Original Article
https://doi.org/10.9758/cpn.24.1261
Clinical Psychopharmacology and Neuroscience 2025;23(3):400-410
pISSN 1738-1088 / eISSN 2093-4327

Integrating Heart Rate Variability Improves Machine Learning-based Prediction of Panic Disorder Symptom Severity

Jin Goo Lee[1,2], Jae-Jin Kim[2,3], Jeong-Ho Seok[2,3], Eunjoo Kim[2,3], Jooyoung Oh[2,3], Chang-Bae Bang[2,3], Byung-Hoon Kim[2,3,4]

[1]Eulji University College of Medicine, Daejeon, Korea
[2]Institute of Behavioral Sciences in Medicine, Yonsei University College of Medicine, Seoul, Korea
[3]Department of Psychiatry, Yonsei University College of Medicine, Seoul, Korea
[4]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Korea

Objective: The association between panic disorder (PD) and heart rate variability (HRV) has long been studied with a focus on the imbalance of the autonomic nervous system. This study aims to demonstrate the predictive capability of HRV in determining PD severity using machine learning.
Methods: Psychometric scales and various HRV components were measured from 507 PD patients who were recruited. We designed three experiments with different sets of input features for comparison. The input features of each experiment were 1) both psychometric scales and HRV together (ExSH), or 2) only the scales (ExS), or 3) only the HRV components. In each experiment, nine machine learning models were used to predict the Panic Disorder Severity Scale. We compared the predictive capability of the three sets of input features by statistically analyzing the performance metrics of the models in the three experiments. SHapley Additive exPlanation (SHAP) was further employed to assess the importance of the input features.
Results: The Random Forest model in ExSH, which incorporated both psychometric scales and HRV, achieved the highest f1-score (76.50%) and sensitivity (75.35%). ExSH showed significantly higher sensitivity and f1-score compared to ExS. For the RF model of ExSH, the highest SHAP importance value was found for the Hamilton Rating Scale for Anxiety, followed by the Hamilton Depression Rating Scale, and the low-frequency power (LF).
Conclusion: Our findings demonstrate that integrating HRV with psychometric scales improves machine learning-based prediction of PD severity. We also highlighted LF as a promising variable among HRV components.

KEY WORDS: Panic disorder; Autonomic nervous system; Heart rate; Machine learning.

INTRODUCTION

Panic disorder (PD) is a common anxiety disorder affecting about 2 to 3% of the entire population [1]. PD is characterized by recurring unexpected panic attacks and persistent concern about experiencing subsequent attacks. A panic attack is defined as a surge of intense fear or discomfort that quickly reaches its peak. It can include various physical or cognitive symptoms such as palpitation, sweating, trembling, chest discomfort, and derealization, which can significantly affect one's daily functioning [2]. The symptoms of panic attacks, such as palpitation or sweating, are known to be related to autonomic nervous system (ANS) dysfunction [3].

Given that heart rate variability (HRV) can measure the level of ANS dysfunction, it has been regarded as one of the potential key indicators of PD in previous studies [4]. In particular, patients with PD exhibit low parasympathetic activity and an imbalance between sympathetic and parasympathetic activities [5]. The neurovisceral integration (NVI) model is a theoretical model that integrates the explanation of these relationships between anxiety, ANS, and HRV [6]. According to the NVI model,

Received: December 10, 2024 / **Revised:** January 31, 2025
Accepted: February 21, 2025 / **Published online:** March 25, 2025
Address for correspondence: Byung-Hoon Kim
Department of Psychiatry and Department of Biomedical Systems Informatics, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea
E-mail: egyptdj@yonsei.ac.kr
ORCID: https://orcid.org/0000-0003-2501-038X

anxiety disorders, including PD, are characterized by a systemic inflexibility grounded in the disinhibition of sympathoexcitatory circuits within the central autonomic network (CAN), leading to a decrease in vagally mediated HRV [6-8]. CAN is an internal regulation system that controls visceromotor, neuroendocrine, and behavioral responses, which are critical for goal-directed behavior and adaptability [9].

This theoretical background of NVI has led to research on the relationship between various components of HRV and PD. Here, the components refer to various measurable quantities related to the variability derived from the raw heart rate recordings, covering 1) time-domain, 2) frequency-domain, and 3) non-linear domain features [10]. The time-domain components quantify the amount of variability in measurements of the inter-beat interval, which is the period in seconds between successive heartbeats. The time-domain components include the mean normal-to-normal interval (MeanNNI), the standard deviation of normal-to-normal interval (SDNN), the square root of the mean squared differences of successive normal-to-normal interval (RMSSD), and the triangular index (TriangularIndex), which represents the integral of the density distribution divided by the maximum of the density distribution [11]. Frequency-domain components estimate the distribution of absolute or relative power between different frequency bands. The frequency-domain components include low-frequency power (LF), high-frequency power (HF), and the ratio of LF-to-HF (LF/HF). Non-linear domain components allow us to quantify the unpredictability of a time series [12]. The non-linear domain components include measures such as the approximate entropy (ApEn) and the sample entropy (SampEn). Some of these components are thought to be related to the existence of PD. For example, the HF can reflect the level of parasympathetic activity, which leads to PD being associated with a lower HF [13]. On the other hand, the LF is thought to reflect sympathetic activity in PD, but some controversies exist [14,15]. The LF/HF is also suggested to be related to PD from several studies [16-18]. Specifically, meta-analysis results indicate that patients with PD show elevated LF/HF when compared to healthy controls [19,20]. The SDNN, one of the time-domain components that represents autonomic influence, was also found to be lower in PD compared to healthy controls [10].

