applied sciences

MDPI

Article

Identification of Geriatric Depression and Anxiety Using Activity Tracking Data and Minimal Geriatric Assessment Scales

Tae-Rim Lee [1], Geon-Ha Kim [2] and Mun-Taek Choi [3,*]

[1] Department of Artificial Intelligence, Sungkyunkwan University, Suwon 16419, Korea; xofla5454@g.skku.edu
[2] Department of Neurology, EWHA Womans University Mokdong Hospital, College of Medicine, EWHA Womans University , Seoul 07985, Korea; geonha@ewha.ac.kr
[3] School of Mechanical Engineering, Sungkyunkwan University, Suwon 16419, Korea
* Correspondence: mtchoi@skku.edu

Abstract: The identification of geriatric depression and anxiety is important because such conditions are the most common comorbid mood problems that occur in older adults. The goal of this study was to build a machine learning framework that identifies geriatric mood disorders of depression and anxiety using low-cost activity trackers and minimal geriatric assessment scales. We collected activity tracking data from 352 mild cognitive impairment patients, from 60 to 90 in age, by having them wear activity trackers on their wrist for more than a month. We then extracted the features of 24-h activity rhythms and sleep patterns from the time-series activity tracking data. To increase the accuracy, we designed a novel method to incorporate additional features from questionnaire-based assessments of the geriatric depression scale and geriatric anxiety inventory into the activity tracking features. In the multi-label classification, we applied the binary relevance method to develop two single-label classifiers for depression and anxiety. The best hyper-parameters of classification algorithms for each label were selected by comparing the classification performance. We finally selected the combination of classifiers for depression and anxiety with the lowest Hamming loss as a multi-label classifier. This study successfully demonstrated the possibility of identifying geriatric depression and anxiety using low-cost activity trackers and minimal geriatric assessment scales for use in the real fields.

Keywords: depression; anxiety; activity tracker; multi-label classification; 24-h activity rhythms; sleep patterns; geriatric mood disorders; binary relevance

check for updates

Citation: Lee, T.-R.; Kim, G.-H.; Choi, M.-T. Identification of Geriatric Depression and Anxiety Using Activity Tracking Data and Minimal Geriatric Assessment Scales. *Appl. Sci.* **2022**, *12*, 2488. https://doi.org/10.3390/app12052488

Academic Editor: Johann Eder

Received: 14 January 2022
Accepted: 24 February 2022
Published: 27 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

It has been reported that 50 million people have been diagnosed with dementia worldwide [1]. Despite efforts toward the development of disease-modifying treatments for dementia, none have been successfully able to treat the cognitive dysfunction in such patients with dementia. Therefore, recent studies have focused on early detection and intervention for modifiable risk factors to mitigate the development of dementia in older adults, which include sedentary lifestyle, cardiovascular disease and mood problems such as depression and anxiety [2].

Specifically, depression and anxiety are one of the most common comorbid mood problems in older adults [3]. Anxiety disorders are characterized by an exaggerated fear response and attempts to reduce, escape or avoid threat, whereas depressive symptoms are characterized by sad mood or loss of interest in activities, sleep dysregulation, feelings of worthlessness, appetite changes and fatigue [3]. Previous studies have shown that older adults with depression and anxiety [1,2] are at an increased risk of dementia. Barnes et al. demonstrated a significant association between depression and the risk of mild cognitive impairment (MCI) [4], whereas it has been noted that anxiety symptoms may increase the risk of dementia in older adults. Therefore, an earlier diagnosis as well as the appropriate management of such mood symptoms could be important in mitigating further cognitive impairment in older adults.

It is challenging for physicians to diagnose patients accurately who present with depression and anxiety. Anxiety and depression have distinct psychological features. Given that patients showed distinct symptoms and need different medical management according to the diagnosis of depression or anxiety in clinical practice, clinicians should differentiate whether patients had depression or anxiety accurately. Clinicians typically use the following self-reporting questionnaires to assess the major geriatric mood disorders: the geriatric depression scale (GDS) for depression symptoms [5] and geriatric anxiety inventory (GAI) for anxiety symptoms [6]. Both GDS and GAI are composed of questionnaires with yes or no questions that are answered by the patients or their families. Based on the assessment results, a patient can be diagnosed as either depressed or anxious, or a combination of both. However, self-reporting methods such as GDS and GAI may contain incorrect information because they rely on answers received from the patients or their families. Additionally, the large number of answers required by such questionnaires can be inconvenient to the patients. To reduce these limitations, there have been efforts to reduce the number of questionnaires. Hoyl et al. verified that the five items in the GDS short form are statistically significant even though the form has a total of fifteen questionnaires [7]. Byrne et al. demonstrated the possibility of using GAI questionnaires with five items instead of twenty [8]. Instead of such self-reporting methods, an accurate diagnosis requires complex procedures using expensive equipment such as functional magnetic resonance imaging (fMRI). In assessing geriatric mood disorders, more convenient, relatively low-cost and reliable systematic approaches through direct measurements are needed.

An increased risk of mood disorders are known to be associated with disorders of Circadian rhythms, endogenous rhythms with a periodicity of approximately 24 h and sleep [9]. In fact, the patterns of 24-h activity rhythms and sleep can be extracted from time-series activity tracking data from activity trackers [10], such as ActiGraph [11] and Fitbit [12], which are increasingly used in clinical and daily life these days. ActiGraph is known to be relatively accurate enough for clinical use but is far more expensive than a low-cost activity tracker such as Fitbit. Mendlowicz et al. [13] found a correlation between daytime activity levels and depression using ActiGraph. This study focused on depression only, and the total number of participants was 32, which was relatively low. Cook et al. [14] used Fitbit to statistically evaluate sleep in depressive disorder. Spira et al. [15] verified that anxiety symptoms affect sleep quality and sleep fragmentation from Actigraph data. Luik et al. [10] collected activity tracking data from 1714 middle-aged and elderly participants using ActiGraph and extracted six features for the 24-h activity rhythms and sleep. They found statistically significant results that the 24-h activity rhythm and sleep are related to geriatric depression and anxiety. Their study was a major clinical motivation for our study to develop a classification model for identifying geriatric depression and anxiety using activity rhythms and sleep patterns. However, in our case, we used data from low-cost activity trackers for practicability.

Some studies have attempted to apply machine learning techniques for identifying mood disorders using biosignals from various sensors. Xiaowei et al. [16] conducted a study to better recognize depression using electroencephalogram (EEG) features and machine learning methods. They studied only depression in a limited number of 28 subjects and used EEG in a controlled environment. Ghandeharioun et al. [17] conducted a study to predict depressive symptoms based on the Hamilton Depression Rating Scale (HDRS) by machine learning techniques applied to data acquired from sensors on E4 wearable wristbands and Android phones. They used additional input data, including electrodermal activity (EDA), location changes, phone-based communication and phone usage patterns. Rykov et al. [18] investigated the possibility of detecting depression using digital biomarkers, such as steps, heart rate, energy expenditure and sleep data, from consumer-grade wearable sensors. They discovered their classification model's potential to assist in depression screening with good accuracy.

