Original Paper

# Forecasting Next-Day Alcohol Craving in Individuals With Problematic Alcohol Use With Multimodal Data and a Tiny Recurrent Neural Network: Prospective Longitudinal Study

Dongwon Lee[1*], BS; Ji-Yoon Lee[2*], PhD; Hyeonmin Lee[3], MA; Mina Kwon[3], PhD; Suyeon Song[3], MA; Jeong-hyun Yoon[4]; Young-Chul Jung[5,6], MD, PhD; Woo-Young Ahn[1,3,7], PhD; Jung-Seok Choi[8], MD, PhD

[1]AI Institute, Seoul National University, Seoul, Republic of Korea
[2]Center for Artificial Intelligence in Healthcare, Seoul National University Bundang Hospital, Seongnam, Gyeonggi-do, Republic of Korea
[3]Department of Psychology, Seoul National University, Seoul, Republic of Korea
[4]Blue Signum, Seoul, Republic of Korea
[5]Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea
[6]Department of Psychiatry, Yonsei University College of Medicine, Seoul, Republic of Korea
[7]Department of Brain and Cognitive Sciences, Seoul National University, Seoul, Republic of Korea
[8]Department of Psychiatry, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
*these authors contributed equally

**Corresponding Author:**

Jung-Seok Choi, MD, PhD
Department of Psychiatry
Samsung Medical Center, Sungkyunkwan University School of Medicine
81 Irwon-ro, Gangnam-gu
Seoul 06351
Republic of Korea
Phone: 82 2-3410-3581
Fax: 82 2-3410-0050
Email: choijs73@skku.edu

## Abstract

**Background:** Alcohol craving is associated with drinking behaviors despite negative consequences and relapse following periods of abstinence or treatment. Accurately forecasting alcohol craving is crucial for preventing alcohol consumption and relapse, thereby mitigating alcohol-related problems. Previous studies have assessed alcohol craving using diverse modalities such as psychological assessments, neurophysiological measures including electroencephalography (EEG) or heart rate variability (HRV) at a single time point, and digital phenotypes collected longitudinally via smartphone apps. However, alcohol craving is inherently heterogeneous and multidimensional across individuals and over time, making it difficult to capture using any single modality alone. In contrast, multimodal AI enables the prediction of alcohol craving by integrating complementary behavioral and neurophysiological features.

**Objective:** This study aimed to develop and evaluate AI models to forecast next-day alcohol craving in individuals with problematic alcohol use with multimodal datasets.

**Methods:** We analyzed multimodal datasets comprising smartphone-based digital phenotypes collected over 28 days, along with baseline EEG, HRV, and psychological measures in a prospective longitudinal study of 60 participants with problematic alcohol use. A tiny multimodal recurrent neural network (RNN) was implemented, in which representations of static baseline features were encoded into a latent representation and used to initialize the hidden state of the RNN. Model performance was evaluated using accuracy, weighted $F_1$-score, macro $F_1$-score, and mean absolute error (MAE).

**Results:** The multimodal model consistently outperformed the unimodal model based on digital phenotypes alone, achieving higher accuracy (0.56 vs 0.45), weighted $F_1$-score (0.51 vs 0.38), and macro $F_1$-score (0.37 vs 0.24), along with lower MAE (0.68 vs 0.90). At the individual level, 72% (43/60) of participants showed improved prediction accuracy with the multimodal model. These improvements were statistically significant (Wilcoxon signed-rank test, $P<.001$; paired $t$ test, $P<.001$).

**Conclusions:** Integrating digital phenotypes with neurophysiological and psychological features enhances the prediction of short-term alcohol craving dynamics. This approach may enable more precise and timely personalized interventions for individuals at risk of relapse.

***J Med Internet Res 2026;28:e101973;*** doi: 10.2196/101973

**Keywords:** alcohol craving; artificial intelligence; digital phenotypes; ecological momentary assessment; multimodality; recurrent neural network; smartphone applications

## Introduction

According to the 2024 World Health Organization Alcohol Fact Sheet, alcohol consumption was responsible for approximately 2.6 million deaths worldwide in 2019 [1]. Alcohol consumption is linked to an increased risk of mental health conditions such as depression, anxiety, and alcohol use disorders (AUDs), as well as liver disease, heart disease, and cancer [1]. AUD is a chronic relapsing disorder that affects physical, psychological, and social functioning, highlighting the need for sustained treatment and ongoing clinical care [2]. The disorder is a leading contributor to illness and death but is frequently improperly diagnosed or treated in clinical settings, and as a result, the burden of disease remains high [2]. Problematic alcohol use spans a wide spectrum ranging from nonclinical binge drinking to AUD [3]. Therefore, early identification and intervention are important before drinking behaviors progress to more severe alcohol-related problems [4].

Alcohol craving is a critical factor in problematic alcohol use, contributing to persistent drinking behavior as well as relapse and the maintenance of alcohol dependence [5,6]. Craving has traditionally been regarded as a primary driver of alcohol use, with cognitive processing models suggesting that alcohol consumption often occurs through automatized behaviors, whereas craving reflects a conscious cognitive response to the disruption or inhibition of drinking processes [7]. Sociodemographic disparities in alcohol craving and AUD outcomes have also been reported in previous studies [8]. Assessing craving may help clarify its association with alcohol relapse in predicting drinking during treatment [9]. Alcohol craving is a time-dependent event that varies across situations and contexts and requires repeated assessments in natural environments [10]. It was shown to predict subsequent alcohol use, and higher craving levels were associated with an increased likelihood of drinking and heavy drinking in the following week [11]. This predictive association appears to be strongest over short time frames and identifies craving as a key indicator of forthcoming drinking risk [11]. However, many previous studies have assessed craving at a single modality and time point using cross-sectional approaches, which may fail to capture the dynamic fluctuations of craving in daily life [12]. Therefore, optimal measurement of alcohol craving requires assessing it at least daily, particularly in high-risk situations for drinking or when craving significantly impacts functioning [12].

Given that alcohol craving fluctuates dynamically across time and contexts, continuous real-time monitoring through digital phenotyping may provide a promising approach for capturing these temporal variations in naturalistic settings. Advances in digital technologies and data analytics have created unparalleled opportunities to assess and modify health behavior and thus accelerate the ability of science to understand and contribute to improved health behavior and health outcomes [13]. Digital phenotyping measures behavior from smartphone sensors, keyboard interaction, and various features of voice and speech. It can capture diverse variables spanning psychomotor activity, mood state, geolocation, and movement in real time, with a potential to assess features related to specific addictive behaviors, such as urges, cravings, lifestyle factors, and high-risk environmental situations [14,15]. In a recent study, alcohol craving exhibited a systematic rhythm over the course of a 24-hour period, and alcohol use frequency was associated with the shape of these rhythms [16]. Findings suggest that consideration of daily rhythms in alcohol craving may further enhance the understanding of mechanisms driving alcohol use [16]. This approach would be useful for assessing craving as a time-dependent event, and further research is needed to validate its utility.

While digital phenotyping provides valuable insights into momentary behavioral states in everyday life, alcohol craving is also influenced by stable individual characteristics related to addiction vulnerability. Neurophysiological measures such as electroencephalography (EEG) and heart rate variability (HRV) have been associated with reward processing, emotional regulation, and autonomic functioning in substance use disorders. Resting-state EEG findings indicate that subjective craving is associated with frontal beta activity, reflecting an electrophysiological correlate of alcohol craving induced by personalized cues [17]. A previous functional magnetic resonance imaging–EEG study showed that alcohol craving was associated with abnormal activity in key brain regions, including the dorsal anterior cingulate cortex, nucleus accumbens, posterior cingulate cortex, pregenual anterior cingulate cortex, amygdala, and parahippocampus, and was also linked to increased beta-band activity in resting-state EEG [18]. Another HRV study showed that high-frequency HRV is an important predictor of alcohol craving, supporting its role as a potential biomarker of craving in individuals with alcohol dependence and its involvement in self-regulatory processes [19]. These trait-level biomarkers may reflect underlying neurobiological susceptibility to craving and relapse and should therefore be considered alongside dynamic behavioral data in alcohol craving prediction.

In this context, psychiatric features are heterogeneous and multidimensional, and vary across individuals, and therefore may not be adequately captured by unimodal approaches

[20]. Digital phenotyping captures temporal fluctuations in real-world behavior but overlooks underlying neurobiological vulnerability, whereas neurophysiological studies identify trait-level markers but may lack sensitivity to time-varying dynamics. To address these limitations, a previous review has highlighted the integration of neuroimaging and digital phenotyping, demonstrating its potential to link behavioral dynamics with underlying neural mechanisms [21]. However, this approach remains in its early stages, with key challenges including temporal misalignment between modalities and the limited use of multivariate analyses and machine learning (ML) approaches [21]. Recent advances in AI and ML provide powerful approaches for modeling complex interactions across multimodal datasets [22]. ML approaches are particularly well-suited to identifying temporal misalignment and modeling complex, time-varying relationships across modalities. This integration may provide a more comprehensive understanding of alcohol craving.