Accurate and reliable measurement of PD symptom severity can be helpful for making clinical decisions in practice, as in determining the treatment response. Conventional assessments make use of a few psychometric scales, such as the Panic Disorder Severity Scale (PDSS) [21], the Hamilton Rating Scale for Anxiety (HAM-A) [22], or the Hamilton Depression Rating Scale (HAM-D) [23]. Although these scales are thoroughly validated and widely used, limitations often arise from the fact that they mostly rely on subjective reporting and judgment. Accordingly, attempts are being made to utilize HRV as an objective biomarker for PD severity. More specifically, research is being conducted on the relationship between the HRV components and PD severity. For example, the MeanNNI exhibited a negative correlation with PDSS [24] and the Beck Anxiety Inventory (BAI) [25,26]. The BAI also showed a negative correlation with the variance of NNI, LF, and HF, while a positive correlation was found with the LF/HF [26]. Moreover, the recent rise of interest in machine learning techniques has led to studies in training a model that can directly predict PD based on HRV components [27,28]. Despite these possibilities and the theoretical background of HRV, research examining the predictive capability of HRV in determining PD severity is lacking.

Therefore, this study aims to demonstrate the predictive capability of HRV in determining PD severity using machine learning. Specifically, we investigate whether 1) HRV data can improve the predictive capability of PD severity when augmented with psychometric scales, and 2) to what extent the predictive capability can be obtained with machine learning models. Utilizing prior research and background knowledge, we designed three experiments with different sets of input features for comparison. The input features of each experiment were 1) both psychometric scales and HRV together, 2) only the scales, or 3) only the HRV components. In each experiment, nine machine learning models are used to classify PD severity into two groups: PDSS scores of five or less, and scores greater than five, which is the cutoff value of PDSS [29]. We compared the predictive capability of the three sets of input features by statistically analyzing the performance metrics (accuracy, sensitivity, positive predictive value, and f1-score) of the models in the three experiments. This stands in contrast to the conventional practice of relying solely on scales or HRV. By doing so, our study not only advances the objective assessment of PD severity but also

lays the groundwork for a broader understanding of the implications of HRV in the evaluation of psychiatric disorders.

METHODS

Participant Recruitment

Subjects aged 20 years or older who were diagnosed with PD by the Korean version of Diagnostic and Statistical Manual of Mental Disorder, fifth edition (DSM-5) were retrospectively recruited from the Psychiatry department inpatient and outpatient clinics of Gangnam Severance Hospital. The recruitment period included from January 2015 to June 2021. Results were analyzed for a total of 507 subjects who satisfied the above criteria and completed all data collection. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures involving human subjects/patients were approved by the Institutional Review Board of Gangnam Severance Hospital, No. 3-2021-0440. Informed consent was waived by the approving ethics committee due to the retrospective nature of the study.

Acquisition of Heart Rate Variability Data

For the measurement of HRV, either SA-6000 (Medicore Co., Ltd.) or QECG-3 (LAXTHA Inc.) devices were used. The participants were seated and rested for 5 minutes before the start of the test. Then, the electrodes were attached, and 3 channels of electrocardiography timeseries signals were collected for five minutes. It is known that short-term measurements of HRV, even as brief as five minutes, are sufficient to yield reliable data [30]. Processing of the raw electrocardiogram timeseries was performed with Python 3 using libraries 'numpy', 'biosppy' and 'hrvanalysis' on a local Linux workstation. The NNI was obtained by applying a set of preprocessing steps to the raw timeseries, which included R-peak extraction, R-R interval calculation, ectopic beat removal, and R-peak interpolation.

Input Feature Selection from Heart Rate Variability Components and Psychometric Scales

We created three sets of input features for machine learning to compare their performance: psychometric scales and HRV (16 features), psychometric scales only (6 features), and HRV only (10 features) (Fig. 1).

Ten components were selected for the HRV input features: MeanNNI, SDNN, pNNI50 (the proportion of NNI50, representing the number of interval differences between successive NNIs greater than 50 ms, to the total number of NNIs), RMSSD, mean heart rate, TriangularIndex, LF, HF, LF/HF, SampEn (Table 1).

We selected these 10 HRV components based on the following anxiety disorders-related studies [24,26-28,31-34]. Among them, the TriangularIndex was chosen due to its robustness to outliers and artifacts [31,32]. Because of the complexity of the physiological systems, it is important to include the nonlinear dynamics of HRV in the considerations [33]. Within the nonlinear domain of HRV