Since patients can have both depression and anxiety symptoms at the same time, multi-label classification arises when applying machine learning. In many fields, including biology, music and linguistics, interest in multi-label classification is increasing [19]. Multi-label classification problems allow one data sample to belong to multiple classes simultaneously [19–21]. In single-label classification problems, a data sample can only belong to one class. The binary relevance (BR) is one simple solution to the multi-label classification problem. BR solves multi-label classification problems by creating an independent binary classifier for each label [20,22]. Several studies have used the BR to solve multi-label classification problems. Zhang et al. applied BR to specific features to perform classification for different labels [23]. They achieved an enhanced performance using label-specific features. Huang et al. suggested sharing label-specific features between correlated labels for the classification for each label [24]. This method outperformed other multi-label classification methods. To overcome the limitation of the BR method in not considering correlations between labels, Alvares-Cherman et al. proposed an approach that allows binary classifiers to discover label dependencies [25]. Comparing the performances of various classification algorithms is an important process for machine learning; however, previous studies have not applied multiple classification algorithms to the binary classification of each label in the BR method.

This study aims to develop a clinically significant classification model to identify comorbid geriatric depression and anxiety using both activity tracking data from low-cost activity trackers and minimal questionnaire-based geriatric assessment scales. This study extends our earlier work [26] using only low-cost activity tracking data that showed potential but needed significant improvement in predictive accuracy for field use. The novelties of our approach are as follows. Full activity rhythms and sleep-related features are extracted from activity tracking data collected from older adults using low-cost activity trackers. A minimal number of GDS and GAI features are selected for the convenience in diagnosis based on feature importance measured from full elderly assessment scale data. A new method is proposed to ensure high classification performance by combining the activity tracking features and minimal GDS and GAI features. A basis framework for a multi-label classification method used to identify compound symptoms of geriatric depression and anxiety is established.

2. Materials and Methods

2.1. Study Participants

All participants of this study were recruited from Ewha Womans University Mokdong Hospital. All participants gave written agreement, and the research protocols were approved by the Ewha Women's University Mokdong Hospital Institutional Review Board (EUMC 2016-09-042-013). The participants were evaluated eligibility using the following inclusion criteria: (1) between 60 and 90 years in age, (2) literate, (3) not diagnosed with dementia, and (4) scored ≥ -1.5 SD of the mean of age and education-matched norm on the Korean version of MMSE (K-MMSE) [27]. Individuals with any of the following have been excluded from the study: (1) visual or hearing impairments severe enough to interfere with the questionnaire response, (2) any major medical problems such as cardiovascular disease or cancer, and (3) who refused to wear activity trackers. Additionally, we excluded cases where it was judged that it would be difficult to fill out the questionnaire or difficult to obtain objective information by completing the questionnaire. We initially recruited 358 older adults without dementia. From those 358 participants, two individuals were excluded from the analysis as they withdrew their consent to participate in the research, while the other four refused to wear the activity trackers for more than one month. Therefore, a total of 352 participants were finally analyzed for this study. The average (±standard deviation) age and the ratio of females were 72.48 (±5.9) years and 73.01%, respectively.

2.2. Geriatric Depression and Anxiety Assessment

All participants were assessed for geriatric depression and anxiety prior to the start of the experiments as follows. The assessment of depression was performed using the Korean version of the short form of Geriatric Depression Scale (K-SGDS)—a questionnaire that is commonly used to screen depressive symptoms in older adults [27,28]. GDS and K-SGDS are used interchangeably in this paper. K-SGDS consists of 15 questions with "yes" or "no" answers, and a score of 8 or more is the cut-off for determining depression [27]. A partial list of the questionnaire items in K-SGDS is listed in Table 1.

Table 1. Partial list of questionnaire items in K-SGDS.

Item No.	Questionnaire
GDS 3	Do you feel that your like is empty?
GDS 5	Are you in good spirits most of the time?
GDS 8	Do you often feel helpless?
GDS 14	Do you feel your situation is hopeless?
GDS 15	Do you think that most people are better off than you are?

Anxiety symptoms of the participants were evaluated using the Korean version of Geriatric Anxiety Inventory (K-GAI) [29]—a commonly used questionnaire to identify anxiety symptoms in older adults. GAI and K-GAI are used interchangeably in this paper. The K-GAI consists of 20 questions with "yes" or "no" answers and a score of 7 or more is the cut off for suggesting anxiety [29]. A partial list of the questionnaire items in K-GAI is listed in Table 2.

Table 2. Partial list of questionnaire items in K-GAI.

Item No.	Questionnaire
GAI 4	I find it hard to relax.
GAI 5	I often cannot enjoy things because of my worries.
GAI 9	I cannot help worrying about even trivial things.
GAI 11	My own thoughts often make me anxious.
GAI 20	I often feel upset.

For this study, an experienced clinician (G. H. Kim) interviewed and evaluated all participants whether the participants had depression or anxiety after they completed the questionnaire of K-GAI and K-GDS. Then, all the participants were subdivided into four groups: none (non-depression and non-anxiety), depression only, anxiety only and both depression and anxiety, which is generally acceptable in clinical practice. We finally used the clinician's diagnosis of depression and anxiety as labels for the data from activity trackers.

2.3. Activity Tracking Data Acquisition

To collect the 24-h activity tracking data from subjects, we used a Fitbit Alta HR2 [12], a low-cost wristwatch wearable activity tracker. Figure 1 shows the data acquisition using the Fitbit Alta HR2 in the overall data acquisition and analysis system for this study. The application on the smart mobile device uploads the data measured by the Fitbit device to the Fitbit cloud. The restful application programming interface, called Fitbit Web API, is used to receive data from the Fitbit cloud [30].

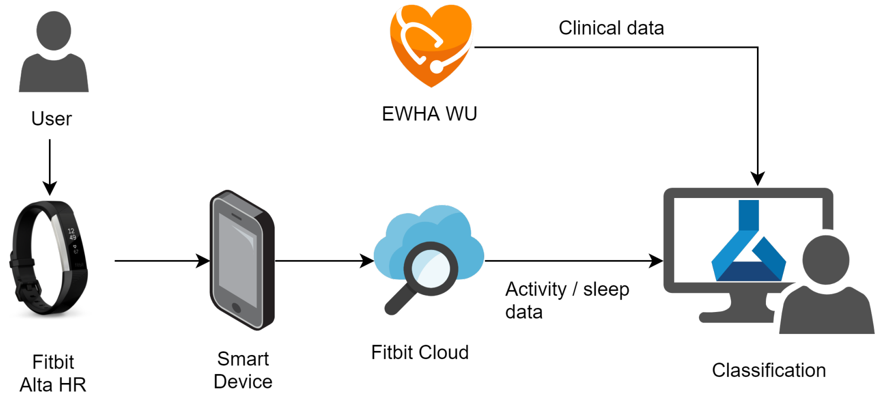


Figure 1. Data acquisition and analysis system.

The Fitbit Alta HR2 senses the heart beats per minute, steps per minute, calories per minute and distance per minute. The device also provides sleep-related data through the sensed data. Sleep-related data include start and end of the sleep times, sleep quality, total sleep time and sleep stages per minute.

