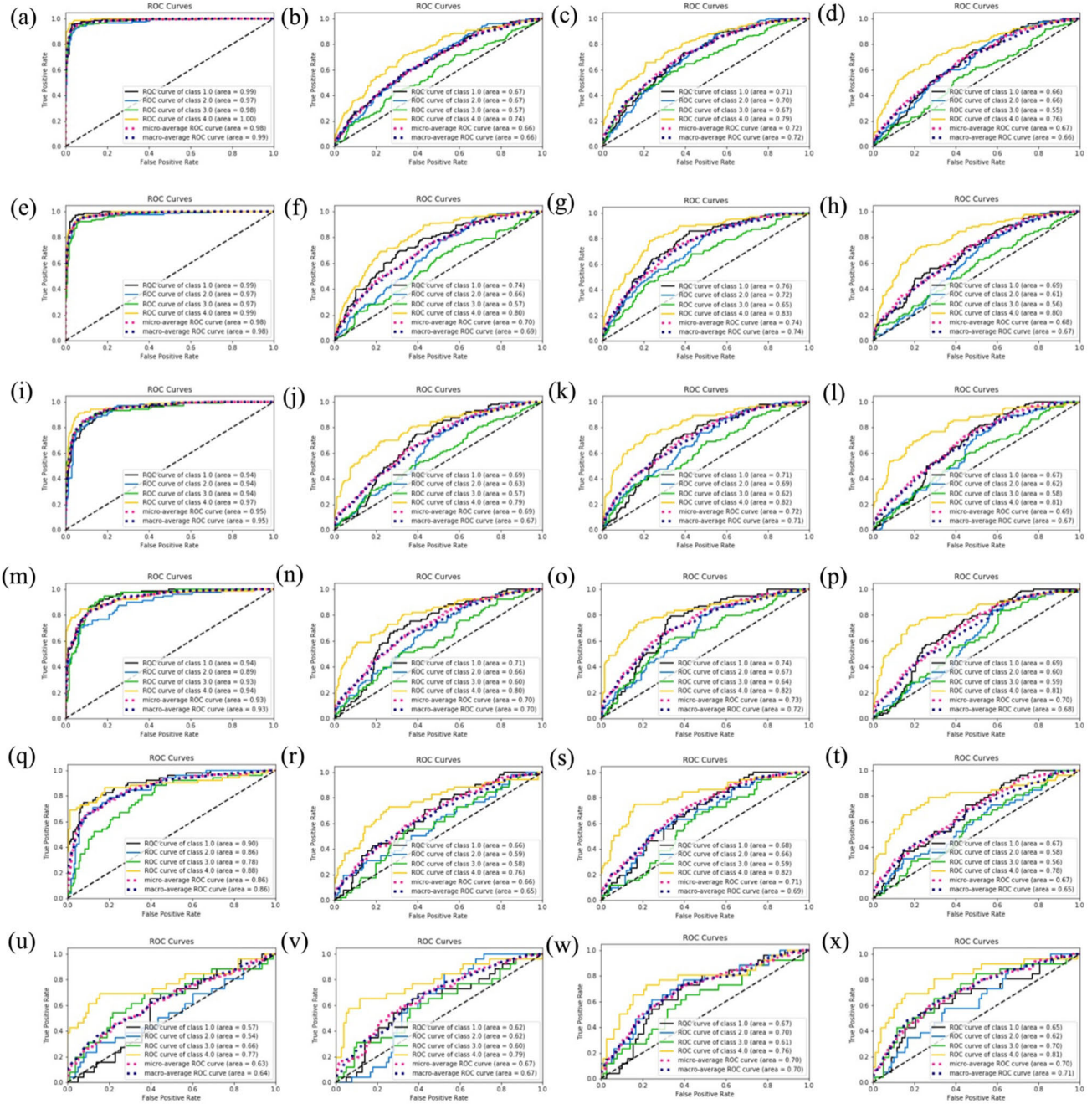
FIGURE 14. ROC curves of classifiers under three-class conditions. (a), (b), (c), (d) Two-day length actigraphy, (e), (f), (g), (h) three-day length actigraphy, (i), (j), (k), (l) four-day length actigraphy, (m), (n), (o), (p) five-day length actigraphy, (q), (r), (s), (t) six-day length actigraphy, (u), (v), (w), (x) seven-day length actigraphy, (a), (e), (i), (m), (q), (u) XGBoost classifier, (b), (f), (j), (n), (r), (v) SVC, (c), (g), (k), (o), (s), (w) MLP, (d), (h), (l), (p), (t), (x) LR.