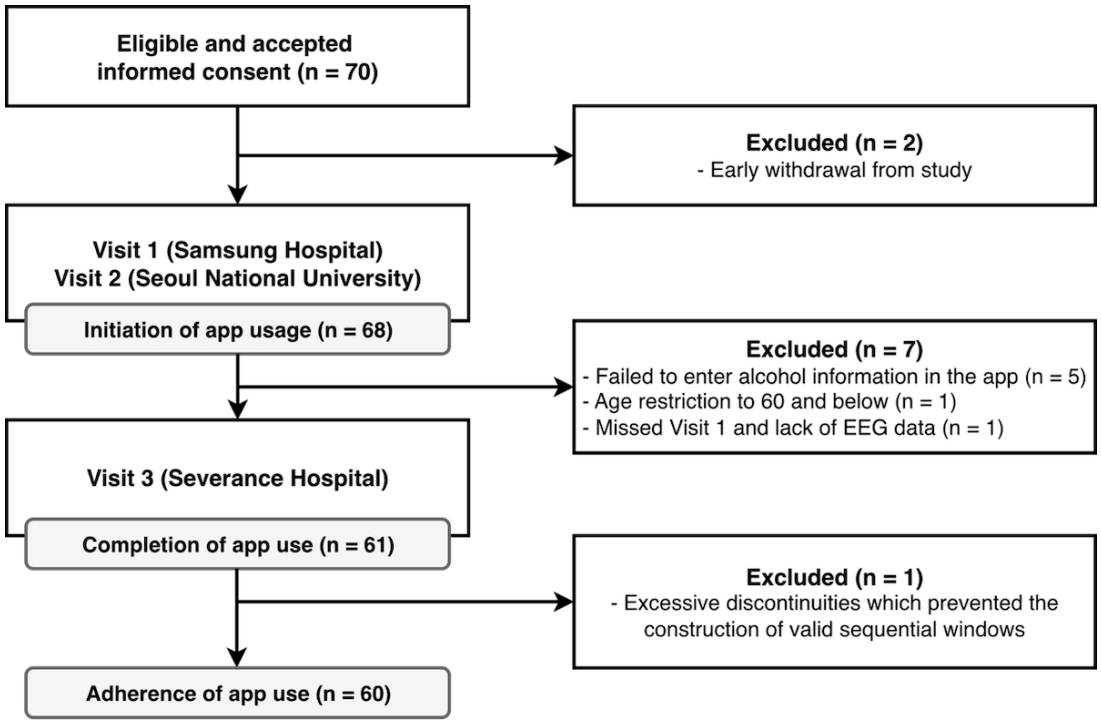
Taken together, this study aims to predict next-day alcohol craving in individuals with problematic alcohol use using multimodal datasets. To account for temporal discrepancies across modalities, we used a tiny recurrent neural network (RNN) approach that explicitly models time-dependent relationships. We compared AI algorithms trained on only longitudinal digital phenotypes collected via a smartphone app over 28 days with those on a fully integrated dataset using digital phenotypes, EEG, HRV, and psychological features. We hypothesized that an AI model with a fully integrated multimodal dataset would outperform those using only longitudinal digital phenotypes.

# Methods

## Participants

The study participants were adults aged 19 to 59 years. The upper age limit was set to minimize potential confounding effects of age-related neurophysiological and structural brain changes [23-26]. The sample size was informed by previous EEG and ML studies in alcohol-related populations with comparable sample sizes [27-29]. A total of 70 participants who self-reported problematic drinking behaviors were recruited. Participants were recruited between July 2023 and July 2024 through an online advertisement posted on AllLiveC, a platform specializing in clinical trial and research participant recruitment, targeting individuals with self-reported problematic drinking behaviors. Participants were not receiving formal treatment for AUD during the study period. Exclusion criteria included inability to understand the consent form or provide informed consent, pregnancy or plans for pregnancy, participation in another clinical trial, use of more than 2 phones, and claustrophobia or contraindications for magnetic resonance imaging scanning. Two participants withdrew early from the study, and 7 participants were excluded after study initiation due to age, failure to enter alcohol information in the mobile app, or missing visit 1. Additionally, 1 participant was excluded due to excessive discontinuities that prevented the construction of valid sequential windows. As a result, the final sample consisted of 60 participants in the final analysis (Figure 1).

**Figure 1.** Participant flow diagram.



## Ethical Considerations

The study protocol was approved by the institutional review board of Severance Hospital, Yonsei University Health System (4-2023-1590) and Samsung Medical Center (2023-02-115), Seoul, South Korea. All participants provided informed written consent before enrollment. To protect participant privacy and confidentiality, participant personal information was managed in a pseudonymized form throughout the study. Participants received compensation according to their level of app-use participation, with compensation averaging approximately KRW 200,000 (US $149) per participant.

## Assessments by Visit

At the baseline visit (visit 1), participants' demographic information was collected, and assessments for HRV, EEG, and psychological and cognitive function were conducted. During this visit at Samsung Medical Center, Seoul, participants also completed interviews with the Mini-International Neuropsychiatric Interview to rule out comorbid psychiatric diagnoses and were evaluated for AUD based on *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* [30] criteria. Cognitive function was assessed using the Korean Wechsler Adult Intelligence Scale-IV and several psychometric assessments. Additionally, EEG and HRV tests were performed.

On visit 2, the Harujjan application was installed on participants' mobile devices, and they received detailed, step-by-step instructions on its usage. Participants were guided on how to log their daily entries, including mood, sleep quality, and alcohol intake, and set reminders to ensure consistent usage throughout the study. Ecological momentary assessment (EMA) assessments were administered once daily via a mobile app and were completed at the end of each day at a participant-selected time. Automated push notifications were provided at the selected time, and each survey took approximately 5 minutes to complete. The daily EMA survey comprised 17 items across 4 domains: mood (overall mood and interest/pleasure), diet (meal consumption and meal quantity), alcohol-related behaviors (alcohol craving, alcohol consumption, drinking time, beverage type, amount consumed, drinking companions, drinking duration, drinking location, drinking motives, and blackout experience), and sleep (sleep duration, sleep satisfaction, and sleep-related daytime interference). Some items were conditionally presented based on participants' responses. If participants consumed additional food or alcohol after completing the survey, they could submit additional entries until the following morning. Surveys were completed daily regardless of whether craving was experienced; no event-contingent assessments were used.

At the final visit (visit 3), after 4-week mobile app usage, participants completed an HRV test and a psychometric assessment.

## Datasets

### Outcome and Predictors

The primary outcome was next-day alcohol craving, assessed using a single-item EMA ("How much did you want to drink alcohol today?"), rated on a 5-point Likert scale ranging from 1 (not at all) to 5 (very much). For modeling, this outcome was treated as a 5-class classification target, with the craving level reported on the following day serving as the prediction target.

Predictors comprised longitudinal EMA-based time-series features, including mood, dietary, sleep, and alcohol consumption measures, and static baseline features, including cognitive, alcohol-use, psychological, EEG, and HRV measures.

### Psychological Features

Psychological features included Drinking Motives Questionnaire (DMQ) [31,32] scores for social, coping, enhancement, and conformity; Alcohol Use Disorders Identification Test (AUDIT) [33,34], Beck Depression Inventory-II (BDI-II) [35,36], Beck Anxiety Inventory (BAI) [37,38]; raw and transformed Barratt Impulsiveness Scale (BIS) [39,40] total score and subscales for attentional, motor, and nonplanning; Pittsburgh Sleep Quality Index (PSQI) [41,42] total score and subscales for subjective sleep quality, sleep latency, sleep duration, sleep efficiency, sleep disturbance, use of sleep medication, and daytime dysfunction; and Stages of Change Readiness and Treatment Eagerness Scale (SOCRATES) [43,44] subscales for ambivalence, recognition, and taking steps. All of those scales were validated for Korean use. Intelligence quotient was assessed by the Korean Wechsler Adult Intelligence Scale-IV [45].

### EEG

Resting-state EEG was recorded in a sound-shielded room for 10 minutes: 5 minutes with eyes closed and 5 minutes with eyes open, using a 64-channel Quik-Cap (Compumedics Neuroscan) following the modified International 10-20 system. All EEG recordings were obtained using SynAmps 2 (Compumedics) and the Neuroscan system (Scan 4.5; Compumedics). The electrode impedance was maintained below 5 kΩ. EEG data were processed using Python (version 3.12.2), MNE-python (version 1.8.0) [46], and autorejection [47]. Artifact-free 2-second epochs from the eyes-closed condition were selected for analysis. The relative power was computed for 5 frequency bands: delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-25 Hz), and gamma (30-40 Hz).

### HRV

HRV data were obtained during a 5-minute resting-state recording. Participants remained seated while electrocardiogram electrodes were placed on both wrists and the left ankle. Electrocardiogram signals were recorded using a ProComp Infiniti system (SA7500 Computerized Biofeedback System; Thought Technology Ltd) equipped with an electrocardiogram-Flex/Pro sensor, with a sampling rate of 256 Hz and

a band-pass filter of 0.26-30 Hz. Participants were asked to abstain from caffeine and nicotine for at least 2 hours before the recording. HRV features included the following: average beats per minute, standard deviation of normal-to-normal intervals, physical stress index, total power, very low-frequency power, low-frequency power, high-frequency power, normalized low-frequency power, normalized high-frequency power, low-frequency to high-frequency ratio, root mean square of successive differences, approximate entropy, short-range detrended fluctuation analysis, total short-range detrended fluctuation analysis, natural logarithm of total power, natural logarithm of very low-frequency power, natural logarithm of low-frequency power, and natural logarithm of high-frequency power.

## Data Preprocessing

### Handling Missing Data

Participants were instructed to complete EMA entries for 28 consecutive days. However, adherence varied substantially across individuals. While some participants provided entries beyond the 28-day period, others reported fewer observations, and many exhibited irregular reporting patterns with intermittent day gaps between consecutive days. Among the final analytic sample of 60 participants, 29 provided EMA responses for exactly 28 days, while the number of EMA response days ranged from 21 to 30 across participants (Figure S3 in Multimedia Appendix 1). Additionally, 4 participants exhibited nonconsecutive reporting patterns with day gaps. Given the sequential nature of our modeling framework, we did not impute missing days to artificially construct continuous time series, as such imputation may introduce unrealistic temporal dependencies and bias the learning process [48]. Instead, we adopted a sliding window framework and excluded invalid sequences based on temporal consistency. Specifically, we removed (1) windows containing gaps greater than one day within the input sequence and (2) windows where the temporal gap between the final input observation and the prediction target exceeded 2 days. This ensured that all training samples preserved meaningful temporal continuity and temporally consistent short-term behavioral patterns. For missing values within individual features, we applied domain-informed imputation strategies tailored to each variable type.

### Feature Selection

After data preprocessing, the dataset consisted of 394 baseline features with 60 samples and 45 time-series features from the app-based EMA data. Given the high-dimensional feature space relative to the limited sample size (ie, $P>>N$), the model was highly susceptible to overfitting, which can degrade generalization performance on unseen datasets [49]. Therefore, feature selection was a critical step to reduce the dimensionality and improve model robustness [50]. Among various methods, random forest–based feature selection is widely used for high-dimensional, low-sample-size settings [51]. Accordingly, random forest–based feature selection was applied separately to the EMA (time-series) data and baseline (static) data.

For the EMA data, feature importance was computed using a random forest classifier trained exclusively on the training portion of the dataset, excluding the final 8 days per participant used for validation and testing to prevent data leakage. Features with importance values above the mean importance threshold were retained. For the baseline data, random forest regression was applied using the participant-level average alcohol craving as the target variable. The average craving was computed using only the training data (excluding the final 8 days) to prevent information leakage. Based on random forest feature importance, 42 baseline features were retained. After feature selection, 11 features remained for the EMA data and 42 features for the baseline data, substantially reducing the feature space while preserving the most informative predictors. One additional feature derived from EEG relative power (epoch count) was manually excluded, as it was not considered a meaningful physiological predictor but rather reflected data preprocessing characteristics.