In this study, data were collected as follows. All participants wore the wearable devices for at least one month during their daily lives. During this period, the data available from the Fitbit device, including the heart rate, steps and sleep, were collected. The subjects were instructed to keep the devices on while sleeping, washing, moving and exercising, among other activities.

2.4. Activity Tracking Feature Extraction

We extracted seven activity tracking features from the time-series data obtained using the Fitbit devices. The features related to the 24-h activity rhythms and sleep patterns include inter-daily stability (IS), intra-daily variability (IV) and dominant rest phase onset (DRO). The sleep-related features include the total sleep time (TST), sleep onset latency (SOL), wake after sleep onset (WASO) and sleep quality [31]. These features were inspired from the results of a study conducted by Luik et al., which demonstrated the association of these features with depression and anxiety in older adults [10].

IS represents the stability of an activity rhythm of a daily life pattern [32], given by

$$IS = \frac{n \sum_{h=1}^{p} (\bar{x}_h - \bar{x})^2}{p \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{1}$$

where i denotes each hourly point, n denotes the total number of data, x denotes the number of movements induced by steps, x_i denotes the individual hourly data, $\bar{x}$ denotes the mean of all data, p denotes the number of hourly data per day and $\bar{x}_h$ denotes the hourly means for the same hour between days.

Next, IV represents the variability of the activity rhythms throughout the day. In other words, IV indicates the frequency of transitions between rest and activity [32], which is given as

$$IV = \frac{n \sum_{i=2}^{n} (x_i - x_{i-1})^2}{(n-1) \sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{2}$$

TST represents the sum of all the time spent sleeping during the day. WASO is the total waking time during sleep hours. SOL represents the time required to fall asleep after going to bed. Sleep quality is defined as the proportion of real sleep during a detected sleep period.

For TST, WASO and sleep quality, we used the values provided by the Fitbit cloud. For DRO and SOL, which are not provided by the Fitbit cloud, we calculated from the step data as follows. The DRO was measured as the start time of the 5-h period with the least activity within 24 h. SOL, the time it takes a subject to fall asleep, was measured from the start time provided by the Fitbit cloud to the time it took for the steps reach to zero.

The characteristic of data and input features by symptoms in the data set are shown in Table 3. To calculate the features described earlier, only data satisfying the following conditions were used. First, we used subjects whose data were collected for one week or longer. Second, the collected data were used only if there were at least 48 h without any missing data. Third, in the case of missing more than 3 h in a row among 24-h data, all data for that date were excluded.

Table 3. Characteristic of Input Features.

	Total **($N = 352$)**	**None** **($N = 241$)**	**Depression Only** **($N = 15$)**	**Anxiety Only** **($N = 49$)**	**Both** **($N = 47$)**
Demographic					
Ages (years)	72.48 ± 5.9	72.29 ± 5.9	74.33 ± 5.73	72.92 ± 6.04	72.45 ± 6.2
Female (%)	73.01	74.27	46.67	77.56	70.21
Activity rhythm					
IS	0.37 ± 0.13	0.38 ± 0.13	0.33 ± 0.11	0.36 ± 0.12	0.37 ± 0.12
IV	1.19 ± 0.25	1.19 ± 0.25	1.16 ± 0.19	1.20 ± 0.26	1.20 ± 0.26
L5	651.42 ± 95.55	652.63 ± 92.68	612.82 ± 80.30	666.64 ± 107.48	641.51 ± 104.07
Sleep					
TST	396.58 ± 92.35	385.94 ± 86.84	482.02 ± 100.96	408.46 ± 84.05	411.49 ± 96.65
WASO	36.28 ± 15.44	35.19 ± 15.53	42.73 ± 15.29	35.57 ± 12.61	40.52 ± 14.97
SOL	19.26 ± 15.57	18.90 ± 13.56	18.11 ± 10.96	19.39 ± 18.80	21.35 ± 20.27
SQ	90.94 ± 3.92	91.09 ± 3.81	90.71 ± 4.67	91.39 ± 3.08	89.71 ± 4.08

2.5. Minimal Geriatric Assessment Feature Selection

As shown in the previous work by Sim et al. [26], the classification model trained using only activity tracking features from low-cost activity trackers was not sufficiently accurate for use in the field. In this study, we designed a novel way to incorporate questionnaire-based assessment scale data as well as activity tracking data in the development of an accurate model. In this study, we suggest a method to select each set of minimal questionnaire items for depression and anxiety as follows for use with the activity tracking data in our multi-label classifier development.

In addition to the current 352 data set described in Section 2.1, Ewha Womans University Mokdong Hospital had previous data sets related to geriatric assessment scales previously collected. We additionally used those data sets to increase the accuracy and generality when selecting minimal assessment features, as described below. The total number of sample data by symptoms used to derive minimal assessment features is presented in Table 4. Note that the previous data set was used only for the selection of minimal assessment features, not for the multi-label classification described in the next section.

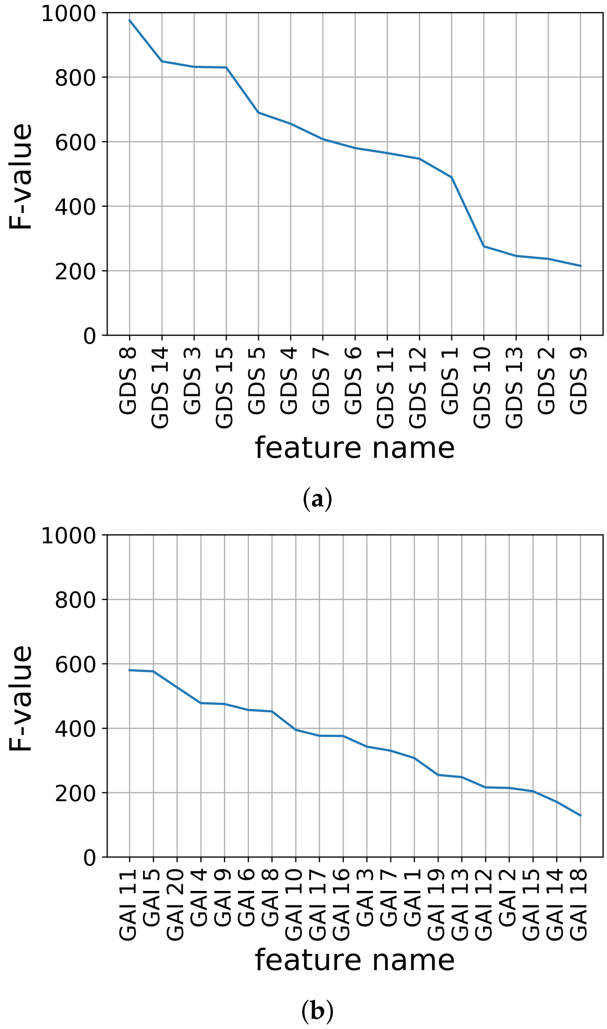
Table 4. Number of samples used to derive minimal assessment features.

