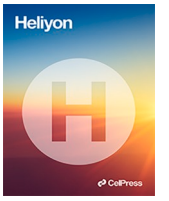
Heliyon 10 (2024) e33485

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Exploring the relationship between response time sequence in scale answering process and severity of insomnia: A machine learning approach

Zhao Su [a,b,1], Rongxun Liu [a,c,1], Keyin Zhou [a], Xinru Wei [a,b], Ning Wang [a,d], Zexin Lin [a], Yuanchen Xie [e], Jie Wang [b], Fei Wang [a,**], Shenzhong Zhang [b,***], Xizhe Zhang [b,*]

[a] Early Intervention Unit, Department of Psychiatry, The Affiliated Brain Hospital of Nanjing Medical University, Nanjing, Jiangsu, China
[b] School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, Jiangsu, China
[c] School of Psychology, Xinxiang Medical University, Xinxiang, Henan, China
[d] School of Public Health, Xinxiang Medical University, Xinxiang, Henan, China
[e] The Fourth School of Clinical Medicine, Nanjing Medical University, Nanjing, Jiangsu, China

ARTICLE INFO

Keywords:
Response time
Machine learning
Insomnia
Behavioral data

ABSTRACT

Utilizing computer-based scales for cognitive and psychological evaluations allows for the collection of objective data, such as response time. This cross-sectional study investigates the significance of response time data in cognitive and psychological measures, with a specific focus on its role in evaluating sleep quality through the Insomnia Severity Index (ISI) scale. A mobile application was designed to administer scale tests and collect response time data from 2729 participants. We explored the relationship between symptom severity and response time. A machine learning model was developed to predict the presence of insomnia symptoms in participants using response time data. The result revealed a statistically significant difference ($p < 0.01$) in the total response time between participants with or without insomnia symptom. Furthermore, a strong correlation was observed between the severity of specific insomnia aspects and the response times at the individual questions level. The machine learning model demonstrated a high predictive Area Under the ROC Curve (AUROC) of 0.824 in predicting insomnia symptoms based on response time data. These findings highlight the potential utility of response time data to evaluate cognitive and psychological measures.

1. Introduction

Insomnia is a prevalent sleep disorder that affects millions of people worldwide, leading to significant impairments in daily functioning, quality of life, and overall health [1]. Insomnia is characterized by difficulties initiating or maintaining sleep, early

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: fei.wang@yale.edu (F. Wang), zsz@njmu.edu.cn (S. Zhang), zhangxizhe@njmu.edu.cn (X. Zhang).
[1] The two authors contributed equally to this work: Zhao Su, Rongxun Liu.

https://doi.org/10.1016/j.heliyon.2024.e33485
Received 3 March 2024; Received in revised form 7 June 2024; Accepted 21 June 2024
Available online 25 June 2024

morning awakenings, and non-restorative sleep [1]. Current diagnostic methods for insomnia primarily rely on subjective reports from patients, sleep diaries, and clinical interviews, objective measures like actigraphy or polysomnography are not yet part of the routine diagnostic canon [2,3]. However, these methods face challenges such as the lack of standardized diagnostic criteria, the subjective nature of self-reported symptoms, and the potential underdiagnosis of insomnia in primary care settings [4,5]. For instance, the Insomnia Severity Index (ISI) scale, employed in this study, may yield divergent assessments of insomnia severity and symptoms even when two individuals report identical scores. With the advent of information technology, participants can conveniently answer assessment scales on their personal device [6], while their behavioral data during the answering process can simultaneously be collected [7]. This behavioral data can offer an objective perspective on the participant's state during the scale completion, and could significantly assist in the interpretation of scale results [8].