Consequently, a total of 41 baseline features were retained for subsequent analyses. This separate feature selection strategy was adopted to account for the fundamentally different statistical properties of time-series and static data. Selected EMA features included mood (Patient Health Questionnaire) items, dietary patterns (meal frequency), sleep quality, and alcohol consumption variables. Selected baseline features included cognitive measures (eg, intelligence quotient and working memory index), alcohol-use measures, psychological scales (eg, DMQ, BIS, and SOCRATES), EEG band power across multiple regions (delta, theta, beta, and gamma), and HRV-derived metrics. A full list of the selected features of the baseline and the EMA data is provided in Tables S1 and S2 in Multimedia Appendix 1.

### Data Splitting Strategy (Sliding Window Approach)

Our objective was to predict the next-day alcohol craving using prior observations. Forecasting refers to the prediction of future outcomes based on past and present information [52,53]. In time-series forecasting, a sequence of past observations, often referred to as a sliding window, is used to predict future targets [54]. When evaluating the performance of time-series models, it is critical to ensure that no future information is used during training, and that only data available at the time of prediction are used for parameter estimation [53]. To preserve the temporal structure of the data and prevent information leakage, a participant-wise time-series split was applied. For each participant, the last 5 days were used as the test set, and the preceding 3 days were used as the validation set. The remaining observations were used for training. This approach ensures that the model is evaluated on future data that were not available during training, thereby reflecting a realistic forecasting scenario while preserving temporal ordering and preventing information leakage from future observations. The choice of lookback-window length involves a trade-off between retaining sufficient temporal context and avoiding irrelevant or noisy historical information

[55]. We therefore examined the robustness of different window lengths, with detailed sensitivity analyses for 2-, 4-, 6-, and 8-day windows provided in Table S4 in Multimedia Appendix 1. A sliding window approach with a window size of 4 was used for the primary analyses, providing a short-term temporal context for next-day forecasting.

## Feature Scaling

Feature scaling is an essential preprocessing step in ML that ensures variables are on a comparable scale, thereby improving model stability and performance [56]. We applied standardization using the standard scaler, a widely used normalization technique [57]. Standardization transforms each feature to have zero mean and unit variance, computed as:

$$z = \frac{X - \mu}{\sigma} \quad (1)$$

where $z$ is the transformed value, $X$ is the original feature value, $\mu$ is the mean, and $\sigma$ is the SD [58]. To prevent data leakage, the scaler was fitted exclusively on the training data. The same transformation parameters were then applied to the validation and test sets.

## Modeling

### RNNs

Traditional clinical approaches often integrate static and time-series data by extracting statistical features from the temporal data, which are then combined with static features. However, such approaches remove the temporal structure of the data and require manual engineering of temporal features, which can be labor-intensive and may result in the loss of temporal information [13,59]. In contrast, RNNs are designed to model sequential data and can capture temporal dependencies directly from raw time-series inputs, enabling more effective use of temporal information [60]. However, clinical datasets typically consist of both static (eg, EEG, HRV, and psychological assessments) and dynamic (EMA time-series) features. Due to their nonsequential nature, static features cannot be incorporated into standard RNN architectures [61,62].

To address this limitation, we adopted a tiny (lightweight) RNN framework that facilitates the integration of multimodal data while maintaining model simplicity. The use of a compact architecture is particularly important given the limited sample size, as it reduces the risk of overfitting while preserving the ability to capture meaningful temporal patterns. Furthermore, we hypothesize that incorporating highly informative multimodal baseline features such as EEG, HRV, and psychological assessments [18,19] allows the model to leverage complementary information that is not directly observable from EMA data alone, thereby improving model performance.

### Training the Unimodal Dataset

To establish a baseline for evaluating the contribution of multimodal integration, the model was first trained using only the EMA time-series data. A tiny vanilla RNN framework with a single hidden layer of 16 units was adopted to model temporal dependencies while minimizing the risk of overfitting given the limited sample size. The hidden state of the vanilla RNN is updated at each time step according to:

$$h_t = tanh(W_{xh}X_t + W_{hh}h_{t-1} + b) \quad (2)$$

where $X_t \in R^n$ is the input at time step $t$, $h_t \in R^m$ is the hidden state, and $b$ is the bias term. The hidden state serves as a dynamic representation of temporal information and is initialized as a zero vector [63]. Given that our target variable is encoded on a Likert scale, the task was formulated as a 5-class classification problem. Therefore, a softmax layer was applied to obtain class probabilities:

$$p_i = \frac{exp\,(z_i)}{\Sigma_i\ exp\,(z_i)} \quad (3)$$

and the predicted class was defined as:

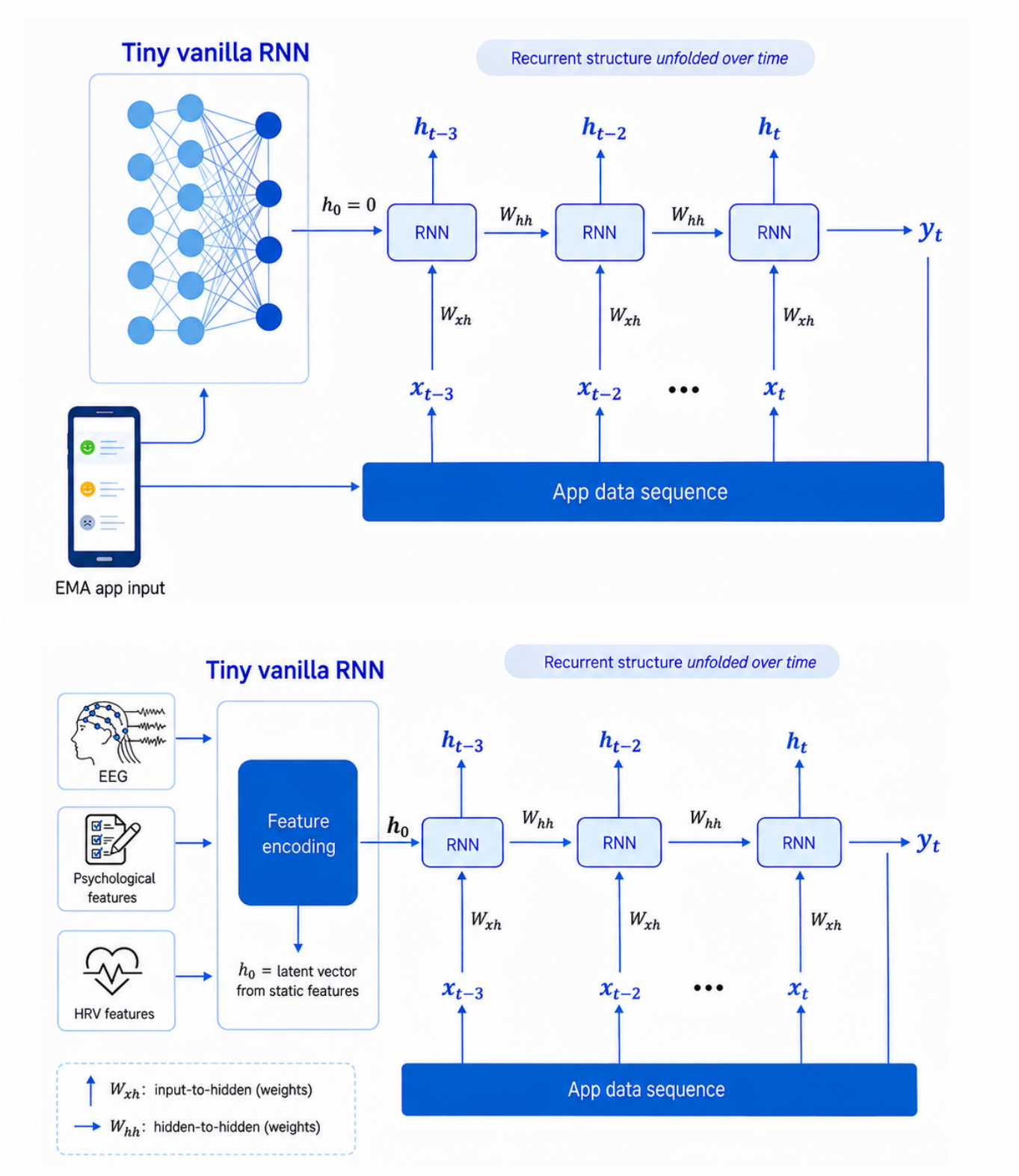
$$y_t = \underset{i}{arg\,max}\ p_i \quad (4)$$

where $z_i$ denotes the output of the final linear layer and $p_i$ is the probability of class $i$. The predicted values for the next-day alcohol craving correspond to the class with the highest probability.

### Training the Multimodal Dataset

To effectively integrate static and time-series data, we adopted a tiny multimodal RNN framework that initializes the hidden state of the RNN using representations learned from static baseline features. This approach enables the model to incorporate subject-specific contextual information into the temporal dynamics of the sequence model, allowing it to capture temporal dependencies while preserving informative baseline characteristics (Figure 2).

**Figure 2.** Model architectures. Top: Unimodal model — a vanilla recurrent neural network (RNN) processes ecological momentary assessment (EMA) time-series with the hidden state initialized as a zero vector. Bottom: Multimodal model — static baseline features (electroencephalography [EEG], heart rate variability [HRV], and psychological measures) are encoded into a latent vector via a feedforward network, which initializes the RNN hidden state, allowing individual-specific baseline characteristics to directly condition temporal predictions.



Specifically, static baseline features (eg, EEG, HRV, and psychological assessments) are first transformed into a latent representation using a single-layer feedforward neural network.

This network consists of one hidden layer with a rectified linear unit activation function [64] and a dropout rate of 20% to mitigate overfitting. The resulting representation is then used to initialize the hidden state of the RNN. Subsequently, EMA time-series data are processed sequentially by the RNN, similar to the unimodal setting.

The key difference is that, instead of initializing the hidden state with a zero vector, the multimodal model initializes it with the learned representation of the static baseline features. This allows the model to condition temporal predictions on individual-specific baseline characteristics. The use of a compact architecture is particularly important given the limited sample size (n=60), as it reduces model complexity while maintaining the ability to capture both temporal dependencies and cross-modal interactions.

A key aspect of our approach is the integration of static baseline features as initial hidden states in the recurrent model, allowing subject-specific contextual information to directly influence temporal predictions.