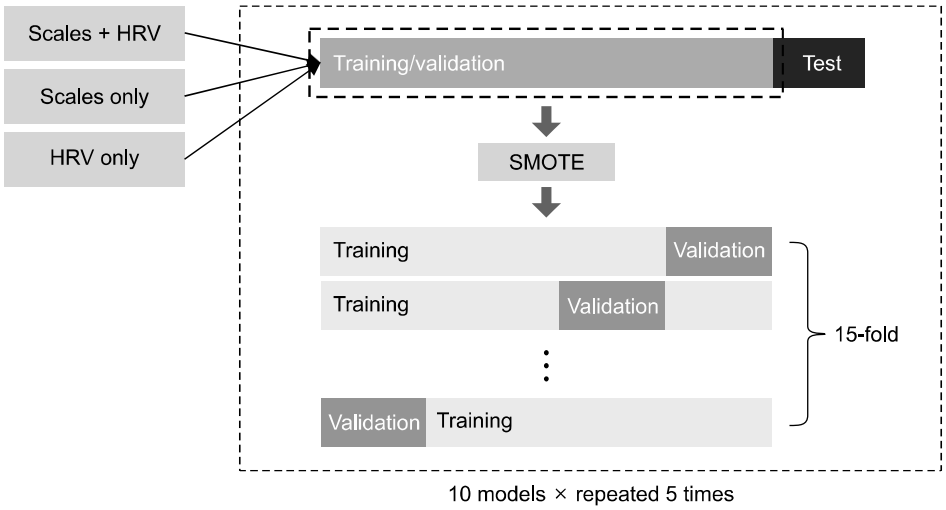

Fig. 1. Machine learning pipeline overview.
For comparison, three separate experiments were conducted, each utilizing one of the three datasets: scales with HRV, scales only, and HRV only. SMOTE was applied to address class imbalance. There are ten models in total, including nine machine learning models and one dummy model. Each experiment was repeated five times.
HRV, heart rate variability; SMOTE, Synthetic Minority Over-sampling Technique.

Table 1. Full list of input features for the classification models

Index	Feature	Abbreviation	Type
1	Hamilton Rating Scale for Anxiety	HAM-A	Psychometric Scales
2	State-Trait Anxiety Inventory-State anxiety	STAI-S	
3	State-Trait Anxiety Inventory-Trait anxiety	STAI-T	
4	Perceived Stress Scale (10-item inventory)	PSS	
5	Hamilton Depression Rating Scale	HAM-D	
6	Korean version of Inventory for Depressive Symptomatology	KIDS-SR	
7	Mean normal-to-normal interval (NNI)	MeanNNI	Heart rate variability components
8	Standard deviation of the NNI	SDNN	
9	The proportion of NNI50, representing the number of interval differences between successive NNIs greater than 50 ms, to the total number of NNIs	pNNI50	
10	Square root of the mean squared differences of successive NNI	RMSSD	
11	Mean heart rate	MeanHr	
12	Low frequency	LF	
13	High frequency	HF	
14	Low frequency/high frequency ratio	LF/HF	
15	Triangular index	TriangularIndex	
16	Sample entropy	SampEn	

analysis, SampEn was preferred over ApEn for its greater theoretical accuracy, consistent measurements, reduced bias in short datasets, more reliable assessment of synchrony in clinical time series, and improved calculation methods [34].

Six psychometric scales were selected for the scale input features: HAM-A, State-Trait Anxiety Inventory-State anxiety (STAI-S), State-Trait Anxiety Inventory-Trait anxiety (STAI-T), HAM-D, Korean version of Inventory for Depressive Symptomatology (KIDS-SR), and Perceived Stress Scale (PSS) (Table 1). The reasons for this selection are as follows. Patients with PD often experience one or more comorbid lifetime psychiatric disorders [35]. The same study found that Major depressive disorder and other anxiety disorders were the most common comorbidities with PD. In a comprehensive study of 9,282 individuals, patients with only PD had an odds ratio of 2.0 to 5.4 for other anxiety disorders and major depression, while patients with both PD and agoraphobia had odds ratios of 2.5 to 25.8 [36]. Considering this characteristic of comorbidity with depression and anxiety disorders, we included the HAM-A and HAM-D, which can be part of the PD severity evaluation criteria, and the KIDS-SR, a reliable and valid self-report measure for assessing depression in Korea, considering that this study targets Koreans [37]. Furthermore, we included the STAI-S and STAI-T, which measure state and trait anxiety [38]. We also included the PSS, which measures perceived self-regulation regarding stress. PSS may be valuable within a clinical setting by facilitating treatment planning and assessing treatment response [39]. We used the PSS 10-item version [40].

To determine if there were any significant differences in input features between the two groups (PDSS scores of five or less, and scores greater than five), we used the *t* test.

Machine Learning

Three experiments were defined based on the three sets of input features (Fig. 1). Experiment with Scales and HRV (ExSH), Experiment with Scales (ExS), and Experiment with HRV (ExH) utilize both scale and HRV features, scale features, and HRV features as their input features, respectively. In all three experiments (ExSH, ExS, ExH), nine machine learning models are trained for solving the same classification problem: classifying between the two groups (PDSS scores of five or less, and scores greater than five, which is the cutoff value of PDSS) [29].

The machine learning models utilized in this study comprised of Logistic Regression (LoR), Support Vector Machine (SVM), Decision Tree (DT), Bagging Decision Tree (BDT), Adaptive Boosting Decision Tree (ABDT), Random Forest Classifier (RF), Multi-layer Perceptron Classifier (MLP), Extreme Gradient Boosting Classifier (XGB), and Cat Boosting Classifier (CB). A baseline model is the DummyClassifier (Dummy). The Dummy, set with

the 'stratified' strategy, provides a baseline by reflecting the class distribution of the training dataset in its predictions.

The dataset was divided into training and testing splits with an 8:2 ratio, deriving 407 and 100 samples within each split, respectively. The distribution of the PDSS scores in our dataset has an imbalance between target classes, which included 156 cases with PDSS scores of five or less, and 351 cases with scores above five. To mitigate the potential issue related to the classifier being biased towards the majority class in the imbalanced dataset [41], we implemented the Synthetic Minority Oversampling Technique for adjusting the imbalanced class distribution [42]. The machine learning pipeline is schematically represented in Figure 1.