Response Time (RT), which represents the duration taken by a participant to respond to a stimulus or complete a task, is a meaningful metric commonly used in cognitive psychology experiments [9]. In cognitive psychology research, RT functions as a dependent variable, influenced by manipulations of independent variables like stimulus exposure duration [10]. It shares a relationship with response accuracy (another primary dependent variable), because participants can often trade off speed for increased accuracy, or conversely vice versa [11]. However, it's important to highlight that RT and accuracy often serve divergent objectives in these studies. The RT used in this study differs from that in traditional psychology experiments. Instead of measuring the RT to a singular event, it records the timing of a sequence of responses, specifically during the process of answering a scale. The data is unobtrusively collected, indicating that it can be used as an objective biomarker for identifying insomnia symptoms. For example, RT can mirror complex thought processes, thereby acting as a reflection of an individual's cognitive abilities [9].

Insomnia, a prevalent sleep disorder, can induce daytime fatigue and decrease cognitive functioning [12]. Several studies have analyzed the cognitive correlate of insomnia or poor sleep quality [12–15]. Additionally, research has identified changes in executive functions associated with insomnia [12,16]. In study [17], a negative correlation was found between insomnia symptoms and performance on the WAIS-IV coding task, highlighting the relationship between insomnia and response speed. These effects might change how quickly a person responds to things. Given this connection, RT measures could serve as a complementary metric to the ISI results, potentially enhancing the scale's accuracy and reliability. Furthermore, RT can independently indicate insomnia severity, thereby contributing another dimension of objectivity to the assessment process.

An expanding body of research recognizes the potential utility of RT data. For instance, researchers have employed RT as a metric for understanding human information processing [11]. Previous studies have examined the correlation between RT and cognitive load, suggesting that extended RTs may signify more complex cognitive processes [18]. In psychological assessments, RTs can signify indecision or emotional conflict, offering a deeper understanding of individual responses [18,19]. An emergent research trend investigates the relationship between mobile phone usage behavior and mental health, including insomnia symptoms [20,21]. However, further research is needed to fully understand the utility and limitations of RT data in these contexts. In this study, we gathered RT data for each question during the participants' response process. The time interval sequences often encapsulate a wealth of information. Prior research has utilized RT data to filter out invalid responses [22–25], underscoring the value of RT as a metric for assessing participants' cognitive and psychological states during different evaluations.

Several studies have explored the relationship between RTs and survey responses. Ferrando and Lorenzo-Seva proposed a measurement model that incorporates RT into the factor-analytic model for graded or continuous items, potentially increasing the accuracy of individual trait estimates [26]. Borger found that longer RTs decreased choice randomness in an online choice experiment, suggesting a link between RT and cognitive effort [22]. In the context of depression assessment, Chung et al. investigated RT as an implicit self-schema indicator and found an inverted U-shaped relationship between depression severity and RTs to depressive symptom items [8]. Iwata et al. also examined RTs in a computerized adaptive testing system for assessing depressive levels and found that RTs to some items significantly correlated with the estimated trait levels [27].

Furthermore, RT data has been used to enhance the prediction of various outcomes. For example, Nock and Banaji demonstrated that a brief RT measure could accurately predict current suicide ideation and attempt status, as well as future suicide ideation, incrementally improving prediction above and beyond known risk factors [28]. Wang et al. explored the use of both item responses and RT information in multidimensional health measurement and found that adding RT information helped reduce the standard error of patients' multidimensional latent trait estimates [29]. In a recent study, Baba and Bunji developed a machine learning model to predict students' mental health problems using data from an annual student health survey, including item responses and answering time. They compared various machine learning models, such as logistic regression, random forest, XGBoost, and LightGBM, and found that the LightGBM model achieved adequate performance in predicting mental health problems [30]. Building upon these findings, we utilize machine learning techniques to construct models that predict final scale outcomes using RT data. Machine learning methods are adept at uncovering latent patterns within data, making them particularly suitable for this task. This data can contribute to a more precise understanding of an individual's mental state and self-perception, ultimately enhancing the accuracy of diagnoses and treatment plans.