### Model Training

All models were implemented using the open-source deep learning library PyTorch [65]. The training data were processed in minibatches of size 32 with shuffling enabled, while validation and test data were evaluated using larger batch sizes of 256 without shuffling. To ensure robustness and reduce sensitivity to random initialization, each model was trained 100 times with different random seed values (ranging from 2026 to 2125). Performance metrics were summarized across the 100 runs using the mean, SD, and 95% CI of the mean. The model parameters were optimized using the Adam optimizer [66] with a learning rate of 0.001. The categorical cross-entropy loss was used for training. For both unimodal and multimodal models, training was conducted for a maximum of 70 epochs. Early stopping was applied based on validation loss, where the model parameters corresponding to the lowest validation loss were selected for final evaluation to prevent overfitting.

### Evaluation Setting

To ensure a realistic forecasting setting, model evaluation was conducted using only information available prior to the prediction time point, preventing the use of future data for parameter estimation [53]. Accordingly, the dataset was split along the temporal dimension for each participant, rather than participant-level splitting. The last 5 days were used as the test set, and the preceding 3 days were used as the validation set for hyperparameter tuning, with the remaining observations used for training.

Model performance was evaluated using multiple metrics to account for class imbalance and the ordinal nature of the target variable. Accuracy is the ratio of the number of correctly classified samples over the total number of samples [67]. Accuracy was reported as an overall performance measure; however, when evaluating imbalanced datasets, accuracy may be misleading, as it considers only the total number of correct predictions without accounting for the distribution of class labels [68]. The imbalanced distribution of our craving labels is shown in Figure S3 in Multimedia Appendix 1.

Therefore, aggregated metrics such as $F_1$-score were also considered to address the issue of class imbalance [69]. $F_1$-score is the weighted harmonic mean of precision and recall [70]. While weighted $F_1$-score is used when there are imbalanced classes, it may be biased toward the majority class [71,72]. On the other hand, macro $F_1$-score treats all classes equally by assigning the same weight to each class, ensuring that minority classes also contribute to the evaluation and that the model is penalized if it performs poorly on them [73]. Therefore, both metrics were reported.

In addition, since the target variable is ordinal, prediction errors of smaller magnitude are more acceptable than larger deviations. In imbalanced ordinal classification settings, metrics that account for the magnitude of prediction errors, such as mean absolute error (MAE), have been shown to provide informative evaluation [74]. Thus, model performance was evaluated using accuracy, weighted $F_1$-score, macro $F_1$-score, and MAE.

To explore participant-level factors potentially associated with prediction performance, post hoc Spearman rank correlations were calculated between participant-wise mean test accuracy and EMA-related characteristics, including EMA response days, within-participant craving SD, craving entropy, the number of unique craving levels, and the number of usable sequential samples. Within-participant craving SD, craving entropy, and the number of unique craving levels were calculated across each participant's observed EMA period. Correlations with the difference between multimodal and unimodal accuracy (Δaccuracy) were also examined.

## Results

### Demographic Features

The demographic characteristics of the participants are presented in Table 1. The average age of the participants was 34.17 (SD 8.62) years, with 25 males and 35 females included.

**Table 1.** Demographic features of the participants.

| Demographic | Value |
|---|---|
| Age (years), mean (SD) | 34.17 (8.62) |
| Sex | |
| Male | 25 (41.7) |
| Female | 35 (58.3) |
| IQ, mean (SD) | 110.83 (14.12) |
| Education (years), mean (SD) | 15.65 (3.11) |
| Marriage, n (%) | |
| Yes | 19 (31.7) |
| No | 41 (68.3) |
| Smoking, n (%) | |
| Yes | 14 (23.3) |
| No | 46 (76.7) |
| AUDIT[a] score, mean (SD) | 18.15 (6.67) |

[a]AUDIT: Alcohol Use Disorders Identification Test.

## EMA Completion and Craving Characteristics

Among the 60 participants included in the final analytic sample, participants provided EMA responses for a mean of 27.88 (SD 1.35; median 28, IQR 28-29; range 21-30) days. A total of 29 (48.3%) participants provided responses for exactly 28 days, while 47 (78.3%) participants provided responses for at least 28 days. Some participants continued to submit EMA entries beyond the planned 28-day period, with 18 (30%) participants providing more than 28 days of data.

Before construction of the next-day prediction samples, a total of 1673 alcohol craving reports were submitted. The mean EMA-reported alcohol craving score was 2.21 (SD 1.13), with a median of 2 (IQR 1-3) and a range of 1-5.

## Model Performance

The unimodal and multimodal models were compared across the aforementioned evaluation metrics, including accuracy, weighted $F_1$-score, macro $F_1$-score, and MAE. The results of the 5-class classification are reported as the mean, SD, and 95% CI of the mean across 100 runs with different random seeds (Table 2).

**Table 2.** Model performance comparison of unimodal and multimodal models. Values are presented as mean (SD) across 100 runs.

| | Train, mean (SD) | Validation, mean (SD) | Test, mean (SD; 95% CI) |
|---|---|---|---|
| Unimodal | | | |
| Accuracy | 0.46 (0.01) | 0.46 (0.02) | 0.45 (0.01; 0.443-0.447) |
| Weighted $F_1$-score | 0.38 (0.02) | 0.39 (0.02) | 0.38 (0.01; 0.372-0.378) |
| Macro $F_1$-score | 0.26 (0.04) | 0.23 (0.02) | 0.24 (0.04; 0.235-0.249) |
| MAE[a] | 0.90 (0.04) | 0.85 (0.04) | 0.90 (0.03; 0.894-0.904) |
| Multimodal | | | |
| Accuracy | 0.59 (0.02) | 0.53 (0.03) | 0.56 (0.02; 0.552-0.560) |
| Weighted $F_1$-score | 0.53 (0.04) | 0.49 (0.03) | 0.51 (0.03; 0.504-0.517) |
| Macro $F_1$-score | 0.43 (0.06) | 0.32 (0.05) | 0.37 (0.05; 0.364-0.381) |
| MAE | 0.62 (0.04) | 0.72 (0.04) | 0.68 (0.04; 0.672-0.686) |

[a]MAE: mean absolute error.

Overall, the multimodal model consistently outperformed the unimodal model across all evaluation metrics. In terms of accuracy, the multimodal model achieved 0.56 (SD 0.02; 95% CI 0.552-0.560) on the test set, compared to 0.45 (SD 0.01; 95% CI 0.443-0.447) for the unimodal dataset. Considering that a random classifier would achieve an accuracy of approximately 20% in a 5-class classification setting, and a majority-class classifier would yield approximately 37%, both models demonstrate meaningful predictive performance, with the multimodal model providing substantial improvement.

Given the imbalanced class distribution, $F_1$-scores were also evaluated. The unimodal model achieved a weighted $F_1$-score of 0.38 (SD 0.01) and a macro $F_1$-score of 0.24 (SD 0.04) on the test set, whereas the multimodal model improved these scores to 0.51 (SD 0.03) and 0.37 (SD 0.05), respectively. The improvement in macro $F_1$-scores suggests that the multimodal model better captures minority class patterns, leading to more balanced performance across classes. Furthermore, MAE was used to account for the magnitude of prediction errors given the ordinal nature of the prediction task. The unimodal model yielded an MAE

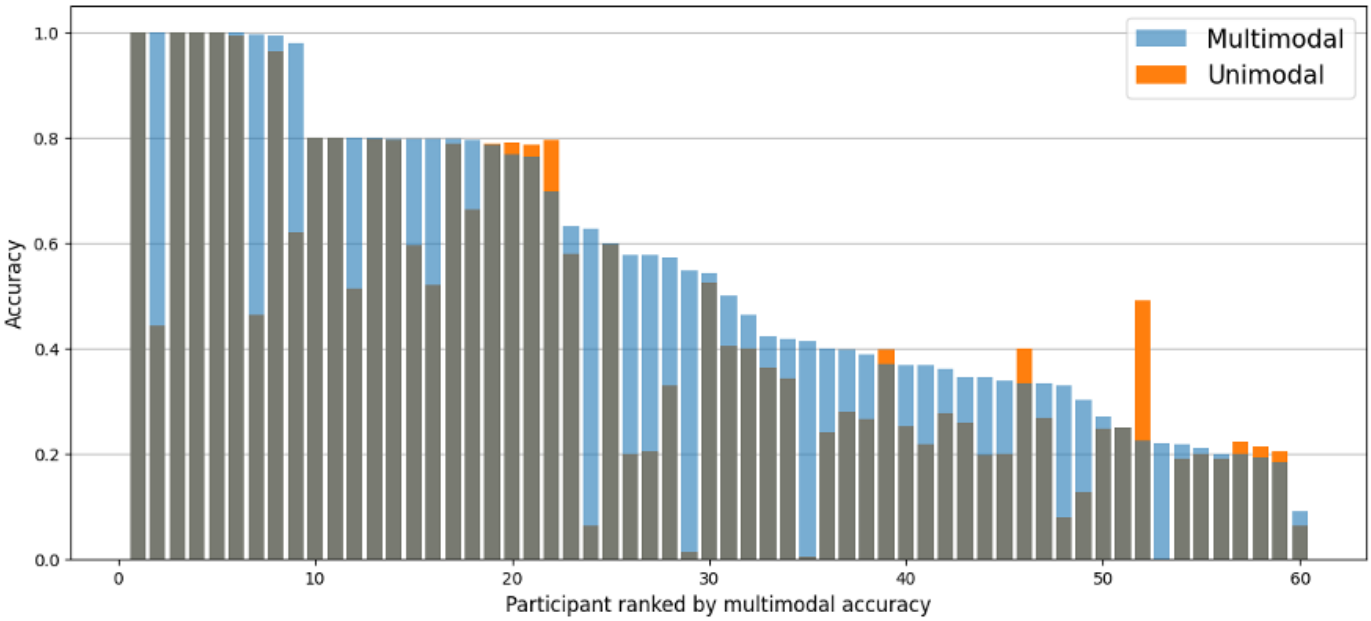
of 0.90 (SD 0.03), whereas the multimodal model reduced this to 0.68 (SD 0.04). This indicates that the multimodal model not only improves the classification accuracy but also reduces the magnitude of prediction errors, producing predictions that are closer to the true ordinal labels. These results demonstrate that incorporating static baseline features into the temporal modeling framework provides complementary information that enhances both classification performance and error sensitivity.

Sample-level confusion matrices based on the modal prediction across the 100 runs for the unimodal and multimodal models, including raw-count and row-normalized representations, are provided in Figure S4 in Multimedia Appendix 1. The multimodal model showed improved class-specific prediction across several craving levels, including some less frequent classes. Notably, correct classification of craving level 5 increased from 0% in the unimodal model to 33% in the multimodal model, although performance for craving level 4 remained limited.