To optimize the hyperparameters for each model, a 15-fold cross-validation Grid search was employed. The full hyperparameter search space is provided in the Supplementary Table 1 (available online). The selected hyperparameters were those that yielded the best average validation f1-score across the folds. The importance of each input feature was assessed by SHapley Additive exPlanation (SHAP) [43]. SHAP, derived from ideas in game theory, can be a robust technique for generating individual explanations by guaranteeing a fair distribution of the effect among the features [44].

The predictive performance was evaluated by the following quantitative metrics for each machine learning model, including accuracy (ACC), sensitivity (SEN), positive predictive value (PPV), and f1-score (F1). To mitigate the unwanted stochasticity affecting our results, we conducted the same experiment five times with varying seeds and averaged from these five iterations. All implementations of machine learning experiments were done using Python 3 with the 'pandas', 'sklearn', 'imblearn', and 'shap' packages [43,45-47].

Statistical Analysis

To compare the performance of the three experiments (ExSH, ExS, and ExH), we conducted Friedman tests, each followed by post-hoc analyses using the Wilcoxon test. Each of the three experiments, including nine machine learning models, was compared based on four performance metrics (ACC, SEN, PPV, and F1) derived from the machine learning models. For example, when comparing ACC, each of the three groups (ExSH, ExS, and ExH) has ACC values for each of the nine machine learning models. We confirmed differences among the three groups in ACC using the Friedman test and conducted post-hoc analysis using the Wilcoxon test to determine the superiority among the three groups. The same process was carried out for the remaining four metrics. To address the issue of multiple comparisons and control the false discovery rate, we applied the Benjamini-Hochberg method.

RESULTS

Demographic Characteristics

A total of 507 participants participated in this study. The mean age ± standard deviation (SD) of the participants was 36.78 ± 16.07 years. Among the 507 subjects, 293 (57.79%) were women. Regarding the input features for psychiatric symptoms, the HAM-D, KIDS-SR, STAI-S, STAI-T, PSS, and HAM-A had mean ± SD values of 16.60 ± 6.92, 16.82 ± 7.83, 56.39 ± 12.61, 54.72 ± 12.11, 21.06 ± 6.41, and 23.87 ± 10.22, respectively.

Predictive Performance of Machine Learning Models

All performance metric results are compiled in Supplementary Table 2 (available online). Among all the experiments, the RF model in ExSH, which incorporated both psychometric scales and HRV, achieved the highest F1 (76.50%). This model also showed the best performance in SEN (75.35%). The CB model in ExSH achieved the second-highest F1 (76.45%) by a slight margin and achieved the highest performance in ACC (67.84%). The LoR model in ExS achieved the highest performance in PPV (81.68%).

For all models in ExSH and ExS, each performance metric (ACC, SEN, PPV, and F1) exceeded those of the baseline dummy model. However, in ExH, which only incorporated HRV components, the performance improvement compared to Dummy was less pronounced than in ExSH and ExS. Notably, some models in ExH (MLP and DT) underperformed relative to Dummy in certain metrics (SEN and F1).

Inter-experimental Comparisons

Among the nine machine learning models, the median values of all metrics in ExSH showed an increase compared to ExS: ACC (from 60.98 to 63.73%), SEN (from 62.92 to 67.75%), PPV (from 77.56 to 78.49%), and F1

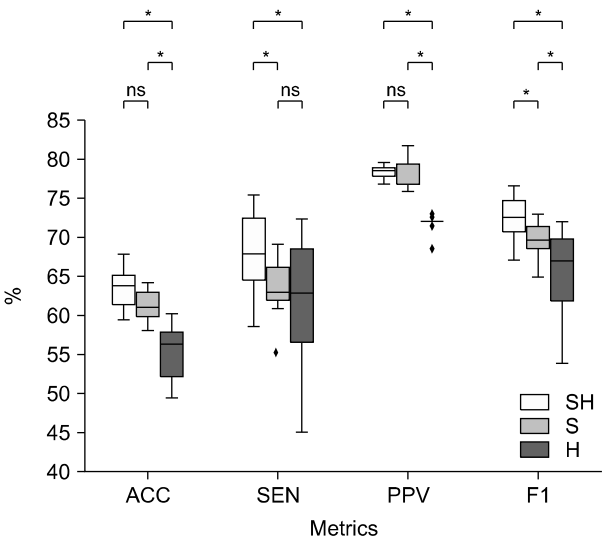

Fig. 2. Box plot of the distribution of performance metrics for each model across the three experiments, excluding the dummy model.
Statistical analysis was performed using Wilcoxon test with Benjamini-Hochberg method.
SH, Experiment SH; S, Experiment S; H, Experiment H; ACC, accuracy; SEN, sensitivity; PPV, positive predictive value; F1, f1-score; ns, not significant.
*$p < 0.05$.