After classification by ML classifiers, we evaluated the results based on both classification performance and feature importance. In terms of classification performance, we compared performance under various conditions (24 conditions). To confirm changes in performance with class conditions, we set four conditions (binary-, three-, four-, and five-class labels). Among the four classifiers (XGBoost, SVC, MLP classifier, and LR), the XGBoost classifier showed the best

performance in all experimental conditions. Furthermore, we compared the performance of our framework with that of the classifiers proposed in previous studies. The performance of each classifier is listed in Table 11. Despite using different data, we confirmed that the XGBoost classifier proposed in our study showed excellent performance compared to that of classifiers developed in previous studies. Moreover, based on these results, we found that circadian characteristics of
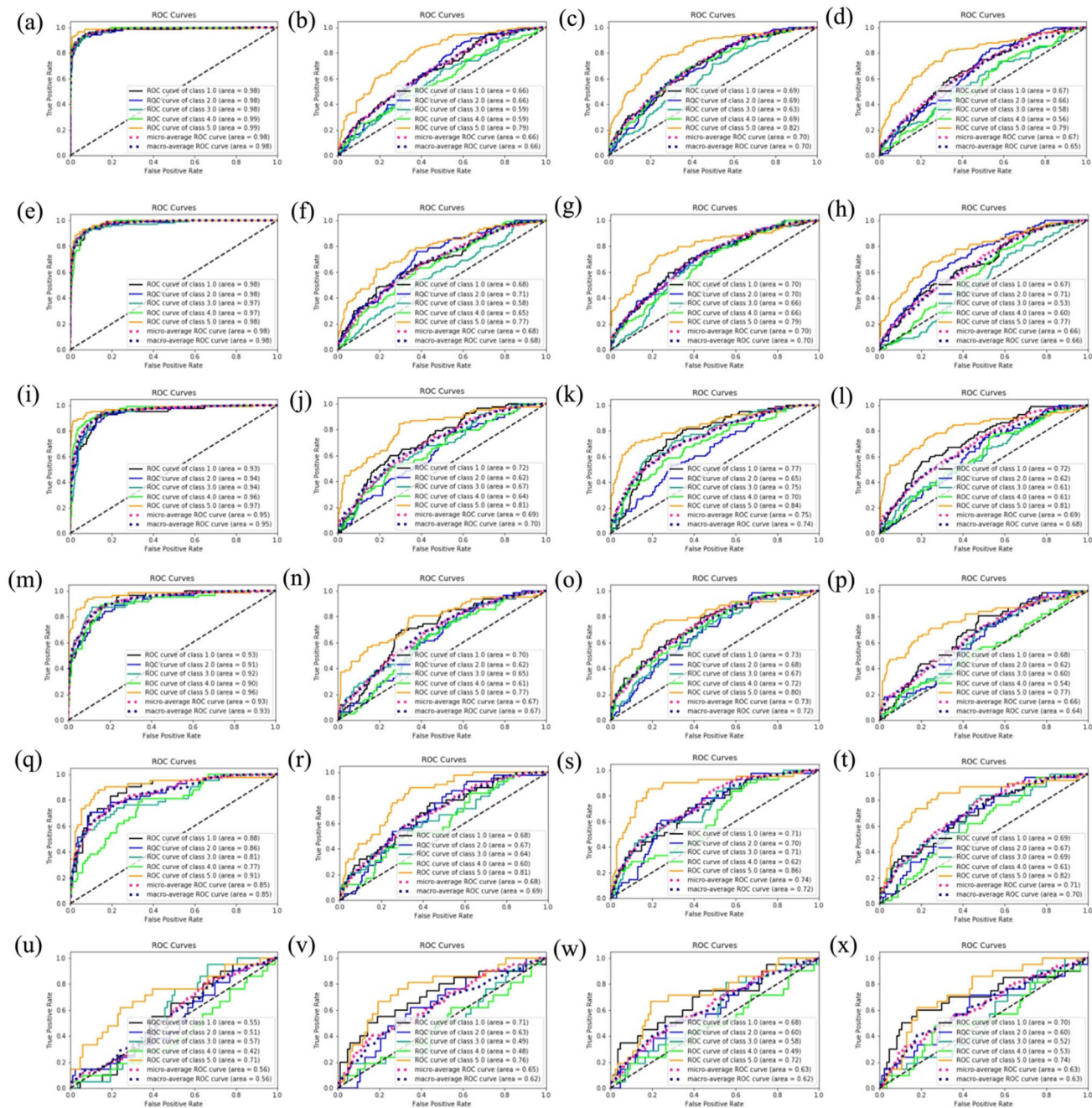
FIGURE 15. ROC curves of classifiers under four-class conditions. (a), (b), (c), (d) Two-day length actigraphy, (e), (f), (g), (h) three-day length actigraphy, (i), (j), (k), (l) four-day length actigraphy, (m), (n), (o), (p) five-day length actigraphy, (q), (r), (s), (t) six-day length actigraphy, (u), (v), (w), (x) seven-day length actigraphy, (a), (e), (i), (m), (q), (u) XGBoost classifier, (b), (f), (j), (n), (r), (v) SVC, (c), (g), (k), (o), (s), (w) MLP, (d), (h), (l), (p), (t), (x) LR.