Label	**Current Data Set**	**Prev. Data Set**	**Total**
None	241	389	630
Depression only	15	319	334
Anxiety only	49	24	73
Both	47	46	93
Total	352	778	1130

The minimal assessment features for depression and anxiety are derived as follows. First, the importance of the assessment data were calculated using the assessment data in the complete data sets. In this study, the *F-value*, a statistical method that highlights the differences between groups [33,34], was used to calculate the feature importance, as follows:

$$F\text{-}value = \frac{variance\,of\,group\,means(MeanSquareBetween)}{mean\,of\,within\,group\,variances(MeanSquaredError)}. \quad (3)$$

The feature importance of K-SGDS and K-GAI in descending order using the *F-values* are shown in Figure 2a,b, respectively.



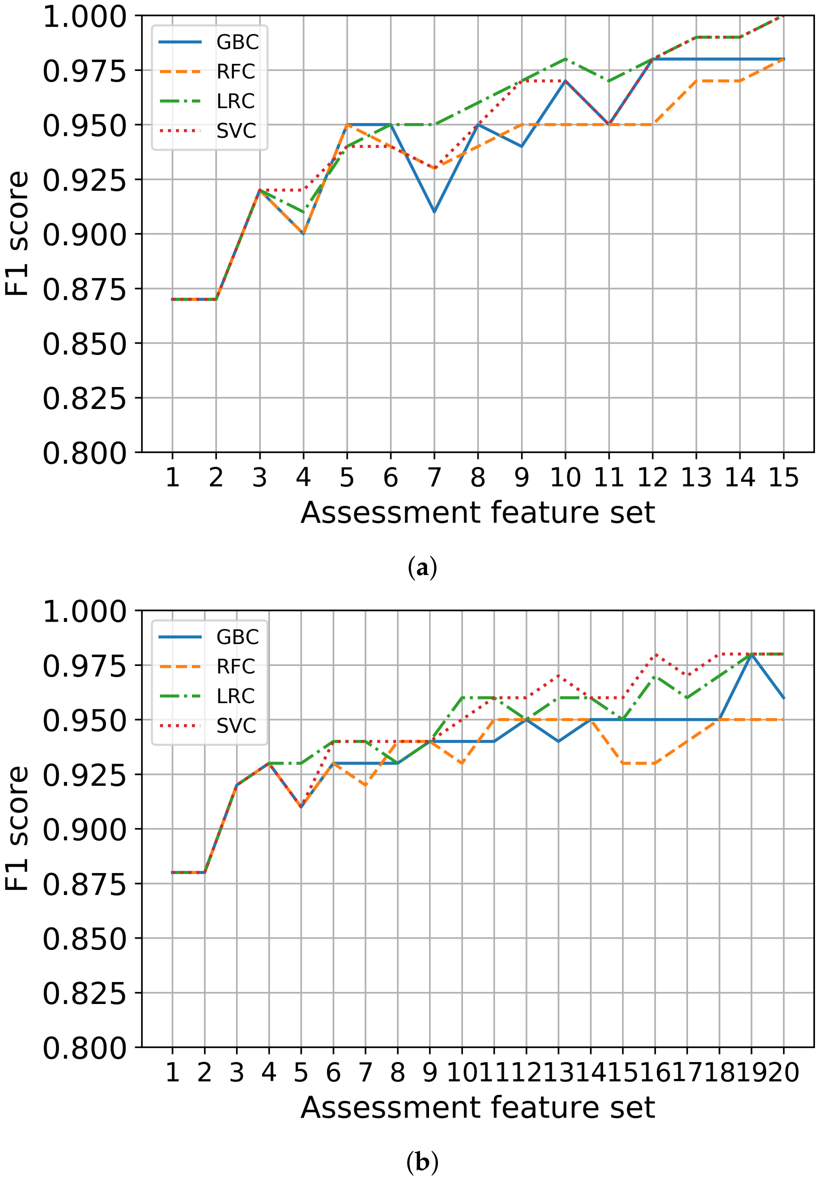
(**a**)

(**b**)

Figure 2. Feature importance of assessment data using *F-value*. (**a**) K-SGDS, (**b**) K-GAI.

Second, the candidate sets of the assessment features for each K-SGDS and K-GAI were constructed by accumulating the feature importance in the order of highest ranking. For example, in K-SGDS, Data Set 1 had GDS 8; Set 2 had GDS 8 and 14; Set 3 had GDS 8, 14 and 3; and in K-GAI, Set 1 had GAI 11; Set 2 had GAI 11 and 5; Set 3 had GAI 11, 5 and 20. Hence, we had 15 and 20 sets of assessment features for depression and anxiety, respectively.

Third, each set in K-SGDS and K-GAI were used to train the binary classifiers of depression and anxiety, respectively. The F_1 scores for the binary classifiers for depression and anxiety using only the assessment features are shown in Figure 3a,b, respectively. We chose feature sets with the F_1 scores of at least 90% and fewer items. Consequently, GDS 8, 14 and 3 out of 15 items in K-SGDS were selected as minimal assessment features for depression, and GAI 11, 5, and 20 out of 20 items of K-GAI were selected as minimal assessment features for anxiety.



(**a**)

(**b**)

Figure 3. Binary classification performance using only assessment feature sets. (**a**) Depression, (**b**) Anxiety.

2.6. Multi-Label Classification

Multi-label classification refers to a scenario in which one data sample can have multiple labels [19–21]. Because each subject may have a combination of depression and anxiety, a multi-label classification is more appropriate for this study. The basic methodology of the multi-label classification in this paper is similar to the previous study by Sim et al. [26].

Label cardinality (LC) and label density (LD) are indicators of the degree to which a data set is multi-label. LC shows the average number of labels per example of the set. LD shows the value divided by the average number of labels per sample. LC is calculated as follows:

$$LC(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}|Y_i| \tag{4}$$

where D is a data set, and Y_i is the label set of the i-th data sample. LD normalizes LC based on the size of the label set, which is given as follows:

$$LD(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i|}{|L|} \tag{5}$$

where L is a label set that can be included in the data. For Data Set 3, LC and LD were 0.46 and 0.23, respectively.

2.6.1. Single-Label Classification for Depression and Anxiety

The target multi-label classification problem was converted into two single-label classification problems. The BR method is adopted to create an independent binary classifier for each label. Thus, we developed classifiers for depression and anxiety, respectively. For each classifier, we applied several classification algorithms to find the one with the highest performance.

For each single-label classifier, we used the following representative algorithms: logistic regression (LR), support vector machine (SVM), random forest (RF) and gradient boosting (GB). LR calculates the probability estimates using a logistic function to classify data into groups with high probabilities [35,36]. LR minimizes the following objective function with respect to parameters, θ:

$$-\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right] \tag{6}$$

where $x^{(i)}$ is the input value, $y^{(i)}$ is the output value, m is the number of data values and h is the cost function. The SVM identifies a boundary for classification by maximizing the margin of the boundary [35,37]. The SVM minimizes the following objective function with respect to parameters, θ:

$$C \sum_{i=1}^{m} \left[y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)})\right] + \frac{1}{2} \sum_{i=1}^{m} \theta_j^2 \tag{7}$$

where C is a hyperparameter, and $cost_n$ is a cost function. RF uses an ensemble of decision trees [38]. It performs the prediction based on the results of multiple decision trees with randomness. It uses a bagging method, i.e., it reuses samples several times to train each model and then aggregates the results. GB is an ensemble algorithm using a boosting method [39]. Boosting is a method for combining weak learners to create strong learners. This algorithm utilizes gradient descent during boosting. It provides a powerful performance but has a disadvantage of requiring long training times.