## Accuracy Comparison at the Individual Level

To further examine whether the multimodal model consistently improves performance at the individual level, we compared participant-wise accuracy between the unimodal and multimodal models. For each participant, the difference in accuracy between the multimodal and unimodal models was computed. The multimodal model consistently outperformed the unimodal model at the individual level, with 72% (43/60) of participants showing higher prediction accuracy with the multimodal model, indicating that the performance improvement was consistent across the majority of individuals. Participant-wise accuracy was significantly higher for the multimodal model than for the unimodal model. This improvement was confirmed by both the Wilcoxon signed-rank test (W=176; $P<.001$) and the paired $t$ test ($t_{59}=5.00$; $P<.001$), treating the accuracy values as paired observations for each participant. The use of both nonparametric and parametric tests was motivated by uncertainty regarding the normality of the distribution. As illustrated in Figure 3, these findings demonstrate that the multimodal approach improves predictive performance not only at the aggregate level but also consistently across individuals.

**Figure 3.** Participant-wise prediction accuracy for the unimodal (orange) and multimodal (blue) models across all n=60 participants. Blue bars indicate the multimodal model, orange bars indicate the unimodal model, and gray bars represent the overlapping region where the two bars coincide. The multimodal model achieves higher accuracy in 72% (43/60) of participants, with statistically significant improvement confirmed by both the Wilcoxon signed-rank test and paired $t$ test (both $P<.001$).




To explore potential factors associated with participant-level prediction performance, we conducted post hoc Spearman correlation analyses. EMA response days were not significantly associated with accuracy in either the unimodal ($\varrho=0.035$; $P=.79$) or multimodal ($\varrho=0.085$; $P=.52$). In contrast, greater within-participant craving variability was associated with lower accuracy in both models (unimodal: $\varrho=-0.549$; $P<.001$; multimodal: $\varrho=-0.671$; $P<.001$). Craving entropy and the number of unique craving levels were likewise negatively associated with accuracy in both models, whereas the number of usable sequential samples was not significantly associated with accuracy. None of the examined participant characteristics was significantly associated with Δaccuracy. Detailed results are provided in Table S3 in Multimedia Appendix 1.

## Modality Ablation

To examine the contribution of each baseline modality, we conducted a modality ablation analysis in which each baseline modality was used individually to initialize the RNN hidden state and compared with the unimodal zero-initialization model and the full multimodal model (Table 3). The modality ablation results provide interpretable insight into which signals drive performance gains. HRV features yield the strongest single-modality improvement, consistent with the established role of autonomic regulation in craving and relapse [19]. EEG features also contribute substantially, while psychological features alone produce smaller gains—possibly due to overlap with EMA-collected mood and behavioral data. These findings suggest that clinicians prioritizing resource-constrained deployment might first consider HRV monitoring as a complement to EMA, given its relative ease of collection compared to full EEG acquisition.

**Table 3.** Modality ablation performance on the test set. Values represent means across 100 runs.

| Model | Accuracy, mean (SD) | Weighted $F_1$-score, mean (SD) | Macro $F_1$-score, mean (SD) | MAE[a], mean (SD) |
|---|---|---|---|---|
| Unimodal (zero initialization) | 0.45 (0.01) | 0.38 (0.01) | 0.24 (0.04) | 0.90 (0.03) |
| + EEG[b] only | 0.52 (0.03) | 0.46 (0.04) | 0.33 (0.05) | 0.76 (0.06) |
| + HRV[c] only | 0.53 (0.02) | 0.48 (0.03) | 0.34 (0.04) | 0.73 (0.04) |
| + PF[d] only | 0.50 (0.02) | 0.45 (0.03) | 0.33 (0.04) | 0.80 (0.04) |
| Multimodal (all) | 0.56 (0.02) | 0.51 (0.03) | 0.37 (0.05) | 0.68 (0.04) |

[a]MAE: mean absolute error. All metrics except MAE are higher-is-better.
[b]EEG: electroencephalography.
[c]HRV: heart rate variability.
[d]PF: psychological features.

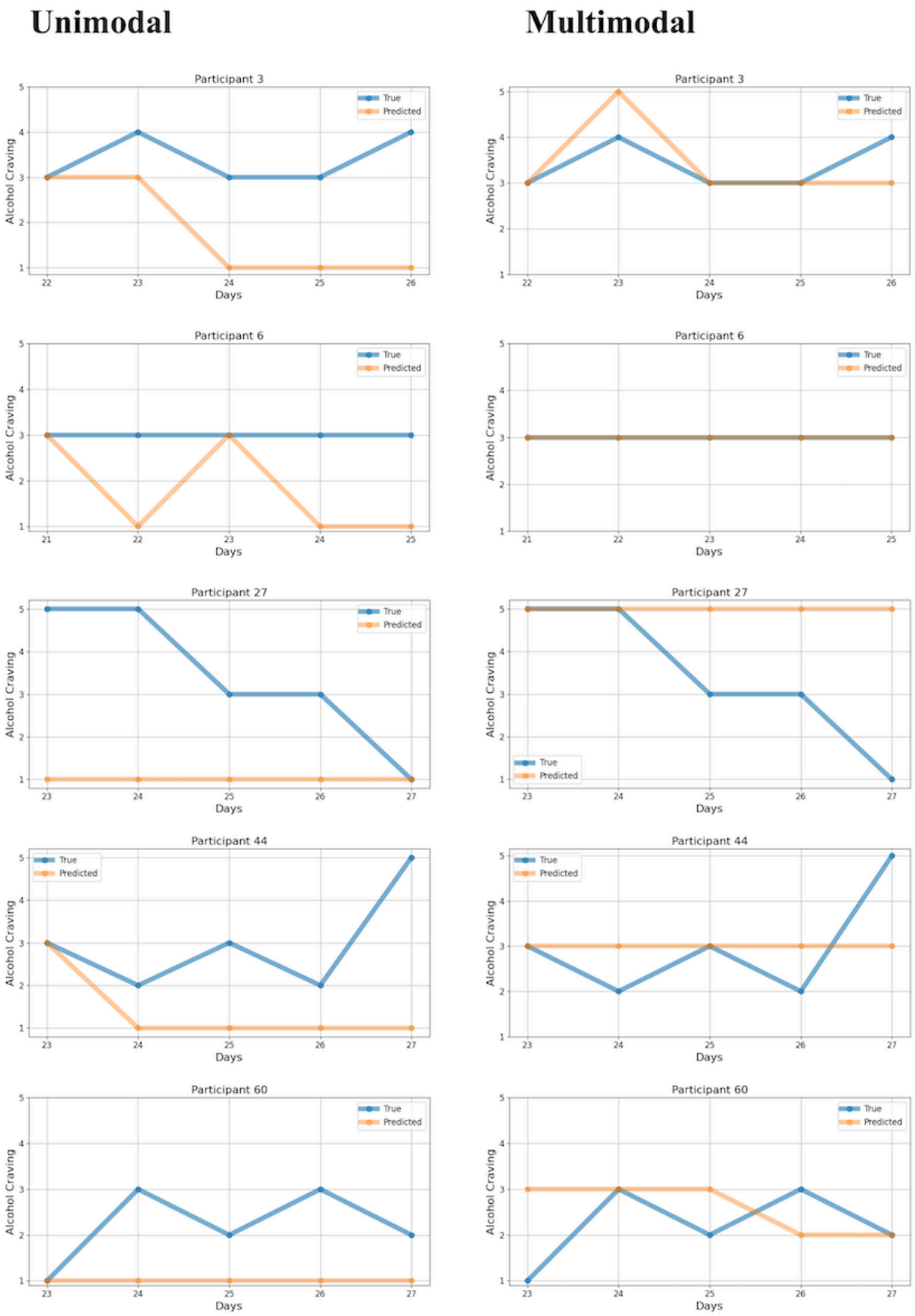
## Predicted Alcohol Craving Values by Both Models

Figure 4 illustrates representative examples comparing predicted alcohol craving levels from the unimodal and multimodal models against the ground truth, respectively. For each prediction time point, the final predicted value was obtained by taking the mode of the 100 model outputs across different random seeds, with ties resolved by selecting the lower class label. Each row illustrates a distinct performance pattern. For participant 3, the multimodal model showed near-perfect tracking, whereas the unimodal model showed consistent deviation. For participant 6, the multimodal model achieved 5 of 5 correct predictions, compared with 2 of 5 for the unimodal model. For participant 27, both models showed comparable overall accuracy, but the multimodal model produced predictions closer to the true ordinal labels. For participant 44, high-craving minority class instances were better captured by the multimodal model. For participant 60, the multimodal model showed substantial accuracy improvement.

Overall, the multimodal model demonstrates more accurate and stable predictions compared to the unimodal model across participants. In several cases, the multimodal model achieves perfect or near-perfect accuracy, whereas the unimodal model exhibits substantial deviations from the true values. Importantly, the advantage of the multimodal model is not limited to overall accuracy. In cases where both models produce incorrect predictions, the multimodal model tends to generate predictions that are closer to the true ordinal label. This behavior is reflected in the lower MAE observed in the multimodal setting and suggests improved sensitivity to the ordinal structure of the target variable.

Furthermore, the multimodal model shows improved performance in capturing the minority class instances, allowing it to better capture less frequent but clinically important states. This is consistent with the observed improvement in the macro $F_1$-score. Although some participants exhibit limited improvement in overall accuracy, the multimodal model still produces predictions that are closer to the ground truth, indicating meaningful improvements beyond simple accuracy gains. Taken together, these results suggest that incorporating static baseline features enables the model to generate more precise, balanced, and clinically informative predictions at the individual level.

**Figure 4.** Representative participant-level comparisons of true and predicted next-day alcohol craving levels for the unimodal and multimodal models across five test days. Blue lines indicate true values, and orange lines indicate predicted values. The left column shows unimodal model predictions, and the right column shows multimodal model predictions. Final predictions were obtained by taking the mode across 100 random seed runs.



# Discussion

## Principal Findings

This study demonstrates that a multimodal AI approach improves the prediction of changes in next-day alcohol cravings over a 28-day period compared to a unimodal model based solely on digital phenotypes. By integrating digital behavioral data with neurophysiological and psychological features, the multimodal model achieved more accurate and consistent predictions across individuals. These findings suggest that static baseline characteristics provide subject-specific contextual information that shapes the interpretation of temporal behavioral patterns. By conditioning temporal predictions on individual-specific features, the model is better able to capture heterogeneity in craving dynamics across participants. Rather than assuming uniform temporal patterns across individuals, the multimodal approach enables more personalized and stable predictions of alcohol craving.