The primary objective of this study is to elucidate the relationship between RT and the results of ISI scale assessments. To probe this relationship, we employ regression analysis to evaluate how RT correlates with the responses on the ISI scale. We develop machine learning models capable of predicting the presence of insomnia symptoms using RT data. Additionally, we conduct a comprehensive analysis of the most impactful features in these predictive models to gain insights into their relative importance and interactions that influence prediction outcomes. Our research explores the utility of RT data as a novel tool for evaluating cognitive and psychological measures, particularly in the context of insomnia symptoms. Ultimately, this study provides preliminary evidence for the potential utility of RT data in enhancing cognitive and psychological assessment practices and contributing to more accurate diagnoses and effective treatment strategies for insomnia.

2. Materials and methods

This study is cross-sectional observational research aimed at exploring the relationship between the behaviors of subjects during questionnaire completion and the final outcomes of the scale. Our sample consisted of 2729 newly enrolled university students. We employed regression analysis to investigate the correlation between behavioral data and scale responses. Furthermore, we utilized machine learning techniques to examine which specific behaviors during the scale answering process are associated with the outcomes of the scale. This study was approved by the Ethics Committee of Nanjing Medical University (Approval No. (2022)793).

2.1. Data acquisition

The methodology of this study revolves around a comprehensive psychological screening program designed specifically for new enrollees at Nanjing Medical University. All the participants were newly enrolled first-year university students, encompassing all new enrollees for the academic year. A substantial sample size was employed, involving a total of 2729 participants. The average age in this group is 19, with a standard deviation of 0.93. Data collection was conducted within the university campus during the fall semester of 2022. Each participant interacted with a dedicated application on their own devices, responding to the 7 items of the ISI scale. Additionally, all participants were given the freedom to choose their preferred time and location for completing the questionnaire online.

In the present study, we utilized the Chinese version of the ISI. The psychometric properties of the Chinese version of the ISI have been previously validated, with a reported Cronbach's alpha coefficient of 0.83 for the total scale [31,32]. The ISI comprises seven questions, each with five response options ranging from 0 to 4 points, where higher scores denote increased severity of insomnia symptoms. The total score is computed by summing the scores from all seven questions, and participants garnering a total score of 8 or higher were deemed to exhibit insomnia symptoms.

Ultimately, there are 15 values per participant recorded in the dataset: 7 categorical variables and 8 continuous variables. For each of the 7 questions in the ISI scale, the chosen response was recorded as a categorical variable. Additionally, the RT for each question was measured in seconds and recorded as a continuous variable. The total ISI score, calculated by summing the scores from all seven questions, was also treated as a continuous variable.

2.2. Response time

The scale was administered via a mobile platform. When a participant moves into a new question, a timestamp is recorded, and another timestamp is taken when the participant selects an answer. The difference between these two timestamps equates to the value of the RT. Fig. 1 illustrates the process of generating the RT sequence. The data collection module in our program is fully implemented at the front end. Moreover, all students complete the questions on campus, sharing the same geographical location. These measures ensure that any variability in RT due to extraneous factors is minimized.

The procedure generated a sequence of seven RTs for each participant. Additionally, the system recorded the chosen response for each question, facilitating the calculation of the total score.