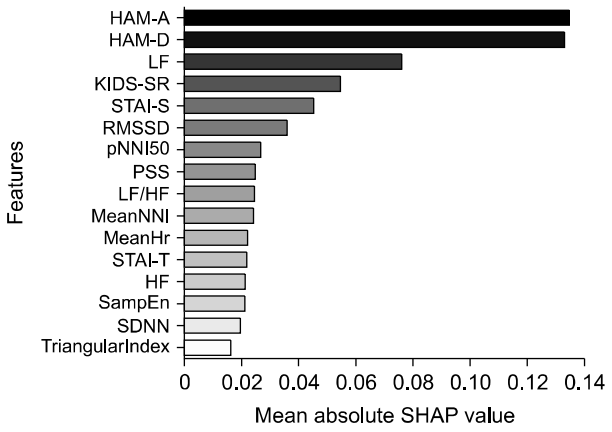

Fig. 3. Mean absolute SHAP values of Random Forest Classifier for Experiment with Scales and heart rate variability.
The final mean absolute SHAP values were calculated by averaging the mean absolute SHAP values across the five repetitions of the experiment.
SHAP, SHapley Additive exPlanation; HAM-A, Hamilton Rating Scale for Anxiety; HAM-D, Hamilton Depression Rating Scale; LF, low frequency; KIDS-SR, Korean version of Inventory for Depressive Symptomatology; STAI-S, State-Trait Anxiety Inventory-State anxiety; RMSSD, square root of the mean squared differences of successive NNI; pNNI50, the proportion of NNI50; PSS, Perceived Stress Scale (10-item inventory); LF/HF, low frequency/high frequency ratio; MeanNNI, mean NNI; MeanHr, mean heart rate; STAI-T, State-Trait Anxiety Inventory-Trait anxiety; HF, high frequency; SampEn, sample entropy; SDNN, standard deviation of the NNI; TriangularIndex, triangular index; NNI, normal-to-normal intervals.

(from 69.58 to 72.51%), as depicted in the box plots (Fig. 2).

All the performance metrics used (p: ACC, PPV $<$ 0.001; SEN, F1 $<$ 0.01) showed significant differences among the three experiments according to the Friedman test (Supplementary Table 3; available online). SEN (p = 0.020) and F1 (p = 0.049) were significantly different between ExSH and ExS. All four metrics (p = 0.020) were significantly different between ExSH and ExH. Between ExS and ExH, all metrics (p = 0.020) except SEN were significantly different. Statistical significances of the differences between all three pairs of experiments are shown together with the box plots in Figure 2.

Interpretation of Input Features

We demonstrate the importance of input features using mean absolute SHAP values of RF for ExSH, which showed the best performance (i.e. f1-score) in this study (Fig. 3). Other results on mean absolute SHAP values are referred to Supplementary Figure 1 (available online). In ExSH's RF, the order of mean absolute SHAP values was HAM-A, HAM-D, LF, and KIDS-SR.

In the t test, all psychometric scales showed significant differences (p: HAM-A, HAM-D, KIDS-SR, STAI-S $<$ 0.0001; STAI-T $<$ 0.001; PSS $<$ 0.01) between the two groups. None of the HRV components showed significant differences in the t test.

DISCUSSION

In this study, we demonstrated that the integration of HRV components with psychometric scales as input features for a machine learning classifier predicting PD severity shows higher sensitivity and f1-score compared to using psychometric scales only. For sensitivity, a maximum increase of 11.57% was seen in SVM, with an average (± SD) increase of 4.53% (± 3.64). For f1-score, a maximum increase of 6.96% was seen in SVM, with an average (± SD) increase of 2.75% (± 2.35). This implies that integrating HRV components with psychometric scales could be beneficial in improving the predictive capability of PD severity. However, relying solely on HRV components showed less effective performance compared to psychometric scales in predicting PD severity, thereby also highlighting the limitations of HRV in this

context. On the other hand, a machine learning classifier predicting PD severity showed the following maximal predictive capabilities: accuracy (67.84%), sensitivity (75.35%), PPV (81.68%), and f1-score (76.50%). Among these metrics, accuracy, sensitivity, and f1-score achieved their maximal performance in ExSH (when both HRV and psychometric scales were considered as input features).

Assessing PD severity holds significant clinical implications, as it can be related to decisions concerning type or duration of the treatment [48]. Therefore, in our study, to increase objectivity and accuracy in this evaluation, we utilized HRV, which has been studied as a 'biomarker' in the field of psychiatry, and indeed observed performance improvement [4,49-51]. Furthermore, our study employs a multimodal prediction, considering both HRV and psychometric scales, which are distinct data types representing the biological and psychological aspects of the patient. This approach is in line with the recent trend of analyzing multimodal data, such as HRV, psychometric scales, natural language, or neuroimaging, to understand psychiatric disorders, including anxiety disorders [27,52-54]. Our study specifically holds its significance as it is the first to explore the integration of HRV with psychometric scales and show that HRV can be a significant feature for improving the performance of the severity prediction of PD. We expect that our results suggest the potential of integrating other features for machine learning-based prediction of anxiety disorders [55].

In the *t* test, all six scales showed significant differences between the two groups, which are PDSS scores of five or less, and scores greater than five. The higher scores on anxiety disorder-related scales (e.g. HAM-A) and depressive disorder-related scales (e.g. HAM-D) in the higher PD severity group may be explained by the high comorbidity of PD with other anxiety disorders and depressive disorder [35].

Although some HRV components (e.g., LF, LF/HF) have been controversial, various components (e.g., RMSSD, SDNN, HF, LF, and LF/HF) have been suggested to be associated with PD in previous studies [4,19,20,26,27,30,56-58]. However, no significant differences between the two groups were found for any HRV components in our study. While previous studies mainly compared the HRV components of PD patients with healthy controls, our study focuses on distinguishing the severity of PD and may not present a significant difference between the two groups from the statistical test. However, it is clear from our study that there was an improvement in machine learning performance when integrating HRV components. This suggests that variables that were not significant in conventional statistical analysis may have the potential to significantly improve performance when integrated into machine learning-based predictions. For a similar reason, Yoo *et al.* [54] leveraged HRV components that were insignificant in their *t* test by using deep neural networks to predict the severity of anxiety disorders.