physical activity not widely used were valuable to classify depression levels.

Among several factors influencing algorithm performance, the length of actigraphy for extracting features was critical. We expected that classification performance would be affected by the length of the actigraphy data from which features were extracted. Furthermore, the length is one of the hyperparameters that the researcher must determine. To

determine the optimal length of actigraphy, six conditions (two, three, four, five, six, and seven days) were evaluated. As a result, all the evaluation metric values showed the highest values when actigraphy monitoring was conducted for two days.

However, the dataset consisted of extracted features of different sizes. Because different dataset sizes can affect the evaluation results, we constructed a dataset with

FIGURE 16. ROC curves of classifiers under five-class conditions. (a), (b), (c), (d) Two-day length actigraphy, (e), (f), (g), (h) three-day length actigraphy, (i), (j), (k), (l) four-day length actigraphy, (m), (n), (o), (p) five-day length actigraphy, (q), (r), (s), (t) six-day length actigraphy, (u), (v), (w), (x) seven-day length actigraphy, (a), (e), (i), (m), (q), (u) XGBoost classifier, (b), (f), (j), (n), (r), (v) SVC, (c), (g), (k), (o), (s), (w) MLP, (d), (h), (l), (p), (t), (x) LR.

1000 samples through stratified random sampling. After experimenting with a different-sized dataset, we found that the tendency in previous experiment results was repeated in these experiments (i.e., performance under a two-day length condition showed the highest evaluation values).

Based on the two experimental results, we found that the optimal duration of actigraphy monitoring to effectively determine depression status was two days, regardless of the dataset size. Similarly, Thomas *et al.* [75] investigated the reliability of actigraphy length using different individuals as a case study and suggested that a two-day period adequately reflected the circadian rhythm of actigraphy. Thus, we concluded that the actigraphy data of two days were sufficient for feature extraction to classify depression levels.

In the case of feature importance, we investigated the ranked features in XGBoost classifiers. In the XGBoost

algorithm, the F1-score was calculated based on the number of times that the decision tree model was used for estimation. A detailed list of important features is presented in Table 12.

The ranked features of the XGBoost classifiers were the same under all conditions. A total of 15 input features were selected by regression models in the feature selection steps and consisted of 10 survey variables and 5 circadian rhythm indices. The factors identified in previous studies on depression and physical activity were justified by the feature ranking in this study. The authors focused on mean activity levels and lower values of physical activity to identify factors associated with depressed individuals. Additionally, they observed lower values of physical activity in the depressed groups [76]–[78]. In summary, our study represents a reliable experimental paradigm in terms of both classification performance and feature importance for classification.

V. CONCLUSION

Classifying depression levels is critical for various fields, including clinical and psychological domains. In this study, we proposed a framework for classifying depression levels using ML algorithms. Based on previous studies on the relationship between depression and physical activity, actigraphy data using an accelerometer were used to extract circadian rhythm indices as features. To evaluate our framework from a diverse perspective, we designed experiments with various class labels and actigraphy length conditions. We found that the XGBoost classifier exhibited the best classification performance and that two days of actigraphy data were suitable for representing the circadian cycle in physical activity.

The first strength of this study was the application of accelerometer-based actigraphy data, which are not widely used to classify depression levels. Second, we determined the ideal length of actigraphy data for feature extraction. Third, a large-scale real-world dataset collected from people living in Korea was used to reflect practical tendencies.

Our study has some limitations. First, actigraphy data included detailed differences (e.g., gap between morning and afternoon, difference between weekdays and weekends). These differences can affect depression levels. However, we considered overall characteristics instead of specific changes to classify depression levels. Second, various classification methodologies including deep learning algorithms can be applied to solve our research questions. To facilitate confirmation of feature importance, we used ML algorithms in our study. Third, we need to consider external validation through datasets collected from other countries to generalize our framework in further study.

APPENDIX A

The ROC curves under other (three, four, five, six, and seven days length) conditions with class conditions (binary, three, four, and five classes). See Figures 9–12.