2.6.2. Performance Metrics

In multi-label classification, there are two methods for evaluating the performance of the classifiers: example-based metrics and label-based metrics [20]. Example-based metrics compare the actual label sets for data samples to the label sets of the prediction results. Label-based metrics calculate the binary evaluation results for each label and then average the results. The evaluated binary classification results are used to generate an evaluation score for all labels.

For a binary classification evaluation, the precision, recall and F_1 scores are calculated [40]. A confusion matrix provides a visual indication of how closely actual data match the predicted results. It also helps us understand how the evaluation indicators are obtained. Table 5 presents a confusion matrix for the binary classification.

Table 5. Confusion matrix for binary classification.

		Predicted:	
		Positive	**Negative**
Actual:	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Accuracy is defined as follows by dividing the number of correctly predicted data by the total number of data, given as

$$Accuracy = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{8}$$

where Y_i is the label set of the i-th data sample, and Z_i is the predicted label set of the i-th data sample. *Precision* represents how much of the actual data are true among the data whose prediction values are true and is defined as

$$Precision = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Z_i|}. \tag{9}$$

Recall represents how much of the predicted data are true among data whose actual values are true, and is defined as

$$Recall = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i|}. \tag{10}$$

The F_1 score represents the harmonic mean of the precision and recall and is defined as

$$F_1 = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|}. \tag{11}$$

To focus on incorrectly predicted labels in multi-label classification, the *Hamming loss* is used and is given as

$$Hamming\text{-}Loss = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \Delta Z_i|}{|L|} \tag{12}$$

where Δ is a symmetric difference between two sets. In this study, we applied micro-averages to the label-based metrics [20]. The micro-averaging is formulated as follows:

$$M_{micro} = M\left(\sum_{\lambda=1}^{|L|} TP_\lambda, \sum_{\lambda=1}^{|L|} FP_\lambda, \sum_{\lambda=1}^{|L|} TN_\lambda, \sum_{\lambda=1}^{|L|} FN_\lambda \right) \tag{13}$$

where λ indicates the elements in the label set.

It is appropriate to measure the performance of a single label classifier based on how accurate it is by measuring the F_1 score, and it is appropriate to determine how low the error is through the Hamming loss for the performance of a multi-label classifier. Therefore, in this paper, we evaluated the performance of single label classifier and multi-label classifier as different indicators [19].

2.6.3. Multi-Label Classifier Training

To develop a multi-label classifier with high accuracy, we individually trained two single-label classifiers for depression and anxiety using the minimal K-SGDS and K-GAI assessment features, respectively, in addition to the activity tracking features. For an objective evaluation of the classification model, we first split the original data to obtain a test set of 20 balanced samples. The test set had five data per symptom, considering a small number of samples for depression, as shown in Data Sets 2 and 3 in Table 4. To train the model with sufficient balanced data per symptom, we resampled using the remaining data to obtain a training set of a total of 352 balanced samples. We used the resampling module in Scikit-learn [41], which adapted bootstrapping with replacement for oversampling and without replacement for undersampling.

We applied several classification algorithms to the binary classifier for each label to find the best performing algorithm with the BR method. We generated a logistic regression classifier (LRC), a support vector machine classifier (SVC), a random forest classifier (RFC) and a gradient boosting classifier (GBC) by applying LR, SVM, RF and GB, respectively, and trained the algorithms using the training set.

During this classifier training process, the optimal hyperparameters for each classification algorithm should be identified because they can affect the classifier performance [42]. We optimized the combination of hyperparameters for each classifier based on F_1 scores using a grid search method in the discrete ranges of hyper-parameter values, as shown in Table 6. Notice that the ranges were given along the Scikit-learn convention. To evaluate classifiers with different hyperparameters, it is necessary to avoid generalization errors through a verification using data that were not used for the model learning. To this end, we adopted a 10-fold stratified cross-validation to avoid data reduction through an additional separation. After optimal binary classifiers were trained for each label, multi-label classifiers were generated from a combination of the binary classifiers. This process and the implementation of classifiers were conducted mainly using Scikit-learn.

Table 6. Hyperparameters for classification using both activity tracking and minimal assessment features.

Label	Classifier	Hyperparameters
Depression	LRC	{'C': [10, 100, 1000]}
	SVC	{'C': [1, 10, 100], 'gamma': [0.01, 0.1, 1], 'kernel': ['rbf', 'linear']}
	RFC	{'n_estimators': [8, 16, 32]}
	GBC	{'n_estimators': [32, 64, 128], 'learning_rate': [0.8, 1.0, 1.2]}
Anxiety	LRC	{'C': [0.1, 1, 10]}
	SVC	{'C': [1, 10, 100], 'gamma': [0.01, 0.1, 1], 'kernel': ['rbf', 'linear']}
	RFC	{'n_estimators': [32, 64, 128]}
	GBC	{'n_estimators': [4, 8, 16], 'learning_rate': [0.4, 0.6, 0.8]}

3. Results

In this section, the model training results of the single-label classifiers for depression and anxiety are shown. To compare the performance, we used the following two feature sets to train the multi-label classifiers: (1) activity tracking data only and (2) activity tracking data and minimal assessment scales.

3.1. Single-Label Classification for Depression

The performance results of the single-label classifier for depression using the training set of (1) only activity tracking features and (2) the activity tracking and minimal K-SGDS assessment features are shown in Tables 7 and 8, respectively. We evaluated multiple classification algorithms using the training set to find the best performing hyperparameters per algorithm. In Table 7, the best performing classifier for depression using only activity tracking features was GBC 69.6%. Since it is practically meaningful if the confidence rate is over 80% [43], only using activity tracking features is not accurate enough for a practical implementation. In Table 8, the best performing classifier for depression using the activity tracking and minimal K-SGDS assessment features was RFC with 96.4%. Most of these algorithms for depression have an F_1 score of 90% or higher, which is extremely accurate. Therefore, it can be justified to use the minimal assessment features in conjunction with the activity tracking features for depression.

Table 7. Performance scores of the depression classifier using only activity tracking features on the training set.

Classifier	Parameters	Accuracy	Precision	Recall	F_1 Score
RFC	{'n_estimators': 32}	0.6847	0.6887	0.6847	0.6823
GBC	{'learning_rate': 0.6, 'n_estimators': 64}	0.6989	0.7072	0.6989	0.6958
SVC	{'C': 1000, 'gamma': 0.01, 'kernel': 'rbf'}	0.6818	0.6969	0.6818	0.6769
LRC	{'C': 0.01}	0.5938	0.5965	0.5938	0.5891

Table 8. Performance scores of the depression classifier using both activity tracking and minimal K-SGDS assessment features on the training set.

Classifier	Parameters	Accuracy	Precision	Recall	F_1 Score
RFC	{'n_estimator': 16}	0.964	0.966	0.964	0.964
GBC	{'learning_rate': 1.0, 'n_estimator': 64}	0.955	0.957	0.955	0.955
SVC	{'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}	0.943	0.945	0.943	0.943
LRC	{'C': 100}	0.892	0.900	0.892	0.891

To further understand the classification characteristic of the best depression classifier (RFC), we investigated the confusion matrix of RFC, as shown in Table 9. According to the table, the FN error of 1 means that the classifier incorrectly classified one in ten people as negative for depression, and the FP errors of 0 means that there are no FP classification.