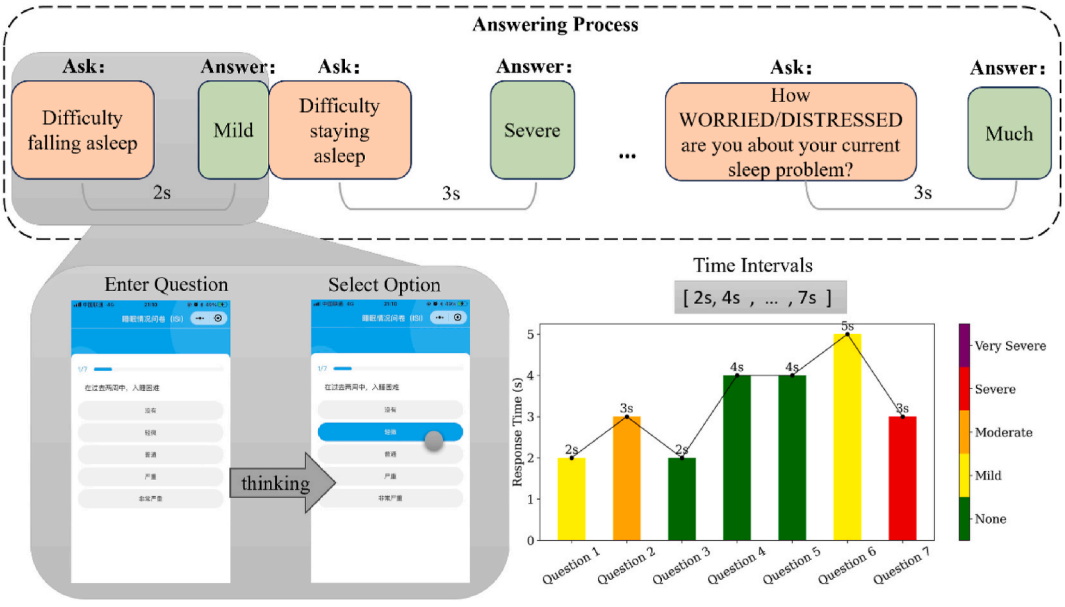


Fig. 1. RT generating in Participant Answer Processing.

2.3. Data processing

Data Cleaning: In this study, the data collection process was dependent on participant responses, which introduce potential issues such as network interruptions or dropout cases, resulting in missing values that could compromise data quality. To address this, we excluded participants with such data irregularities from the study. Additionally, we identified outliers in the RTs using the median absolute deviation (MAD) method. MAD is a robust measure of dispersion that is less sensitive to extreme values compared to standard deviation. We calculated the MAD for each participant's RTs and excluded responses that exceeded a threshold of four times the MAD above the median RT. This statistically justified method ensured that the data used in the analysis was free of significant irregularities, thereby reducing the risk of biased or inconsistent results.

Exclusion of Careless Responses: We also took measures to exclude careless responses from the dataset. The methods used to identify such responses involve detecting RTs that were unusually low. Specifically, if a participant's average RT is below 1.5 s, that participant's data will be excluded from the analysis to avoid confounding results. These thresholds were set based on the understanding that careless responses often manifest as consistently quick (hence the low average RT threshold) [33]. By doing so, we can ensure that only data from participants who demonstrated an adequate level of engagement and conscientiousness during the study are included for analysis.

2.4. Statistics analysis

To compare differences in RT between the two groups – individuals with insomnia symptoms and those without – we employed a non-parametric statistical test, the Mann-Whitney *U* test. This choice of this test was informed by the non-normal distribution of the data and the ordinal nature of the variables under investigation. This test was used to compare the outcome scores between two independent groups. A significant level of 0.01 was set as a priori. Following this, we implemented a one-way Analysis of Variance (ANOVA) with an F-test to examine differences in RT across multiple groups. These groups were determined based on the five different options that participants could select during the task. To further explore, we applied quadratic regression analysis to explore the relationship between option selection (treated as an ordinal variable) and RT. We chose a quadratic model to allow for the possibility that the relationship between option selection and RT might be nonlinear. All statistical analyses were performed using Python 3.9.13 and the statsmodels 0.13.2 [34].

It is important to note that age, which was uniform across our sample, was not considered a potential source of bias in our analyses. Furthermore, we conducted statistical tests to assess the impact of gender on ISI scores and found no significant differences, indicating that gender did not influence the insomnia severity in our study population. This allowed us to proceed without having to adjust for gender as a confounding factor, simplifying our analysis and focusing on the direct relationship between RT and insomnia severity without the need for stratification or additional adjustments for these demographic variables.