APPENDIX B

Experimental results under 1000-rows dataset. See Tables 13–16 and Figures 13–16.

REFERENCES

[1] A. Doris, K. Ebmeier, and P. Shajahan, "Depressive illness," *Lancet*, vol. 354, no. 9187, pp. 1369–1375, 1999.

[2] G. I. Papakostas, T. Petersen, Y. Mahal, D. Mischoulon, A. A. Nierenberg, and M. Fava, "Quality of life assessments in major depressive disorder: A review of the literature," *Gen. Hospital Psychiatry*, vol. 26, no. 1, pp. 13–17, Jan. 2004.

[3] N. Breslau and G. C. Davis, "Chronic stress and major depression," *Arch. Gen. Psychiatry*, vol. 43, no. 4, pp. 309–314, 1986.

[4] M. L. Hackett, C. Yapa, V. Parag, and C. S. Anderson, "Frequency of depression after stroke: A systematic review of observational studies," *Stroke*, vol. 36, no. 6, pp. 1330–1340, Jun. 2005.

[5] A. G. Billings, R. C. Cronkite, and R. H. Moos, "Social-environmental factors in unipolar depression: Comparisons of depressed patients and nondepressed controls," *J. Abnormal Psychol.*, vol. 92, no. 2, pp. 119–133, 1983.

[6] S. Ayad, A. K. Khanna, S. U. Iqbal, and N. Singla, "Characterisation and monitoring of postoperative respiratory depression: Current approaches and future considerations," *Brit. J. Anaesthesia*, vol. 123, no. 3, pp. 378–391, Sep. 2019.

[7] A. Charbonneau, A. K. Rosen, R. R. Owen, A. Spiro, A. S. Ash, D. R. Miller, L. Kazis, B. Kader, F. Cunningham, and D. R. Berlowitz, "Monitoring depression care: In search of an accurate quality indicator," *Med. Care*, vol. 42, no. 6, pp. 522–531, 2004.

[8] I. P. Clara, B. J. Cox, and M. W. Enns, "Confirmatory factor analysis of the depression-anxiety-stress scales in depressed and anxious patients," *J. Psychopathology Behav. Assessment*, vol. 23, no. 1, pp. 61–67, 2001.

[9] P. Muris, H. Schmidt, H. Merckelbach, and E. Schouten, "Anxiety sensitivity in adolescents: Factor structure and relationships to trait anxiety and symptoms of anxiety disorders and depression," *Behav. Res. Therapy*, vol. 39, no. 1, pp. 89–100, Jan. 2001.

[10] Y. Mohan, S. S. Chee, D. K. P. Xin, and L. P. Foong, "Artificial neural network for classification of depressive and normal in EEG," in *Proc. IEEE EMBS Conf. Biomed. Eng. Sci. (IECBES)*, Dec. 2016, pp. 286–290.

[11] S. D. Puthankattil and P. K. Joseph, "Classification of EEG signals in normal and depression conditions by ANN using RWE and signal entropy," *J. Mech. Med. Biol.*, vol. 12, no. 4, Sep. 2012, Art. no. 1240019.

[12] S. S. Mikkelsen, J. S. Tolstrup, E. M. Flachs, E. L. Mortensen, P. Schnohr, and T. Flensborg-Madsen, "A cohort study of leisure time physical activity and depression," *Preventive Med.*, vol. 51, no. 6, pp. 471–475, Dec. 2010.

[13] A. K. Lebedeva, E. Westman, T. Borza, M. K. Beyer, K. Engedal, D. Aarsland, G. Selbaek, and A. K. Haberg, "MRI-based classification models in prediction of mild cognitive impairment and dementia in late-life depression," *Frontiers Aging Neurosci.*, vol. 9, p. 13, Feb. 2017.

[14] S. C. Gorber and S. M. Tremblay, *The Objective Monitoring of Physical Activity: Contributions of Accelerometry to Epidemiology, Exercise Science and Rehabilitation*. New York, NY, USA: Springer, 2016, pp. 369–376.

[15] S. Bauhoff, "Systematic self-report bias in health data: Impact on estimating cross-sectional and treatment effects," *Health Services Outcomes Res. Methodol.*, vol. 11, nos. 1–2, pp. 44–53, Jul. 2011.

[16] W. Mumtaz, A. S. Malik, M. A. M. Yasin, and L. Xia, "Review on EEG and ERP predictive biomarkers for major depressive disorder," *Biomed. Signal Process. Control*, vol. 22, pp. 85–98, Sep. 2015.