To identify which variables played crucial roles in machine learning-based predictions, we employed the SHAP method, which is known to be a robust method for model explanation [43,44]. The mean absolute SHAP value, serving as a measure of feature importance, corresponds to the relative importance ranking of features [59]. In our study, LF exhibited the highest mean absolute SHAP value among HRV components in the model with the highest f1-score (RF in ExSH). While many studies have explored the relationship between PD and LF, shedding light on parasympathetic and sympathetic activities [26,27,56-58], our research further supports this association. HAM-A and HAM-D had higher mean absolute SHAP values than LF in the same model. This, together with the fact that ExS performed significantly better than ExH, may indicate that HRV alone is less predictive of PD severity than psychometric scales.

RF showed the best f1-score for ExSH, and CB for ExS and ExH. Ensemble-based classifiers (i.e. BDT, ABDT, RF, XGB, and CB), which are a set of classifiers whose individual decisions are combined in some way [60], have the potential to outperform their individual base classifier (i.e. DT) [61,62]. This may explain why RF, CB, and some other ensemble models outperformed DT in our study. However, this does not mean that ensemble-based classifiers, including RF and CB, are the best models for predicting PD severity. Determining which model is best at making disease-related predictions is complex [61]. For example, MLP and LoR outperformed ensemble models in machine-learning-based studies related to PD, although comparisons are difficult due to the heterogeneity of the input features and the dependent variable being predicted [27,28].

This study, while providing valuable insights, does have certain limitations that should be considered. First, we selected the HRV components as input features based

on our literature review. This feature selection process has the limitation of introducing some bias. However, we tried to overcome the curse of dimensionality by selecting features that utilize domain knowledge, given the lack of machine learning-based research on predicting PD severity. Secondly, our study has a low maximum accuracy. Although hyperparameter tuning based on accuracy could have yielded better accuracy, we chose f1-score because it reflects both sensitivity and positive predictive value, which are in a trade-off relationship, and is more suitable for imbalanced datasets [63]. Based on previous literature, we selected the HRV components to use in our study, but there may still be some bias in this approach. We expect that in the future, filter methods of feature selection like minimum Redundancy Maximum Relevance selection algorithm [64] or feature extraction techniques like Principal Component Analysis [65] may improve the accuracy [66]. Additionally, our study involved 507 subjects, which may pose technical constraints for applying machine learning, and the 5-minute HRV measurement time may not have fully reflected the patients' actual states. These aspects also leave room for future improvements. Predicting PD severity directly is a first attempt and is a more difficult task than predicting the presence of PD. However, in studies targeting only patients with PD, panic-related distress and the duration of PD were related to HF [67]. These findings suggest that HRV has potential for classification within PD and warrant further research. Thirdly, due to the retrospective design, we were unable to account for coexisting diseases and the use of medication in PD patients. Considering the non-specific nature of HRV in mental illnesses, these two factors could influence the results. Conducting a prospective study that incorporates these variables would allow for more precise outcomes. Fourthly, we performed a binary classification, but in a real clinical situation, a more fine-grained classification or regression may be more appropriate. Making such predictions is usually a more difficult task than binary classification.

In conclusion, our study holds significance in exploring the potential of HRV and machine learning in predicting PD severity. With over 500 samples, we extensively investigated the effect of HRV integration across nine different machine learning models. Our findings demonstrate that integrating HRV with psychometric scales yields higher predictive capability compared to considering psychometric scales alone. This suggests the potential of multimodal approaches in machine learning research within the field of psychiatry. Furthermore, our study explored the extent to which predictive performance can be achieved in predicting PD severity using machine learning and, through the reliable methodology of SHAP, highlighted LF as a promising variable among HRV components. To further enhance prediction performance and achieve accurate and objective evaluations in clinical settings, future research in psychiatry should continue to explore multimodal machine learning studies, including HRV and other modalities.

■ Funding

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2022R1I1A1A01069589). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2021M3E5D9025019).

■ Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

■ Author Contributions

Conceptualization: Jin Goo Lee, Byung-Hoon Kim. Data acquisition: Chang-Bae Bang, Jae-Jin Kim, Jeong-Ho Seok, Eunjoo Kim, Jooyoung Oh, Byung-Hoon Kim. Data processing: Jin Goo Lee, Chang-Bae Bang, Byung-Hoon Kim. Formal analysis: Jin Goo Lee. Supervision: Jae-Jin Kim, Byung-Hoon Kim. Writing—original draft: Jin Goo Lee. Writing—review & editing: Byung-Hoon Kim.

■ ORCID

Jin Goo Lee https://orcid.org/0009-0004-2528-3630
Jae-Jin Kim https://orcid.org/0000-0002-1395-4562
Jeong-Ho Seok https://orcid.org/0000-0002-9402-7591
Eunjoo Kim https://orcid.org/0000-0003-3061-2051
Jooyoung Oh https://orcid.org/0000-0001-6721-399X
Chang-Bae Bang https://orcid.org/0000-0001-6244-7666
Byung-Hoon Kim https://orcid.org/0000-0003-2501-038X