Table 9. Confusion matrix of RFC for the depression classifier.

		Predicted:	
		Depression	**Non-Depression**
Actual:	Depression	9	1
	Non-depression	0	10

The learning curves of the RFC for the depression classifier are shown in Figure 4. The graph that shows how the error changes in each case as the number of training sets increases is called a learning curve. The training score is the score when testing the model with train data, and the cross-validation score is the score when testing the test set by resampling it to estimate the general performance of the model. The training score for depression was 1.0, and the cross-validation score was about 0.85, while the training score for anxiety was 1.0, and the cross-validation score was about 0.9. There is no small gap between the training and cross-validation score curves, which indicates a high variance. It would be helpful to collect more sample data for depression.

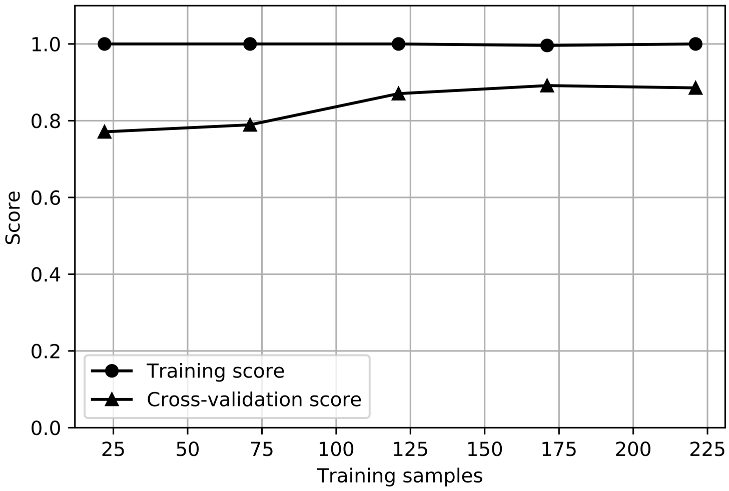


Figure 4. Learning curves of depression classifier.

3.2. Single-Label Classification for Anxiety

The performance results of the single-label classifier for anxiety using the training set of (1) only activity tracking features and (2) the activity tracking and minimal K-GAI assessment features are shown in Tables 10 and 11, respectively. In Table 10, the best performing classifier for anxiety using only activity tracking features was SVC with 59.0%. Thus, only using activity tracking features is not accurate enough for a practical implementation. In Table 11, the best performing classifier for anxiety using the activity tracking and minimal K-GAI assessment features was RFC with 94.6%. Therefore, it can also be justified to use the minimal assessment features instead of the only activity tracking features ones for anxiety as well.

Table 10. Performance scores of the anxiety classifier using only activity tracking features on the training set.

Classifier	Parameters	Accuracy	Precision	Recall	F_1 Score
RFC	{'n_estimators': 32}	0.5824	0.5876	0.5824	0.5767
GBC	{'learning_rate': 1.0, 'n_estimators': 32}	0.5767	0.5772	0.5767	0.5744
SVC	{'C': 1, 'gamma': 1, 'kernel': 'rbf'}	0.5994	0.6103	0.5994	0.5901
LRC	{'C': 0.01}	0.5625	0.5637	0.5625	0.5589

Table 11. Performance scores of the anxiety classifier using both activity tracking and minimal K-GAI assessment features on the training set.

Classifier	Parameters	Accuracy	Precision	Recall	F_1 Score
RFC	{'n_estimator': 64}	0.946	0.948	0.946	0.946
GBC	{'learning_rate': 0.6, 'n_estimator': 8}	0.937	0.940	0.937	0.937
SVC	{'C': 10, 'gamma': 0.1, 'kernel': 'linear'}	0.913	0.918	0.913	0.912
LRC	{'C': 1}	0.910	0.914	0.910	0.909

To further understand the classification characteristic of the best anxiety classifier (RFC), we investigated the confusion matrix of RFC, as shown in Table 12. According to the table, the FN error of 1 means that the classifier incorrectly classified one in ten people as negative for anxiety, and the FP errors of 1 means that the classifier incorrectly classified one in ten people as positive for non-anxiety.

Table 12. Confusion matrix of RFC for the anxiety classifier using activity tracking and minimal GAI assessment features.

		Predicted:	
		Anxiety	**Non-Anxiety**
Actual:	Anxiety	9	1
	Non-anxiety	1	9

The learning curve of RFC for the anxiety classifier is shown in Figure 5. There is a no small gap between the training and cross-validation score curves, which indicates a high variance. It would be helpful to collect more sample data for anxiety.

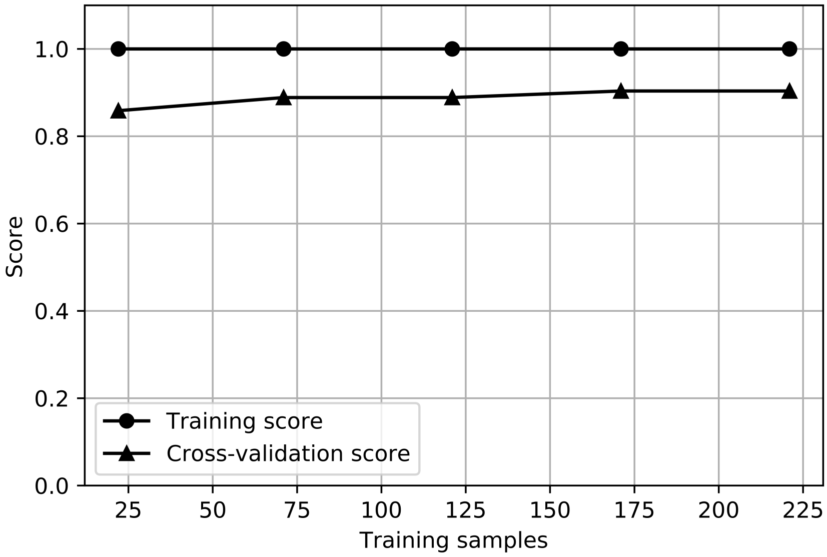


Figure 5. Learning curves of anxiety classifier.

3.3. Multi-Label Classification

Combinations of the classification models for depression and anxiety using both the activity tracking and minimal assessment features were candidates for a multi-label classifier. We evaluated the combinations by calculating the Hamming loss using the training set. A total of 16 combinations were compared using each of the four classifiers of depression and anxiety, and only the top five classifier combinations are shown in this paper. The five combinations are presented in Table 13 in order of lowest Hamming loss. The combinations for depression and anxiety that show the lowest Hamming loss are RFC–RFC and GBC–RFC.

Table 13. Partial list of multi-label performance scores on training set.