[17] S. Olbrich, A. Tränkner, T. Chittka, U. Hegerl, and P. Schönknecht, "Functional connectivity in major depression: Increased phase synchronization between frontal cortical EEG-source estimates," *Psychiatry Res., Neuroimag.*, vol. 222, nos. 1–2, pp. 91–99, Apr. 2014.

[18] K. Imaoka, H. Inoue, Y. Inoue, H. Hazama, T. Tanaka, and N. Yamane, "R-R intervals of ECG in depression," *Psychiatry Clin. Neurosci.*, vol. 39, no. 4, pp. 485–487, 1985.

[19] Z.-X. Zhang, X.-W. Tian, and J. S. Lim, "New algorithm for the depression diagnosis using HRV: A neuro-fuzzy approach," in *Proc. Int. Symp. Bioelectron. Bioinf.*, Nov. 2011, pp. 283–286.

[20] Y. Tazawa, K.-C. Liang, M. Yoshimura, M. Kitazawa, Y. Kaise, A. Takamiya, A. Kishi, T. Horigome, Y. Mitsukura, M. Mimura, and T. Kishimoto, "Evaluating depression with multimodal wristband-type wearable device: Screening and assessing patient severity utilizing machine-learning," *Heliyon*, vol. 6, no. 2, Feb. 2020, Art. no. e03274.

[21] A. Salguero, R. Martínez-García, O. Molinero, and S. Márquez, "Physical activity, quality of life and symptoms of depression in community-dwelling and institutionalized older adults," *Arch. Gerontol. Geriatrics*, vol. 53, no. 2, pp. 152–157, Sep. 2011.

[22] S. Weyerer, "Physical inactivity and depression in the community," *Int. J. Sports Med.*, vol. 13, no. 6, pp. 492–496, Aug. 1992.

[23] M. L. McPhie and J. S. Rawana, "The effect of physical activity on depression in adolescence and emerging adulthood: A growth-curve analysis," *J. Adolescence*, vol. 40, pp. 83–92, Apr. 2015.

[24] J. K. Vallance, E. A. H. Winkler, P. A. Gardiner, G. N. Healy, B. M. Lynch, and N. Owen, "Associations of objectively-assessed physical activity and sedentary time with depression: NHANES (2005–2006)," *Preventive Med.*, vol. 53, nos. 4–5, pp. 284–288, Oct. 2011.

[25] D. Currier, R. Lindner, M. J. Spittal, S. Cvetkovski, J. Pirkis, and D. R. English, "Physical activity and depression in men: Increased activity duration and intensity associated with lower likelihood of current depression," *J. Affect. Disorders*, vol. 260, pp. 426–431, Jan. 2020.

[26] N. Mutrie and M. K. Hannah, "The importance of both setting and intensity of physical activity in relation to non-clinical anxiety and depression," *Int. J. Health Promotion Educ.*, vol. 45, no. 1, pp. 24–32, Jan. 2007.

[27] A. C. Wilson and T. M. Palermo, "Physical activity and function in adolescents with chronic pain: A controlled study using actigraphy," *J. Pain*, vol. 13, no. 2, pp. 121–130, Feb. 2012.

[28] S. Kashikar-Zuck, S. R. Flowers, E. Verkamp, T. V. Ting, A. M. Lynch-Jordan, T. B. Graham, M. Passo, K. N. Schikler, P. J. Hashkes, S. Spalding, G. Banez, M. M. Richards, S. W. Powers, L. M. Arnold, and D. Lovell, "Actigraphy-based physical activity monitoring in adolescents with juvenile primary fibromyalgia syndrome," *J. Pain*, vol. 11, no. 9, pp. 885–893, Sep. 2010.

[29] A. Artese, D. Ehley, A. R. Sutin, and A. Terracciano, "Personality and actigraphy-measured physical activity in older adults," *Psychol. Aging*, vol. 32, no. 2, pp. 131–138, Mar. 2017.

[30] I. Dolu and N. O. Nahcivan, "Feasibility of the duration of actigraphy data to illustrate circadian rhythm among cognitively intact older people in nursing home: Cosinor analysis," *Sleep Biol. Rhythms*, vol. 18, no. 1, pp. 59–64, Jan. 2020.

[31] G. A. Dowling, R. L. Burr, E. J. W. Van Someren, E. M. Hubbard, J. S. Luxenberg, J. Mastick, and B. A. Cooper, "Melatonin and bright-light treatment for rest-activity disruption in institutionalized patients with Alzheimer's disease," *J. Amer. Geriatrics Soc.*, vol. 56, no. 2, pp. 239–246, Feb. 2008.