REFERENCES

1. Kessler RC, Chiu WT, Demler O, Merikangas KR, Walters EE.

Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005;62:617-627.
2. American Psychiatric Association, DSM-5 Task Force. *Diagnostic and statistical manual of mental disorders. 5th ed. American Psychiatric Publishing, Inc.;2013.*
3. Boland R, Verduin M, Ruiz P. *Kaplan & Sadock's synopsis of psychiatry. 12th ed. Lippincott Williams & Wilkins;2021.*
4. Chalmers JA, Quintana DS, Abbott MJ, Kemp AH. *Anxiety disorders are associated with reduced heart rate variability: a meta-analysis. Front Psychiatry 2014;5:80.*
5. Yeragani VK, Pohl R, Berger R, Balon R, Ramesh C, Glitz D, *et al. Decreased heart rate variability in panic disorder patients: a study of power-spectral analysis of heart rate. Psychiatry Res 1993;46:89-103.*
6. Thayer JF, Lane RD. *A model of neurovisceral integration in emotion regulation and dysregulation. J Affect Disord 2000;61:201-216.*
7. Friedman BH. *An autonomic flexibility-neurovisceral integration model of anxiety and cardiac vagal tone. Biol Psychol 2007;74:185-199.*
8. Thayer JF, Lane RD. *Claude Bernard and the heart-brain connection: further elaboration of a model of neurovisceral integration. Neurosci Biobehav Rev 2009;33:81-88.*
9. Benarroch EE. *The central autonomic network: functional organization, dysfunction, and perspective. Mayo Clin Proc 1993;68:988-1001.*
10. Shaffer F, Ginsberg JP. *An overview of heart rate variability metrics and norms. Front Public Health 2017;5:258.*
11. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Heart rate variability: standards of measurement, physiological interpretation and clinical use. Circulation 1996;93:1043-1065.*
12. Stein PK, Reddy A. *Non-linear heart rate variability and risk stratification in cardiovascular disease. Indian Pacing Electrophysiol J 2005;5:210-220.*
13. Pittig A, Arch JJ, Lam CW, Craske MG. *Heart rate and heart rate variability in panic, social anxiety, obsessive-compulsive, and generalized anxiety disorders at baseline and in response to relaxation and hyperventilation. Int J Psychophysiol 2013;87:19-27.*
14. Billman GE. *Heart rate variability - a historical perspective. Front Physiol 2011;2:86.*
15. Reyes del Paso GA, Langewitz W, Mulder LJ, van Roon A, Duschek S. *The utility of low frequency heart rate variability as an index of sympathetic cardiac tone: a review with emphasis on a reanalysis of previous studies. Psychophysiology 2013;50:477-487.*
16. Shaffer F, McCraty R, Zerr CL. *A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. Front Psychol 2014;5:1040.*
17. Heathers JA. *Everything Hertz: methodological issues in short-term frequency-domain HRV. Front Physiol 2014;5:177.*
18. Billman GE. *The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance. Front Physiol 2013;4:26.*
19. Zhang Y, Zhou B, Qiu J, Zhang L, Zou Z. *Heart rate variability changes in patients with panic disorder. J Affect Disord 2020;267:297-306.*
20. Wang Z, Luo Y, Zhang Y, Chen L, Zou Y, Xiao J, *et al. Heart rate variability in generalized anxiety disorder, major depressive disorder and panic disorder: a network meta-analysis and systematic review. J Affect Disord 2023;330:259-266.*
21. Shear MK, Brown TA, Barlow DH, Money R, Sholomskas DE, Woods SW, *et al. Multicenter collaborative panic disorder severity scale. Am J Psychiatry 1997;154:1571-1575.*
22. Hamilton M. *The assessment of anxiety states by rating. Br J Med Psychol 1959;32:50-55.*
23. Hamilton M. *A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23:56-62.*
24. Garakani A, Martinez JM, Aaronson CJ, Voustianiouk A, Kaufmann H, Gorman JM. *Effect of medication and psychotherapy on heart rate variability in panic disorder. Depress Anxiety 2009;26:251-258.*
25. Beck AT, Epstein N, Brown G, Steer RA. *An inventory for measuring clinical anxiety: psychometric properties. J Consult Clin Psychol 1988;56:893-897.*
26. Chang HA, Chang CC, Tzeng NS, Kuo TB, Lu RB, Huang SY. *Decreased cardiac vagal control in drug-naive patients with panic disorder: a case-control study in Taiwan. Asia Pac Psychiatry 2013;5:80-89.*
27. Jang EH, Choi KW, Kim AY, Yu HY, Jeon HJ, Byun S. *Automated detection of panic disorder based on multimodal physiological signals using machine learning. ETRI J 2023;45:105-118.*
28. Na KS, Cho SE, Cho SJ. *Machine learning-based discrimination of panic disorder from other anxiety disorders. J Affect Disord 2021;278:1-4.*
29. Furukawa TA, Katherine Shear M, Barlow DH, Gorman JM, Woods SW, Money R, *et al. Evidence-based guidelines for interpretation of the Panic Disorder Severity Scale. Depress Anxiety 2009;26:922-929.*
30. Cheng YC, Su MI, Liu CW, Huang YC, Huang WL. *Heart rate variability in patients with anxiety disorders: a systematic review and meta-analysis. Psychiatry Clin Neurosci 2022;76:292-302.*
31. Vollmer M. *A robust, simple and reliable measure of heart rate variability using relative RR intervals. Comput Cardiol (2010) 2015;42:609-612.*
32. Malik M, Xia R, Odemuyiwa O, Staunton A, Poloniecki J, Camm AJ. *Influence of the recognition artefact in automatic analysis of long-term electrocardiograms on time-domain measurement of heart rate variability. Med Biol Eng Comput 1993;31:539-544.*
33. Dimitriev DA, Saperova EV, Dimitriev AD. *State anxiety and*