Depression	**Anxiety**	**Hamming Loss**
RFC	RFC	0.0000
GBC	RFC	0.0000
RFC	GBC	0.0030
GBC	GBC	0.0030
SVC	RFC	0.0045

To determine the general performance, the test set was used to calculate the Hamming loss of the two combinations of RFC–RFC and GBC–RFC, yielding 0.075 and 0.125, respectively. Hence, we finally selected the combination of RFC for depression and RFC for anxiety as a multi-label classifier.

4. Discussion

In the design of a good multi-label classifier, it is important to consider not only classification performance but also statistical type errors. Since the F_1 score is the average value of depression and anxiety, different combinations of the best algorithms in the single-label classifiers can be selected depending on which symptom is more focused, referencing confusion matrices.

As mentioned in Section 1, there have been studies to reduce the number of effective questionnaires in geriatric assessment scales [7,8]. We showed a novel method using machine learning to systematically find the minimal set of questionnaires for both depression and anxiety. It is a method devised to increase the accuracy of classification models, but since the number of questionnaires can be reduced, it may be helpful for convenient application in the clinical field. Additionally, some K-SGDS and K-GAI questionnaires are still required for practical performance in the field. The best classifiers for depression and anxiety have to be re-selected if the data is changed.

Motivated by the work of Luik et al. [10], we developed an accurate multi-label classification model for identifying geriatric depression and anxiety using both low-cost activity trackers and minimal set of assessment scale data. They used additional input data including electrodermal activity (EDA), location changes, phone-based communication and phone usage patterns. Ghandeharioun et al. [17] and Rykov et al. [18] achieved a successful work in developing a classification model for depression using digital biomarkers such as steps, heart rate, energy expenditure and sleep data from consumer-grade wearable sensors. Their work is limited to finding depression only, and, to the best of authors' knowledge, there have been no other machine learning-based studies to identify the comorbid problem of geriatric depression and anxiety using a low-cost activity tracker. Sim et al. [26] attempted an earlier development of a multi-label classifier using low-cost activity trackers. They showed the potential, but the prediction accuracy needs to be improved to be used in the field. This study extended the work of Sim et al. [26] to improve single-label classification performances from less than 70% to higher than 90% by combining activity tracking and minimal assessment scale features. In addition, in this study, we adapted more appropriate performance metrics, namely the Hamming loss, which penalizes the prediction error to select the best performing model for multi-label classification.

The limitations of this study are given as follows. First, we proposed minimal assessment scale features in addition to activity tracking features to classify geriatric mood disorders with high accuracy. In order to use only activity tracking data, one may need to use activity trackers with much higher resolutions such as ActiGraph at the cost of practicability. Second, we used the BR method without considering inter-label relationships between depression and anxiety. It would also be a good research topic to try different multi-label methods, such as the adapted algorithm, to consider the label relationships. Third, it is well-known that the seasonal changes have been associated with mood symptoms as well as the amount of physical activities and sleep [44]. Although we have not considered the effects of seasons on affective symptoms or physical activities in this study, it would be important to consider the effects of the seasonal variations on the development of symptoms in depression or anxiety. Further studies based on seasonal activity data would be warranted to investigate the effects of seasonal variations on the accuracy of identifying mood symptoms based on the data from activity trackers. Fourth, as can be seen in the learning curves, collecting and using more sample data will help reduce the variance of the model prediction.

5. Conclusions

This study demonstrated the possibility of accurately identifying geriatric depression and anxiety using low-cost activity trackers and minimal geriatric assessment scales. The main contributions of this study can be emphasized as follows. We extracted full activity rhythms and sleep-related features known to be related to geriatric depression and anxiety from activity tracking data collected from older adults using low-cost activity trackers. We designed and validated a novel method that combines the minimal GDS and GAI features, selected from the elderly assessment scale data, with the activity tracking features. We set up a framework for the multi-label classification of geriatric and anxiety, primarily using low-cost activity trackers, which can serve as the basis for diagnosis assistance systems for clinicians. Although the multi-label classifier proposed in this paper shows good classification performance, this classifier can be another useful tool to clinicians that they can use in conjunction with other tools to improve accuracy and reliability in screening geriatric depression and anxiety. The ability to screen for geriatric depression and anxiety using a low-cost wrist-worn activity tracker will provide practical benefits to both physicians and patients. In the case of a family of an elderly patient living alone, the burden of monitoring the patient's condition on a regular basis can be alleviated. Furthermore, we derived the sets of the minimal questionnaires that can still identify geriatric depression and anxiety with high accuracy, which can be effectively used for patient's convenience in related fields.

Author Contributions: Conceptualization, M.-T.C. and G.-H.K.; methodology, M.-T.C. and T.-R.L.; software, T.-R.L. and M.-T.C.; validation, T.-R.L. and M.-T.C.; formal analysis, T.-R.L. and M.-T.C.; data curation, G.-H.K.; writing–original draft preparation, T.-R.L. and M.-T.C.; writing–review and editing, M.-T.C. and G.-H.K. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: The research protocols were approved by the Ewha Women's University Mokdong Hospital Institutional Review Board (EUMC 2016-09-042-013).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by the Technology Innovation Program (10063300 and 20015190) funded by the Korea Ministry of Trade, Industry and Energy.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. World Health Organization. Towards a Dementia Plan: A WHO Guide 2018. Available online: https://www.who.int/publications/i/item/towards-a-dementia-plan-a-who-guide (accessed on 25 February 2022).
2. Santabárbara, J.; Lopez-Anton, R.; De la Cámara, C.; Lobo, E.; Gracia-García, P.; Villagrasa, B.; Bueno-Notivol, J.; Marcos, G.; Lobo, A. Clinically significant anxiety as a risk factor for dementia in the elderly community. *Acta Psychiatr. Scand.* **2019**, *139*, 6–14. [CrossRef] [PubMed]
3. Lenze, E.J.; Mulsant, B.H.; Shear, M.K.; Alexopoulos, G.S.; Frank, E.; Reynolds, C.F. Comorbidity of depression and anxiety disorders in later life. *Depress. Anxiety* **2001**, *14*, 86–93. [CrossRef] [PubMed]
4. Lee, S.C.; Kim, W.H.; Chang, S.M.; Kim, B.S.; Lee, D.W.; Bae, J.N.; Cho, M.J. The use of the Korean version of Short Form Geriatric Depression Scale (SGDS-K) in the community dwelling elderly in Korea. *J. Korean Geriatr. Psychiatry* **2013**, *17*, 37–43.
5. Yesavage, J.A.; Brink, T.L.; Rose, T.L.; Lum, O.; Huang, V.; Adey, M.; Leirer, V.O. Development and validation of a geriatric depression screening scale: A preliminary report. *J. Psychiatr. Res.* **1982**, *17*, 37–49. [CrossRef]
6. Pachana, N.A.; Byrne, G.J.; Siddle, H.; Koloski, N.; Harley, E.; Arnold, E. Development and validation of the Geriatric Anxiety Inventory. *Int. Psychogeriatr.* **2007**, *19*, 103–114. [CrossRef]
7. Hoyl, M.T.; Alessi, C.A.; Harker, J.O.; Josephson, K.R.; Pietruszka, F.M.; Koelfgen, M.; Mervis, J.R.; Fitten, L.J.; Rubenstein, L.Z. Development and testing of a five-item version of the Geriatric Depression Scale. *J. Am. Geriatr. Soc.* **1999**, *47*, 873–878. [CrossRef]
8. Byrne, G.J.; Pachana, N.A. Development and validation of a short form of the Geriatric Anxiety Inventory–the GAI-SF. *Int. Psychogeriatr.* **2011**, *23*, 125–131. [CrossRef]
9. Zee, P.C.; Attarian, H.; Videnovic, A. Circadian rhythm abnormalities. *Contin. Lifelong Learn. Neurol.* **2013**, *19*, 132. [CrossRef]
10. Luik, A.I.; Zuurbier, L.A.; Direk, N.; Hofman, A.; Van Someren, E.J.; Tiemeier, H. 24-h activity rhythm and sleep disturbances in depression and anxiety: A population-based study of middle-aged and older persons. *Depress. Anxiety* **2015**, *32*, 684–692. [CrossRef]
11. ActiGraph, Corp. *Actigraph GT9X Link*. Available online: https://actigraphcorp.com/actigraph-link/ (accessed on 25 February 2022).
12. Fitbit, Inc. *Fitbit alta HR*. Available online: https://www.fitbit.com/gb/shop/altahr (accessed on 25 February 2022).