## Comparison to Prior Work

This study represents a novel attempt to integrate digital phenotypes collected through a smartphone app with various modalities using AI methods. A previous review study showed that smartphone-based digital phenotyping studies in AUD primarily use passive sensor data (eg, accelerometer, GPS, and app usage) to capture behavioral patterns such as movement and sociability, and rely mainly on self-reported alcohol use, craving, or mood [75]. Another study explored the association between smartphone-derived behavioral features and alcohol craving using EMAs [76]. That study demonstrated that sensor-derived features, such as location entropy and accelerometer-based activity, were significantly correlated with alcohol craving and mood states in individuals with AUD [76]. By contrast, this study advances this line of research by adopting a multimodal AI approach that enables individualized forecasting of daily alcohol craving. This study suggests that integrating all modalities provides better performance than using any single modality alone. This may be because the integration of these modalities captures diverse aspects of craving. This approach may not only improve predictive performance but also enhance the potential for clinical interventions.

While several approaches have been proposed to integrate static data with time-series data through feature concatenation [62,77], only a limited number of studies have explored incorporating static information by using feed-forward networks to summarize static features and initialize the hidden state of recurrent models [61,78]. While prior clinical studies adopting this framework predominantly rely on long short-term memory (LSTM) or gated recurrent unit (GRU)–based architectures, we demonstrate that a tiny lightweight vanilla RNN can achieve stable predictive performance in a small-sample multimodal setting. Given the short input sequence (window size=4) and limited sample size, we prioritized a lightweight vanilla RNN to minimize model complexity, consistent with the prior work noting that simpler RNN architectures require fewer computational resources and may be suitable for smaller datasets or short-term prediction settings [79]. LSTM architectures were originally developed to address difficulties in learning dependencies over extended temporal intervals [80], whereas gated recurrent architectures introduce additional parameters relative to vanilla RNNs [81]. Prior clinical time-series forecasting work has shown that vanilla RNN cells can remain competitive with LSTM and GRU cells while requiring fewer parameters and computational resources [81]. Accordingly, our primary aim was to evaluate multimodal integration within a deliberately compact recurrent framework rather than to identify the optimal sequence-learning architecture.

## Interpretation of the Findings

Combining all modalities yielded the greatest improvement in predictive performance compared with models based on each unimodality in predicting daily alcohol craving levels. In additional analyses, combining digital phenotype data with EEG or HRV improved model performance compared with using psychological features alone. A previous digital phenotype study showed that a combination of multimodal measurement and an AI approach has enhanced the performance of digital markers in phenotyping and diagnosis of major depressive disorder [82]. Prior work has similarly demonstrated that multimodal approaches outperform single-modality models in the diagnosis of internet gaming disorder and AUD, suggesting that integrating self-report–based neuropsychological measures with objective EEG-based neurophysiological data can enhance clinical assessment [83]. The improved performance of the multimodal AI model likely reflects the inherently multidimensional nature of alcohol craving, encompassing psychological, behavioral, and neurophysiological aspects [8,12,84]. Integrating frequency-specific EEG relative power features that reflect neurophysiological processes associated with craving, autonomic nervous system activity reflected by HRV, and subjective self-reports may provide complementary information that is critical for accurate prediction. Future research should continue to use multimodal frameworks to better account for the complex mechanisms underlying alcohol craving.

Post hoc analyses suggested that greater within-person craving variability and more diverse craving-level distributions were associated with lower participant-level prediction accuracy, whereas EMA completion and the number of usable sequential samples were not significantly associated with accuracy. Participants with greater craving variability and more diverse craving-level distributions may represent a more challenging forecasting problem. However, these associations were exploratory and should not be interpreted as causal.

Recently, just-in-time adaptive interventions have gained increasing attention as an approach to delivering interventions in real time by leveraging dynamically changing information about an individual's state and context [85]. They adapt intervention strategies based on both dynamic data (eg, emotions, behaviors, and context) and stable characteristics (eg, demographics and baseline conditions) [85]. Our model may serve as a predictive component within the just-in-time adaptive interventions by identifying high-risk moments based on dynamic and individual-level data. For example,

when predicted craving exceeds a predefined threshold, the system may trigger personalized interventions such as behavioral prompts or clinician alerts. Given that HRV alone provided substantial performance gains, a cost-effective deployment using wearable devices may be feasible without requiring full EEG acquisition. Although the sample size was relatively small, future studies with larger populations may help further validate our multimodal AI approaches, enhancing its potential for real-world clinical application. Moreover, the use of a tiny RNN to address temporal misalignment across modalities may represent a promising strategy for future studies.

The use of a tiny RNN architecture may have contributed to stable model performance by reducing the risk of overfitting in a relatively small sample size and short sequence settings. While more complex architectures such as LSTM or transformer models may offer additional flexibility, given the limited data and the use of short sliding windows, these models may introduce additional model complexity and a large number of parameters. Despite its simplicity, the present approach was sufficient to capture meaningful temporal dependencies in multimodal data.

## Limitations

Several limitations of this study should be noted. First, the sample size was relatively small to be considered representative of all problematic alcohol drinkers, which may limit the generalizability of the findings. Therefore, further investigations with larger sample sizes should be warranted. Second, this study was conducted in a single South Korean cohort without external validation using an independent dataset. This may limit the generalizability of the findings to other populations and settings. Future studies should therefore validate the model in independent and more diverse cohorts. Third, alternative sequence-learning and machine-learning approaches, including LSTM, GRU, temporal convolutional networks, and gradient-boosting methods, were not systematically evaluated in this study. We intentionally prioritized a lightweight vanilla RNN given the short temporal window and limited sample size; however, direct comparisons across alternative model classes are needed to determine whether the observed multimodal benefits generalize across architectures. Fourth, although only 29 (48.3%) of 60 participants provided EMA responses for exactly 28 days, all participants in the final analytic sample contributed at least 21 days of EMA data (range 21-30 days). Within this observed range, EMA response days were not significantly associated with participant-level accuracy. Therefore, this study could not determine a lower minimum EMA completion threshold, such as 8 days, required for satisfactory prediction performance. Fifth, the overall predictive performance was modest, as reflected in the accuracy and weighted and macro $F_1$-scores, indicating the need for further model optimization and validation. Sixth, next-day forecasting may not fully capture individual and circadian variability in craving dynamics; therefore, future studies should consider more frequent assessments and alternative forecasting intervals [16].

## Future Directions

Validation using larger and more diverse cohorts is necessary to enhance the generalizability of the model across broader populations of individuals with problematic alcohol use, particularly by incorporating heterogeneous clinical profiles and varying levels of craving variability. Second, future research should focus on improving model performance for individuals with high intra-individual variability in craving, potentially through advanced modeling approaches that better capture temporal dynamics, such as sophisticated sequence models or attention-based architectures.

Finally, integrating predicted high-risk moments of craving with timely and personalized interventions—such as alerts, behavioral strategies, or therapeutic support—may enable the model to serve as a core component of a digital therapeutic system for preventing alcohol use and relapse. Building on this study, future work will focus on developing an app-based intervention program for alcohol craving regulation that uses the predictive model to deliver real-time personalized support.

## Conclusion

This study attempts a novel approach to predicting daily alcohol craving in individuals with problematic alcohol use using AI algorithms that integrate digital phenotypes and neurophysiological and psychological data. The multimodal model outperformed single-modality approaches in predicting daily craving levels. Notably, integrating all modalities showed better performance than models combining digital phenotypes with either EEG, HRV, or psychological features. This study emphasizes the importance of multimodal approaches for daily alcohol craving prediction and suggests that these approaches may enable timely interventions to prevent problematic alcohol use.

### Acknowledgments

The authors sincerely appreciate all participants for their time and involvement in this study.

### Generative AI Use

The authors used ChatGPT (OpenAI) to assist with language editing, refinement of figure captions, and formatting checks during manuscript preparation. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

## Funding

This work was supported by the National Research Foundation of Korea (NRF; RS-2024-00435727 and RS-2025-00516410 [to WYA], RS-2022-KH125035 [to YCJ, WYA, and JSC]), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT; RS-2026-25507282; to WYA), the Korea Mental Health R&D Project, funded by the Ministry of Health & Welfare, Republic of Korea (HI22C0404; to JSC), the Seoul National University Artificial Intelligence Graduate School Program (RS-2021-II211343; to WYA), K-NeuroMind NEXUS: Neurocognitive-Affective Next-Generation Cross-modal Foundation Platform (RS-2026-25518987; to WYA), and the BK21 FOUR Program (5199990314123; to WYA). The funders had no role in study design, data collection, data analysis, interpretation of the data, or preparation of the manuscript.

## Data Availability

The data supporting the findings of this study are not publicly available due to privacy and ethical restrictions but are available from the corresponding author upon reasonable request and subject to institutional approval. The analytical code for data preprocessing, feature selection, model training, and evaluation is publicly available on GitHub [86].

## Authors' Contributions

DL and JYL contributed equally to this work. WYA and JSC are co-corresponding authors of this work. DL and JYL were responsible for conceptualization, data curation, methodology, investigation, visualization, writing–original draft, and writing–review and editing. HL, MK, and SS contributed to investigation, data curation, and validation. JY performed data curation and resources. YCJ was responsible for conceptualization, funding acquisition, validation, and writing–review and editing. WYA contributed to funding acquisition, methodology, project administration, resources, supervision, validation, and writing–review and editing. JSC contributed to conceptualization, methodology, project administration, resources, funding acquisition, supervision, validation, and writing–review and editing.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Supplementary materials.
[DOCX File (Microsoft Word File), 502 KB-Multimedia Appendix 1]