[32] S. Hopkins, P. L. Morgan, L. J. M. Schlangen, P. Williams, D. J. Skene, and B. Middleton, "Blue-enriched lighting for older people living in care homes: Effect on activity, actigraphic sleep, mood and alertness," *Current Alzheimer Res.*, vol. 14, no. 10, pp. 1053–1062, Sep. 2017.

[33] E. Garcia-Ceja, M. Riegler, P. Jakobsen, J. Torresen, T. Nordgreen, K. J. Oedegaard, and O. B. Fasmer, "Motor activity based classification of depression in unipolar and bipolar patients," in *Proc. IEEE 31st Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2018, pp. 316–321.

[34] A. C. Tonon, D. F. P. Fuchs, W. B. Gomes, R. Levandovski, M. P. de Almeida Fleck, M. P. L. Hidalgo, and L. da Silva Alencastro, "Nocturnal motor activity and light exposure: Objective actigraphy-based marks of melancholic and non-melancholic depressive disorder. Brief report," *Psychiatry Res.*, vol. 258, pp. 587–590, Dec. 2017.

[35] K. H. White, M. E. Rumble, and R. M. Benca, "Sex differences in the relationship between depressive symptoms and actigraphic assessments of sleep and rest-activity rhythms in a population-based sample," *Psychosomatic Med.*, vol. 79, no. 4, pp. 479–484, 2017.

[36] A. M. Berger, K. Wielgus, M. Hertzog, P. Fischer, and L. Farr, "Patterns of circadian activity rhythms and their relationships with fatigue and anxiety/depression in women treated with breast cancer adjuvant chemotherapy," *Supportive Care Cancer*, vol. 18, no. 1, pp. 105–114, Jan. 2010.

[37] P. Rock, G. Goodwin, C. Harmer, and K. Wulff, "Daily rest-activity patterns in the bipolar phenotype: A controlled actigraphy study," *Chronobiol. Int.*, vol. 31, no. 2, pp. 290–296, Mar. 2014.

[38] E. M. C. Feliciano, M. Quante, J. Weng, J. A. Mitchell, P. James, C. R. Marinac, S. Mariani, S. Redline, J. Kerr, S. Godbole, A. Manteiga, D. Wang, and J. A. Hipp, "Actigraphy-derived daily rest–activity patterns and body mass index in community-dwelling adults," *Sleep*, vol. 40, no. 12, Dec. 2017, Art. no. zsx168.

[39] K. Suibkitwanchai, A. M. Sykulski, G. P. Algorta, D. Waller, and C. Walshe, "Nonparametric time series summary statistics for high-frequency accelerometry data from individuals with advanced dementia," 2020, *arXiv:2005.01171*. [Online]. Available: http://arxiv.org/abs/2005.01171

[40] S. Vinzio, A. Ruellan, A.-E. Perrin, J.-L. Schlienger, and B. Goichot, "Actigraphic assessment of the circadian rest-activity rhythm in elderly patients hospitalized in an acute care unit," *Psychiatry Clin. Neurosci.*, vol. 57, no. 1, pp. 53–58, Feb. 2003.

[41] A. Germain and D. J. Kupfer, "Circadian rhythm disturbances in depression," *Human Psychopharmacol., Clin. Experim.*, vol. 23, no. 7, pp. 571–585, Oct. 2008.

[42] T. A. Wehr, D. Sack, N. Rosenthal, W. Duncan, and J. C. Gillin, "Circadian rhythm disturbances in manic-depressive illness," *Fed. Proc.*, vol. 42, no. 11, pp. 2809–2814, 1983.

[43] A. Khademi, Y. EL-Manzalawy, L. Master, O. M. Buxton, and V. G. Honavar, "Personalized sleep parameters estimation from actigraphy: A machine learning approach," *Nature Sci. Sleep*, vol. 11, pp. 387–399, Dec. 2019.

[44] (2014). *The Sixth Korea National Health and Nutrition Examination Survey (KNHANES VI-2). Korea Centers for Disease Control and Prevention*. Accessed: Apr. 14, 2021. [Online]. Available: https://knhanes.cdc.go.kr/knhanes/eng/index.do

[45] (2016). *The Sixth Korea National Health and Nutrition Examination Survey (KNHANES VI-3). Korea Centers for Disease Control and Prevention*. Accessed: Apr. 14, 2021. [Online]. Available: https://knhanes.cdc.go.kr/knhanes/eng/index.do

[46] P. G. Sokolove and W. N. Bushell, "The chi square periodogram: Its utility for analysis of circadian rhythms," *J. Theor. Biol.*, vol. 72, no. 1, pp. 131–160, 1978.