nonlinear dynamics of heart rate variability in students. PLoS One 2016;11:e0146131.
34. Richman JS, Moorman JR. *Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol 2000;278:H2039-H2049.*
35. Tilli V, Suominen K, Karlsson H. *Panic disorder in primary care: comorbid psychiatric disorders and their persistence. Scand J Prim Health Care 2012;30:247-253.*
36. Kessler RC, Chiu WT, Jin R, Ruscio AM, Shear K, Walters EE. *The epidemiology of panic attacks, panic disorder, and agoraphobia in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2006;63:415-424.*
37. Yoon JH, Jon DI, Hong HJ, Hong N, Seok JH. *Reliability and validity of the Korean version of Inventory for Depressive Symptomatology. Mood Emot 2012;10:131-151.*
38. Spielberger CD, Gonzalez-Reigosa F, Martinez-Urrutia A, Natalicio LF, Natalicio DS. *The State-Trait Anxiety Inventory. Interam J Psychol 1971;5:145-158.*
39. Roberti JW, Harrington LN, Storch EA. *Further psychometric support for the 10-item version of the Perceived Stress Scale. J Coll Couns 2006;9:135-147.*
40. Cohen S. *Perceived stress in a probability sample of the United States. In: Spacapan S, Oskamp S, editors. The social psychology of health. Sage Publications, Inc.;1988.*
41. Kaur H, Pannu HS, Malhi AK. *A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput Surv 2019;52:1-36.*
42. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. *SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357.*
43. Lundberg SM, Lee SI. *A unified approach to interpreting model predictions. In: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017); Dec 4-9, 2017; Long Beach, CA, USA.*
44. ElShawi R, Sherif Y, Al Mallah M, Sakr S. *Interpretability in healthcare: a comparative study of local machine learning interpretability techniques. Comput Intell 2021;37:1633-1650.*
45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830.*
46. Lemaître G, Nogueira F, Aridas CK. *Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 2017;18:1-5.*
47. McKinney W. *pandas: a foundational Python library for data analysis and statistics [Internet]. Python for high performance and scientific computing; 2011 Nov 18 [cited at 2023 Nov 1]. Available from: https://scholar.google.com/scholar?oi=bibs&cluster=3748557446307443113&btnI=1&hl=pt-BR*
48. Zimmerman M, Morgan TA, Stanton K. *The severity of psychiatric disorders. World Psychiatry 2018;17:258-275.*
49. Strimbu K, Tavel JA. *What are biomarkers? Curr Opin HIV AIDS 2010;5:463-466.*
50. Sheridan DC, Baker S, Dehart R, Lin A, Hansen M, Tereshchenko LG, *et al. Heart rate variability and its ability to detect worsening suicidality in adolescents: a pilot trial of wearable technology. Psychiatry Investig 2021;18:928-935.*
51. Chang HA, Chang CC, Tzeng NS, Kuo TB, Lu RB, Huang SY. *Generalized anxiety disorder, comorbid major depression and heart rate variability: a case-control study in taiwan. Psychiatry Investig 2013;10:326-335.*
52. Lee DY, Kim N, Park C, Gan S, Son SJ, Park RW, *et al. Explainable multimodal prediction of treatment-resistance in patients with depression leveraging brain morphometry and natural language processing. Psychiatry Res 2024;334:115817.*
53. Poirot MG, Ruhe HG, Mutsaerts HMM, Maximov II, Groote IR, Bjørnerud A, *et al. Treatment response prediction in major depressive disorder using multimodal MRI and clinical data: secondary analysis of a randomized clinical trial. Am J Psychiatry 2024;181:223-233.*
54. Yoo JH, Jeong H, An JH, Chung TM. *Mood disorder severity and subtype classification using multimodal deep neural network models. Sensors (Basel) 2024;24:715.*
55. Schwarz E. *Advancing psychiatric biomarker discovery through multimodal machine learning. Biol Psychiatry 2022;91:524-525.*
56. Cohen H, Benjamin J, Geva AB, Matar MA, Kaplan Z, Kotler M. *Autonomic dysregulation in panic disorder and in post-traumatic stress disorder: application of power spectrum analysis of heart rate variability at rest and in response to recollection of trauma or panic attacks. Psychiatry Res 2000;96:1-13.*
57. Kang EH, Lee IS, Park JE, Kim KJ, Yu BH. *Platelet serotonin transporter function and heart rate variability in patients with panic disorder. J Korean Med Sci 2010;25:613-618.*
58. Martinez JM, Garakani A, Kaufmann H, Aaronson CJ, Gorman JM. *Heart rate and blood pressure changes during autonomic nervous system challenge in panic disorder patients. Psychosom Med 2010;72:442-449.*
59. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. *Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. Clin Transl Sci 2024;17:e70056.*
60. Dietterich TG. *Ensemble methods in machine learning. In: Kittler J, Roli F, editors. Multiple classifier systems. Springer; 2000.*
61. Uddin S, Khan A, Hossain ME, Moni MA. *Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 2019;19:281.*
62. Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, *et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. NPJ Digit Med 2021;4:87.*
63. Chawla NV. *Data mining for imbalanced datasets: an overview. In: Maimon O, Rokach L, editors. Data mining and knowledge discovery handbook. 2nd ed. Springer New York; 2010. p.875-886.*

64. Ding C, Peng H. *Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 2005;3:185-205.*
65. Pearson K. *LIII. On lines and planes of closest fit to systems of points in space. Lond Edinb Dubl Phil Mag 1901;2:559-572.*
66. Dwyer DB, Falkai P, Koutsouleris N. *Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 2018;14:91-118.*
67. Hovland A, Pallesen S, Hammar Å, Hansen AL, Thayer JF, Tarvainen MP, *et al. The relationships among heart rate variability, executive functions, and clinical variables in patients with panic disorder. Int J Psychophysiol 2012;86:269-275.*