## References

1. World Health Organization. WHO fact sheet on alcohol. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/alcohol
2. Haber PS. Identification and treatment of alcohol use disorder. N Engl J Med. 2025;392(3):258-266. [doi: 10.1056/NEJMra2306511] [Medline: 39813644]
3. Wilson J, Tanuseputro P, Myran DT, et al. Characterization of Problematic Alcohol Use Among Physicians: A Systematic Review. JAMA Netw Open. Dec 1, 2022;5(12):e2244679. [doi: 10.1001/jamanetworkopen.2022.44679] [Medline: 36484992]
4. Kaner E, Campbell C, Pienaar ED, Heather N, Schlesinger C, Saunders J. Brief interventions for excessive drinkers in primary care health settings. Cochrane Database Syst Rev. 2003;2. [doi: 10.1002/14651858.CD004148]
5. Tuchman FR, Hallgren KA, Richards DK, et al. Reductions in WHO risk drinking levels correlate with alcohol craving among individuals with alcohol use disorder. Alcohol Clin Exp Res. 2024;48(2):420-429. [doi: 10.1111/acer.15257]
6. Addolorato G, Leggio L, Abenavoli L, Gasbarrini G, Alcoholism Treatment Study Group. Neurobiochemical and clinical aspects of craving in alcohol addiction: a review. Addict Behav. 2005;30(6):1209-1224. [doi: 10.1016/j.addbeh.2004.12.011] [Medline: 15925129]
7. Tiffany ST, Conklin CA. A cognitive processing model of alcohol craving and compulsive alcohol use. Addiction. 2000;95(8s2):145-153. [doi: 10.1046/j.1360-0443.95.8s2.3.x]
8. Yoon G, Thompson K, Hakes JK, Westermeyer J, Petrakis IL, Kim SW. Alcohol craving and psychiatric disorders among current drinkers. Am J Addict. 2021;30(1):34-42. [doi: 10.1111/ajad.13083] [Medline: 32667738]
9. Flannery BA, Poole SA, Gallop RJ, Volpicelli JR. Alcohol craving predicts drinking during treatment: an analysis of three assessment instruments. J Stud Alcohol. Jan 2003;64(1):120-126. [doi: 10.15288/jsa.2003.64.120] [Medline: 12608492]
10. Oslin DW, Cary M, Slaymaker V, Colleran C, Blow FC. Daily ratings measures of alcohol craving during an inpatient stay define subtypes of alcohol addiction that predict subsequent risk for resumption of drinking. Drug Alcohol Depend. 2009;103(3):131-136. [doi: 10.1016/j.drugalcdep.2009.03.009] [Medline: 19443131]
11. McHugh RK, Fitzmaurice GM, Griffin ML, Anton RF, Weiss RD. Association between a brief alcohol craving measure and drinking in the following week. Addiction. 2016;111(6):1004-1010. [doi: 10.1111/add.13311] [Medline: 26780476]

12. Kavanagh DJ, Statham DJ, Feeney GFX, et al. Measurement of alcohol craving. Addict Behav. 2013;38(2):1572-1584. [doi: 10.1016/j.addbeh.2012.08.004] [Medline: 23142210]
13. Marsch LA. Digital health data-driven approaches to understand human behavior. Neuropsychopharmacology. 2021;46(1):191-196. [doi: 10.1038/s41386-020-0761-5] [Medline: 32653896]
14. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA. 2017;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
15. Liu S, Heinzel S, Dolan RJ. Digital phenotyping and mobile sensing in addiction psychiatry. Pharmacopsychiatry. 2021;54(6):287-288. [doi: 10.1055/a-1559-4003] [Medline: 34407554]
16. Hisler GC, Pedersen SL, Hasler BP. The 24-hour rhythm in alcohol craving and individual differences in sleep characteristics and alcohol use frequency. Alcohol Clin Exp Res. 2022;46(6):1084-1093. [doi: 10.1111/acer.14826] [Medline: 35383960]
17. Kim J, Nam S, Kim DH, et al. Frontal EEG response to alcohol craving elicited by individually tailored video cues. Alcohol. 2023;112:1-7. [doi: 10.1016/j.alcohol.2023.05.005] [Medline: 37225110]
18. Huang Y, Mohan A, De Ridder D, Sunaert S, Vanneste S. The neural correlates of the unified percept of alcohol-related craving: a fMRI and EEG study. Sci Rep. 2018;8(1):923. [doi: 10.1038/s41598-017-18471-y]
19. Quintana DS, Guastella AJ, McGregor IS, Hickie IB, Kemp AH. Heart rate variability predicts alcohol craving in alcohol dependent outpatients: further evidence for HRV as a psychophysiological marker of self-regulation. Drug Alcohol Depend. 2013;132(1-2):395-398. [doi: 10.1016/j.drugalcdep.2013.02.025] [Medline: 23601925]
20. Cohen AS, Cox CR, Masucci MD, et al. Digital phenotyping using multimodal data. Curr Behav Neurosci Rep. 2020;7(4):212-220. [doi: 10.1007/s40473-020-00215-4]
21. Camacho E, Brady RO, Lizano P, Keshavan M, Torous J. Advancing translational research through the interface of digital phenotyping and neuroimaging: a narrative review. Biomarkers in Neuropsychiatry. 2021;4:100032. [doi: 10.1016/j.bionps.2021.100032]
22. Tendulkar M, Tendulkar R. Amalgamating artificial intelligence & neuropsychiatry: multimodal approach for precision diagnosis. Asian J Psychiatr. 2026;115:104795. [doi: 10.1016/j.ajp.2025.104795] [Medline: 41435526]
23. Topiwala A, Allan CL, Valkanova V, et al. Moderate alcohol consumption as risk factor for adverse brain outcomes and cognitive decline: longitudinal cohort study. BMJ. 2017;357:j2353. [doi: 10.1136/bmj.j2353] [Medline: 28588063]
24. Topiwala A, Taschler B, Ebmeier KP, et al. Alcohol consumption and telomere length: Mendelian randomization clarifies alcohol's effects. Mol Psychiatry. 2022;27(10):4001-4008. [doi: 10.1038/s41380-022-01690-9] [Medline: 35879401]
25. Nunes PT, Kipp BT, Reitz NL, Savage LM. Aging with alcohol-related brain damage: critical brain circuits associated with cognitive dysfunction. Int Rev Neurobiol. 2019;148:101-168. [doi: 10.1016/bs.irn.2019.09.002] [Medline: 31733663]
26. Pfefferbaum A, Sullivan EV, Pohl KM, et al. Brain volume in fetal alcohol spectrum disorders over a 20-year span. JAMA Netw Open. 2023;6(11):e2343618. [doi: 10.1001/jamanetworkopen.2023.43618] [Medline: 37976065]
27. Mumtaz W, Vuong PL, Xia L, Malik AS, Rashid RBA. An EEG-based machine learning method to screen alcohol use disorder. Cogn Neurodyn. 2017;11(2):161-171. [doi: 10.1007/s11571-016-9416-y] [Medline: 28348647]
28. Mumtaz W, Saad M, Kamel N, Ali SSA, Malik AS. An EEG-based functional connectivity measure for automatic detection of alcohol use disorder. Artif Intell Med. 2018;84:79-89. [doi: 10.1016/j.artmed.2017.11.002] [Medline: 29169647]
29. Singh G, Campbell EM, Hogeveen J, Witkiewitz K, Claus ED, Cavanagh JF. Affective imagery boosts the reward related delta power in hazardous drinkers. Psychiatry Res Neuroimaging. 2023;334:111685. [doi: 10.1016/j.pscychresns.2023.111685] [Medline: 37506424]
30. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. American Psychiatric Association; 2013.
31. Cooper ML. Motivations for alcohol use among adolescents: development and validation of a four-factor model. Psychol Assess. 1994;6(2):117-128. [doi: 10.1037/1040-3590.6.2.117]
32. Shin HW, Han SY. A preliminary study for developing drinking motives scale. Korean Psychol J Cult Soc Issues. 1999;5(1):77-92.
33. Saunders JB, Aasland OG, Babor TF, de la Fuente JR, Grant M. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative project on early detection of persons with Harmful Alcohol Consumption--II. Addiction. 1993;88(6):791-804. [doi: 10.1111/j.1360-0443.1993.tb02093.x] [Medline: 8329970]
34. Joe K, Chai S, Park A, Lee H, Shin I, Min S. Optimum cut-off score for screening of hazardous drinking using the Korean version of Alcohol use Disorder Identification Test (AUDIT-K). J Korean Acad Addict Psychiatry. 2009;13(1):34-40.

35. Beck AT, Steer RA, Brown G. Beck Depression Inventory–II. In: Psychological Assessment. APA PsycNet; 1996. [doi: 10.1037/t00742-000]
36. Lim SY, Lee EJ, Jeong SW, Kim HC, Jeong CH, Jeon TY, et al. The validation study of Beck Depression Scale 2 in Korean version. Anxiety and mood. 2011;7(1):48-53.
37. Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: psychometric properties. J Consult Clin Psychol. Dec 1988;56(6):893-897. URL: https://doi.org/10.1037/0022-006X.56.6.893 [doi: 10.1037//0022-006x.56.6.893] [Medline: 3204199]
38. Yook SP, Kim ZS. A clinical study on the Korean version of Beck Anxiety Inventory: comparative study of patient and non-patient. Korean J Clin Psychol. 1997;16(1):185-197.
39. Patton JH, Stanford MS, Barratt ES. Factor structure of the Barratt impulsiveness scale. J Clin Psychol. 1995;51(6):768-774. [doi: 10.1002/1097-4679(199511)51:63.0.CO;2-1] [Medline: 8778124]
40. Heo SY, Oh JY. The Korean version of the Barratt Impulsiveness Scale, 11th version: its reliability and validity. Korean J Psychol Gen. 2012;31(3):769-782.
41. Buysse DJ, Reynolds CF III, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res. 1989;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
42. Sohn SI, Kim DH, Lee MY, Cho YW. The reliability and validity of the Korean version of the Pittsburgh Sleep Quality Index. Sleep Breath. 2012;16(3):803-812. [doi: 10.1007/s11325-011-0579-9] [Medline: 21901299]
43. Miller WR, Tonigan JS. Assessing Drinkers' Motivation for Change: The Stages of Change Readiness and Treatment Eagerness Scale (SOCRATES). Washington, DC, US: American Psychological Association; 1997:355-369. [doi: 10.1037/10248-014] ISBN: 1-55798-468-9
44. Chun YM. Assessing alcohol dependents' motivation for change: the development study on the Korean version of the stages of change readiness and treatment eagerness scale. Korean J Clin Psychol. 2005;24(1):207-224.
45. Hwang S, Kim J, Park G, Chey J, Hong S. K-WAIS-IV Administration and Scoring Manual. Korean Psychology; 2012.
46. Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C. MEG and EEG data analysis with MNE-Python. Front Neurosci. 2013;7. [doi: 10.3389/fnins.2013.00267]
47. Jas M, Engemann DA, Bekhti Y, Raimondo F, Gramfort A. Autoreject: automated artifact rejection for MEG and EEG data. Neuroimage. 2017;159:417-429. [doi: 10.1016/j.neuroimage.2017.06.030] [Medline: 28645840]
48. Yao P, Liu M, Cheng X, Shi F, Li H, Liu X, et al. An end-to-end model for time series classification in the presence of missing values. arXiv. Preprint posted online on Aug 11, 2024. URL: https://doi.org/10.48550/arXiv.2408.05849
49. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: A data perspective. ACM computing surveys (CSUR). 2017;50(6):1-45. [doi: 10.1145/3136625]
50. Tsai CF, Sung YT. Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches. Knowl Based Syst. Sep 2020;203:106097. [doi: 10.1016/j.knosys.2020.106097]
51. Li S, Harner EJ, Adjeroh DA. Random KNN feature selection - a fast and stable alternative to Random Forests. BMC Bioinformatics. Nov 18, 2011;12(1):450. [doi: 10.1186/1471-2105-12-450] [Medline: 22093447]
52. Bontempi G, Ben Taieb S, Le Borgne YA. Machine Learning Strategies for Time Series Forecasting European Big Data Management and Analytics Summer. Springer; 2012:62-77. [doi: 10.1007/978-3-642-36318-4_3]
53. Hyndman RJ, Athanasopoulos G. Forecasting: Principles and Practice. OTexts; 2018. ISBN: 0987507117
54. Busk J, Faurholt-Jepsen M, Frost M, Bardram JE, Vedel Kessing L, Winther O. Forecasting mood in bipolar disorder from smartphone self-assessments: hierarchical Bayesian approach. JMIR Mhealth Uhealth. 2020;8(4):e15028. [doi: 10.2196/15028] [Medline: 32234702]
55. Tong S, Yuan J. Efficiently enhancing long-term series forecasting via adaptive lookback with wavelets. AAAI. 2026;40(31):25966-25974. [doi: 10.1609/aaai.v40i31.39797]
56. de Amorim LBV, Cavalcanti GDC, Cruz RMO. The choice of scaling technique matters for classification performance. Appl Soft Comput. 2023;133:109924. [doi: 10.1016/j.asoc.2022.109924]
57. Thara DK, PremaSudha BG. Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. Pattern Recognit Lett. 2019;128:544-550. [doi: 10.1016/j.patrec.2019.10.029]
58. Kim YS, Kim MK, Fu N, Liu J, Wang J, Srebric J. Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models. Sustainable Cities and Society. 2025;118:105570. [doi: 10.1016/j.scs.2024.105570]
59. Yi K, Zhang Q, Fan W, et al. A survey on deep learning based time series analysis with frequency transformation. Presented at: KDD '25; Toronto ON Canada. Aug 3, 2025.URL: https://dl.acm.org/doi/proceedings/10.1145/3711896 [doi: 10.1145/3711896.3736571]

60. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. 2018;8(1):6085. [doi: 10.1038/s41598-018-24271-9] [Medline: 29666385]
61. Miebs G, Mochol-Grzelak M, Karaszewski A, Bachorz RA. Efficient strategies of static features incorporation into the recurrent neural network. Neural Process Lett. 2020;51(3):2301-2316. [doi: 10.1007/s11063-020-10195-x]
62. Esteban C, Staeck O, Baier S, Yang Y, Tresp V. Predicting clinical events by combining static and dynamic information using recurrent neural networks. Presented at: 2016 IEEE International Conference on Healthcare Informatics (ICHI); Chicago, IL, USA. [doi: 10.1109/ICHI.2016.16]
63. Stérin T, Farrugia N, Gripon V. An intrinsic difference between vanilla RNNs and GRU models. Cognitive. 2017;84. URL: https://www.thinkmind.org/articles/cognitive_2017_5_50_48008.pdf [Accessed 2026-11-06]
64. Nair V, Hinton GE, editors. Rectified linear units improve restricted boltzmann machines. Presented at: 27th International Conference on International Conference on Machine Learning; Jun 21-24, 2010; Haifa, Israel. 2010.URL: https://dl.acm.org/doi/10.5555/3104322.3104425 [Accessed 2026-11-06]
65. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. Presented at: 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019; Vancouver, BC. 2019.URL: https://dl.acm.org/doi/abs/10.5555/3454287.3455008 [Accessed 2026-11-06]
66. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online on Dec 22, 2014. URL: https://doi.org/10.48550/arXiv.1412.6980 [Accessed 2026-11-06]
67. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6. [doi: 10.1186/s12864-019-6413-7] [Medline: 31898477]
68. Ghanem M, Ghaith AK, El-Hajj VG, et al. Limitations in evaluating machine learning models for imbalanced binary outcome classification in spine surgery: a systematic review. Brain Sci. 2023;13(12):1723. [doi: 10.3390/brainsci13121723] [Medline: 38137171]
69. Wegier W, Ksieniewicz P. Application of imbalanced data classification quality metrics as weighting methods of the ensemble data stream classification algorithms. Entropy (Basel). 2020;22(8):849. [doi: 10.3390/e22080849] [Medline: 33286620]
70. Manning CD. Introduction to Information Retrieval. Syngress Publishing; 2008.
71. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45(4):427-437. [doi: 10.1016/j.ipm.2009.03.002]
72. Mandl T, Modha S, Majumder P, Patel D, Dave M, Mandlia C. Overview of the HASOC track at FIRE 2019: hate speech and offensive content identification in Indo-European languages. Presented at: 11th Annual Meeting of the Forum for Information Retrieval Evaluation; Dec 12-15, 2019; Kolkata, India. 2019.URL: https://dl.acm.org/doi/10.1145/3368567.3368584 [Accessed 2026-11-06]
73. Opitz J, Burst S. Macro F1 and macro F1. arXiv. Preprint posted online on Nov 8, 2019. URL: https://doi.org/10.48550/arXiv.1911.03347 [Accessed 2026-11-06]
74. Gaudette L, Japkowicz N, editors. Evaluation methods for ordinal classification. In: Canadian Conference on Artificial Intelligence. Springer; 2009. [doi: 10.1007/978-3-642-01818-3_25]
75. Dumas A, Hokayem J, Goodman G, Venkatasubramanian K, Chai P. Smartphone-based digital phenotyping across health conditions: scoping review. J Med Internet Res. 2026;28:e84146. [doi: 10.2196/84146] [Medline: 41877492]
76. Wu T, Sherman G, Giorgi S, et al. Smartphone sensor data estimate alcohol craving in a cohort of patients with alcohol-associated liver disease and alcohol use disorder. Hepatol Commun. 2023;7(12):e0329. [doi: 10.1097/HC9.0000000000000329] [Medline: 38055637]
77. Che Z, Purushotham S, Khemani R, Liu Y, editors. Interpretable deep models for ICU outcome prediction. Presented at: AMIA Annual Symposium Proceedings; Nov 7-11, 2026; Dallas, TX. [Medline: 28269832]
78. Li D, Lyons P, Klaus J, Gage B, Kollef M, Lu C. Integrating static and time-series data in deep recurrent models for oncology early warning systems. Presented at: CIKM '21: The 30th ACM International Conference on Information and Knowledge Management; Nov 1-5, 2021; Virtual event in Queensland, Australia. 2021.URL: https://dl.acm.org/doi/10.1145/3459637.3482441 [Accessed 2026-11-06] [doi: 10.1145/3459637.3482441]
79. Waqas M, Humphries UW. A critical review of RNN and LSTM variants in hydrological time series predictions. MethodsX. Dec 2024;13:102946. [doi: 10.1016/j.mex.2024.102946] [Medline: 39324077]
80. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
81. Zhu T, Li K, Chen J, Herrero P, Georgiou P. Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. J Healthc Inform Res. 2020;4(3):308-324. [doi: 10.1007/s41666-020-00068-2] [Medline: 35415447]
82. Chen J, Chan NY, Li CT, et al. Multimodal digital assessment of depression with actigraphy and app in Hong Kong Chinese. Transl Psychiatry. 2024;14(1):150. [doi: 10.1038/s41398-024-02873-4] [Medline: 38499546]

83. Lee JY, Song MS, Yoo SY, et al. Multimodal-based machine learning approach to classify features of internet gaming disorder and alcohol use disorder: A sensor-level and source-level resting-state electroencephalography activity and neuropsychological study. Compr Psychiatry. 2024;130:152460. [doi: 10.1016/j.comppsych.2024.152460] [Medline: 38335572]
84. Seo D, Sinha R. Chapter 21 - the neurobiology of alcohol craving and relapse. In: Sullivan EV, Pfefferbaum A, editors. Handbook of Clinical. Elsevier; 2014:355-368. [doi: 10.1016/B978-0-444-62619-6.00021-5]
85. Nahum-Shani I, Murphy SA. Just-in-time adaptive interventions: where are we now and what is next? Annu Rev Psychol. 2026;77(1):679-703. [doi: 10.1146/annurev-psych-121024-044244] [Medline: 40939059]
86. Forecasting next-day alcohol craving in individuals with problematic alcohol use drinkers with multimodal data and a tiny recurrent neural network: prospective longitudinal study. GitHub. URL: https://github.com/CCS-Lab/Project_Multimodal_Public/ [Accessed 2026-11-06]

## Abbreviations

**AUD:** alcohol use disorder
**AUDIT:** Alcohol Use Disorders Identification Test
**BAI:** Beck Anxiety Inventory
**BDI-II:** Beck Depression Inventory-II
**BIS:** Barratt Impulsiveness Scale
**DMQ:** Drinking Motives Questionnaire
**EEG:** electroencephalography
**EMA:** ecological momentary assessment
**GRU:** gated recurrent unit
**HRV:** heart rate variability
**LSTM:** long short-term memory
**MAE:** mean absolute error
**ML:** machine learning
**PSQI:** Pittsburgh Sleep Quality Index
**RNN:** recurrent neural network
**SOCRATES:** Stages of Change Readiness and Treatment Eagerness Scale

*Edited by Taiane de Azevedo Cardoso; peer-reviewed by Md Zakir Hossain, Tammy Chung; submitted 21.May.2026; final revised version received 23.Sep.2026; accepted 07.Oct.2026; published 11.Nov.2026*

*Please cite as:*
*Lee D, Lee JY, Lee H, Kwon M, Song S, Yoon J, Jung YC, Ahn WY, Choi JS*
*Forecasting Next-Day Alcohol Craving in Individuals With Problematic Alcohol Use With Multimodal Data and a Tiny Recurrent Neural Network: Prospective Longitudinal Study*
*J Med Internet Res 2026;28:e101973*
*URL: https://www.jmir.org/2026/1/e101973*
*doi: 10.2196/101973*

© Dongwon Lee, Ji-Yoon Lee, Hyeonmin Lee, Mina Kwon, Suyeon Song, Jeong-hyun Yoon, Young-Chul Jung, Woo-Young Ahn, Jung-Seok Choi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.