[47] M. C. Moore-Ede, F. M. Sulzman, F. M. Sulzman, and C. A. Fuller, *The Clocks That Time Us: Physiology of the Circadian Timing System*. Cambridge, MA, USA: Harvard Univ. Press, 1982, pp. 79–80.

[48] E. J. W. Van Someren, C. Lijzenga, M. Mirmiran, and D. F. Swaab, "Long-term fitness training improves the circadian rest-activity rhythm in healthy elderly males," *J. Biol. Rhythms*, vol. 12, no. 2, pp. 146–156, Apr. 1997.

[49] E. J. Van Someren, B. F. Vonk, W. A. Thijssen, J. D. Speelman, P. R. Schuurman, M. Mirmiran, and D. F. Swaab, "A new actigraph for long-term registration of the duration and intensity of tremor and movement," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 3, pp. 386–395, Mar. 1998.

[50] A. C. Brown, M. H. Smolensky, G. E. D'Alonzo, and D. P. Redman, "Actigraphy: A means of assessing circadian patterns in human activity," *Chronobiol. Int.*, vol. 7, no. 2, pp. 125–133, Jan. 1990.

[51] R. Tibshirani, "Regression shrinkage and selection via the lasso: A retrospective," *J. Roy. Statist. Soc. B, Statist. Methodol.*, vol. 73, no. 3, pp. 273–282, 2011.

[52] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[53] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[54] Q. Wu and D. X. Zhou, "Analysis of support vector machine classification," *J. Comput. Anal. Appl.*, vol. 8, no. 2, pp. 1–21, 2006.

[55] M. M. Sani, H. Norhazman, H. A. Omar, N. Zaini, and S. A. Ghani, "Support vector machine for classification of stress subjects using EEG signals," in *Proc. IEEE Conf. Syst., Process Control (ICSPC)*, Dec. 2014, pp. 127–131.

[56] P.-L. Wu, M. Lee, and T.-T. Huang, "Effectiveness of physical activity on patients with depression and Parkinson's disease: A systematic review," *PLoS ONE*, vol. 12, no. 7, Jul. 2017, Art. no. e0181515.

[57] C. M. Teixeira, J. Vasconcelos-Raposo, H. M. Fernandes, and R. J. Brustad, "Physical activity, depression and anxiety among the elderly," *Social Indicators Res.*, vol. 113, no. 1, pp. 307–318, Aug. 2013.

[58] B. Roshanaei-Moghaddam, W. J. Katon, and J. Russo, "The longitudinal effects of depression on physical activity," *Gen. Hospital Psychiatry*, vol. 31, no. 4, pp. 306–315, Jul. 2009.

[59] P. W. Ku, K. R. Fox, L. J. Chen, and P. Chou, "Physical activity and depressive symptoms in older adults: 11-year follow-up," *Amer. J. Preventive Med.*, vol. 42, no. 4, pp. 355–362, 2012.

[60] S. Albahli, H. T. Rauf, M. Arif, M. T. Nafis, and A. Algosaibi, "Identification of thoracic diseases by exploiting deep neural networks," *Comput., Mater. Continua*, vol. 66, no. 3, pp. 3139–3149, 2021.

[61] S. Albahli, H. T. Rauf, A. Algosaibi, and V. E. Balas, "AI-driven deep CNN approach for multi-label pathology classification using chest X-rays," *PeerJ Comput. Sci.*, vol. 7, Apr. 2021, Art. no. e495.

[62] A. M. Chekroud, R. J. Zotti, R. Shehzad, R. Gueorguieva, M. K. Johnson, M. H. Trivedi, T. D. Cannon, J. H. Krystal, and P. R. Corlett, "Cross-trial prediction of treatment outcome in depression: A machine learning approach," *Lancet Psychiatry*, vol. 3, no. 3, pp. 243–250, Mar. 2016.

[63] I. Bhakta and A. Sau, "Prediction of depression among senior citizens using machine learning classifiers," *Int. J. Comput. Appl.*, vol. 144, no. 7, pp. 11–16, Jun. 2016.

[64] B. Hosseinifard, M. H. Moradi, and R. Rostami, "Classifying depression patients and normal subjects using machine learning techniques and non-linear features from EEG signal," *Comput. Methods Programs Biomed.*, vol. 109, no. 3, pp. 339–345, 2013.

[65] Y. Mohammadi, M. Hajian, and M. H. Moradi, "Discrimination of depression levels using machine learning methods on EEG signals," in *Proc. 27th Iranian Conf. Electr. Eng. (ICEE)*, Apr. 2019, pp. 1765–1769.