13. Mendlowicz, M.V.; Jean-Louis, G.; von Gizycki, H.; Zizi, F.; Nunes, J. Actigraphic predictors of depressed mood in a cohort of non-psychiatric adults. *Aust. N. Z. J. Psychiatry* **1999**, *33*, 553–558. [CrossRef]
14. Cook, J.D.; Prairie, M.L.; Plante, D.T. Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: A comparison against polysomnography and wrist-worn actigraphy. *J. Affect. Disord.* **2017**, *217*, 299–305. [CrossRef] [PubMed]
15. Spira, A.P.; Stone, K.; Beaudreau, S.A.; Ancoli-Israel, S.; Yaffe, K. Anxiety symptoms and objectively measured sleep quality in older women. *Am. J. Geriatr. Psychiatry* **2009**, *17*, 136–143. [CrossRef]
16. Li, X.; Zhang, X.; Zhu, J.; Mao, W.; Sun, S.; Wang, Z.; Xia, C.; Hu, B. Depression recognition using machine learning methods with different feature generation strategies. *Artif. Intell. Med.* **2019**, *99*, 101696. [CrossRef] [PubMed]
17. Ghandeharioun, A.; Fedor, S.; Sangermano, L.; Ionescu, D.; Alpert, J.; Dale, C.; Sontag, D.; Picard, R. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017; pp. 325–332.
18. Rykov, Y.; Thach, T.Q.; Bojic, I.; Christopoulos, G.; Car, J. Digital Biomarkers for Depression Screening with Wearable Devices: Cross-sectional Study with Machine Learning Modeling. *JMIR MHealth UHealth* **2021**, *9*, e24872. [CrossRef]
19. Tsoumakas, G.; Katakis, I. Multi-label classification: An overview. *Int. J. Data Warehous. Min. (IJDWM)* **2007**, *3*, 1–13. [CrossRef]
20. Prajapati, P.; Thakkar, A.; Ganatra, A. A survey and current research challenges in multi-label classification methods. *Int. J. Soft Comput. Eng.* **2012**, *2*, 248–252.
21. Tsoumakas, G.; Katakis, I.; Vlahavas, I. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*; Springer: Boston, MA, USA, 2009; pp. 667–685.
22. Serby, M.; Yu, M. Overview: Depression in the elderly. *Mt. Sinai J. Med. N. Y.* **2003**, *70*, 38–44.
23. Zhang, M.L.; Wu, L. Lift: Multi-label learning with label-specific features. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 107–120. [CrossRef]
24. Huang, J.; Li, G.; Huang, Q.; Wu, X. Learning label specific features for multi-label classification. In Proceedings of the 2015 IEEE International Conference on Data Mining, Atlantic City, NJ, USA, 14–17 November 2015; pp. 181–190.
25. Alvares-Cherman, E.; Metz, J.; Monard, M.C. Incorporating label dependency into the binary relevance framework for multi-label classification. *Expert Syst. Appl.* **2012**, *39*, 1647–1655. [CrossRef]
26. Sim, J.K.; Kim, G.H.; Choi, M.T. Binary-Relevance Classification of Depression and Anxiety in the Elderly Using Low-Cost Activity Trackers. *J. Med. Imaging Health Inform.* **2020**, *10*, 1423–1428. [CrossRef]
27. Han, C.; Jo, S.A.; Jo, I.; Kim, E.; Park, M.H.; Kang, Y. An adaptation of the Korean mini-mental state examination (K-MMSE) in elderly Koreans: Demographic influence and population-based norms (the AGE study). *Arch. Gerontol. Geriatr.* **2008**, *47*, 302–310. [CrossRef] [PubMed]
28. Reisberg, B.; Ferris, S.H.; de Leon, M.J.; Crook, T. The Global Deterioration Scale for assessment of primary degenerative dementia. *Am. J. Psychiatry* **1982**, 1136–1139.
29. Kim, J.; Park, M.S.; Oh, D.N. Reliability and validity of Korean geriatric anxiety inventory (K-GAI). *J. Muscle Jt. Health* **2014**, *21*, 75–84. [CrossRef]
30. Fitbit, Inc. *Web API Reference*. Available online: https://dev.fitbit.com/build/reference/web-api/ (accessed on 25 February 2022).
31. Luik, A.I.; Zuurbier, L.A.; Hofman, A.; Van Someren, E.J.; Tiemeier, H. Stability and fragmentation of the activity rhythm across the sleep-wake cycle: The importance of age, lifestyle, and mental health. *Chronobiol. Int.* **2013**, *30*, 1223–1230. [CrossRef]
32. Van Someren, E.J.; Hagebeuk, E.E.; Lijzenga, C.; Scheltens, P.; de Rooij, S.E.; Jonker, C.; Pot, A.M.; Mirmiran, M.; Swaab, D.F. Circadian rest—Activity rhythm disturbances in Alzheimer's disease. *Biol. Psychiatry* **1996**, *40*, 259–270. [CrossRef]
33. Archdeacon, T.J. *Correlation and Regression Analysis: A Historian's Guide*; University of Wisconsin Press: Madison, WI, USA, 1994.
34. David, F.N.; Johnson, N. The effect of non-normality on the power function of the F-test in the analysis of variance. *Biometrika* **1951**, *38*, 43–57. [CrossRef]
35. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006; Volume 128.
36. Yu, H.F.; Huang, F.L.; Lin, C.J. Dual coordinate descent methods for logistic regression and maximum entropy models. *Mach. Learn.* **2011**, *85*, 41–75. [CrossRef]
37. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2011**, *2*, 1–27. [CrossRef]
38. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
39. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]
40. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
41. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
42. Claesen, M.; De Moor, B. Hyperparameter search in machine learning. *arXiv* **2015**, arXiv:1502.02127.
43. Nunnally, J.C. Psychometric theory—25 years ago and now. *Educ. Res.* **1975**, *4*, 7–21.
44. Wirz-Justice, A. Seasonality in affective disorders. *Gen. Comp. Endocrinol.* **2018**, *258*, 244–249. [CrossRef] [PubMed]