[66] M. F. Rizwan, R. Farhad, F. Mashuk, F. Islam, and M. H. Imam, "Design of a biosignal based stress detection system using machine learning techniques," in *Proc. Int. Conf. Robot., Elect. Signal Process. Techn. (ICREST)*, Jan. 2019, pp. 364–368.

[67] X. Zhong, H. Shi, Q. Ming, D. Dong, X. Zhang, L.-L. Zeng, and S. Yao, "Whole-brain resting-state functional connectivity identified major depressive disorder: A multivariate pattern analysis in two independent samples," *J. Affect. Disorders*, vol. 218, pp. 346–352, Aug. 2017.

[68] M. T. De Mello, V. D. A. Lemos, H. K. M. Antunes, L. Bittencourt, R. Santos-Silva, and S. Tufik, "Relationship between physical activity and depression and anxiety symptoms: A population study," *J. Affect. Disorders*, vol. 149, nos. 1–3, pp. 241–246, Jul. 2013.

[69] S. Sabia, A. Dugravot, J.-F. Dartigues, J. Abell, A. Elbaz, M. Kivimäki, and A. Singh-Manoux, "Physical activity, cognitive decline, and risk of dementia: 28 year follow-up of whitehall II cohort study," *BMJ*, vol. 357, Jun. 2017, Art. no. j2709.

[70] T. J. Harris, C. G. Owen, C. R. Victor, R. Adams, U. Ekelund, and D. G. Cook, "A comparison of questionnaire, accelerometer, and pedometer: Measures in older people," *Med. Sci. Sports Exerc.*, vol. 41, no. 7, pp. 1392–1402, 2009.

[71] U. R. Acharya, V. K. Sudarshan, H. Adeli, J. Santhosh, J. E. W. Koh, S. D. Puthankatti, and A. Adeli, "A novel depression diagnosis index using nonlinear features in EEG signals," *Eur. Neurol.*, vol. 74, nos. 1–2, pp. 79–83, 2015.

[72] H. Rubin-Falcone, F. Zanderigo, B. Thapa-Chhetry, M. Lan, J. M. Miller, M. E. Sublette, M. A. Oquendo, D. J. Hellerstein, P. J. McGrath, J. W. Stewart, and J. J. Mann, "Pattern recognition of magnetic resonance imaging-based gray matter volume measurements classifies bipolar disorder and major depressive disorder," *J. Affect. Disorders*, vol. 227, pp. 498–505, Feb. 2018.

[73] T. Roh, S. Hong, and H.-J. Yoo, "Wearable depression monitoring system with heart-rate variability," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 562–565.

[74] A. Y. Kim, E. H. Jang, S. Kim, K. W. Choi, H. J. Jeon, H. Y. Yu, and S. Byun, "Automatic detection of major depressive disorder using electrodermal activity," *Sci. Rep.*, vol. 8, no. 1, pp. 1–9, Dec. 2018.

[75] K. A. Thomas and R. L. Burr, "Circadian research in mothers and infants: How many days of actigraphy data are needed to fit cosinor parameters?" *J. Nursing Meas.*, vol. 16, no. 3, pp. 201–206, Dec. 2008.

[76] O. Minaeva, S. H. Booij, F. Lamers, N. Antypa, R. A. Schoevers, M. Wichers, and H. Riese, "Level and timing of physical activity during normal daily life in depressed and non-depressed individuals," *Transl. Psychiatry*, vol. 10, no. 1, pp. 1–11, Dec. 2020.

[77] T. C. Camacho, R. E. Roberts, N. B. Lazarus, G. A. Kaplan, and R. D. Cohen, "Physical activity and depression: Evidence from the alameda county study," *Amer. J. Epidemiol.*, vol. 134, no. 2, pp. 220–231, Jul. 1991.

[78] P. J. O'Connor, L. E. Aenchbacher, and R. K. Dishman, "Physical activity and depression in the elderly," *J. Aging Phys. Activity*, vol. 1, no. 1, pp. 34–58, Oct. 1993.

JUNG-GU CHOI was born in 1993. He is currently pursuing the Ph.D. degree in cognitive science with the Applied Brain Cognitive Laboratory, Yonsei University Graduate Program. His research interests include brain science and data mining.

INHWAN KO is currently pursuing the M.S. degree in psychology with Yonsei University. He has a diverse working experience from a human resource management practitioner in multinational companies to a certified public labor attorney. His research interests include converging human resource management and cognitive neuroscience.

SANGHOON HAN was born in 1977. He is currently a Professor with the Department of Psychology, Yonsei University. His research interests include decision making and cognitive science.

● ● ●