2.5. Feature calculation

Our analysis aimed to derive deeper insights into the RT sequence data we collected. For this purpose, we calculated a range of statistical features that would help to describe the distribution of RT. Specifically, we computed the mean, variance, maximum, minimum, median, skewness, kurtosis, range, and coefficient of variation. These features are commonly used in exploratory data analysis and can provide a detailed overview of the data. We also compute features called 'freq_x', which denotes the frequency of a given value 'x'. For instance, 'freq_1′ signifies the proportion of RT falling within the 1–2 s range in relation to the total RT sequence. Similarly, 'big_than_x' features were also calculated, which denote the frequency of values in the sequence that are greater than 'x'. Lastly, the quadratic transformation of each RT was also calculated as a feature.

Moreover, we also employed dimensionality reduction methods to reduce the high-dimensional RT data down to its principal components. We used two techniques, Principal Component Analysis (PCA) [35] and t-distributed Stochastic Neighbor Embedding (t-SNE) [36] to identify hidden patterns and to visualize the data in a lower-dimensional space. Further, the information extracted from the RT data was used as input features for subsequent machine learning models.

2.6. Prediction model

In this study, two different types of training data were utilized: features extracted from RT sequences and raw RT sequence data. Results were reported for both approaches. Utilizing extracted features for model building enhanced interpretability, which is particularly important in the medical field. The participants' scores were evaluated from 7 questions and summed up to constitute the total score. Participants who obtained a score of 8 or more were determined to exhibit symptoms of insomnia and were labeled as '1′, while all other participants were assigned the labeled '0'.

To address label imbalance, the entire dataset was downsampled 10 times, resulting in a total sample size of 282, with 141 samples for each label. The models were validated using 10-fold cross-validation. Five algorithms were compared in this study, including logistic regression, decision trees, support vector machines, K-nearest neighbors and neural networks (MLP). For hyperparameters tuning, we used Optuna in first training fold, an automatic optimization tool based on the Bayesian optimization algorithm [37]. The model construction experiment utilized scikit-learn 1.4.0 [38], imblearn 0.23.1 [39], and mlxtend 0.8.1 [40]. The hyperparameters for each model, as determined by tuning, are shown in Supplement Table 1. A detailed overview of the modeling steps is presented in Fig. 2.

2.7. Feature selection

To select the most predictive feature subset from the original feature set, we first reduced collinearity by excluding features with a correlation coefficient above 0.8, resulting in 28 key features (Supplement Table 3). We then employed the sequential feature selection algorithm for feature selection [41]. This algorithm iteratively selects and refines features based on the current feature subset, gradually reducing the size of the feature set and ultimately obtaining the most predictive feature subset. We utilized logistic regression as the estimator with a limitation of 10 features. The procedure was conducted using backward feature selection without implementing floating feature selection, and the model performance was assessed using the r2 metric, validated via 3-fold cross-validation.

3. Results

We first conducted a comprehensive analysis of the dataset, including descriptive statistics for RTs, scores, and other relevant variables. We then analyzed the correlation among these variables to gain a deeper understanding of the distribution patterns and association strengths of the data. Additionally, we explored the relationship between RT and score at the question level, as well as the potential correlation between the two variables. Following this, we constructed multiple classification models, evaluated them using 10-fold cross-validation, and assessed their performance based on Area Under the ROC Curve (AUROC).

3.1. Demographic statistics

A total of 2729 participants contributed data to this study. After excluding subjects with missing and outlier values, 2320 participants remained. Of these, 1851 met the inclusion criteria for the study, 141 were labeled as having insomnia, while 1710 were labeled as not having insomnia. Details regarding the study participants can be found in Table 1 and Table 2. Among the analyzed participants, 869 were men (74 displaying symptoms of insomnia), and 982 were women (67 displaying symptoms of insomnia). All participants in the experiment were first-year university students aged between 18 and 21 years old. In total, 12,957 values were generated and analyzed. The t-tests or F-tests were employed to examine whether the distribution of ISI scores was consistent across different demographic groups within our sample population. Our analysis did not reveal any significant deviations in the distribution of insomnia according to gender. Further details are available in Table 2.

3.2. The relationship between score and response time

Fig. 3a and c illustrate the distribution of RT and item scores at the individual question level, respectively. Fig. 3b and d presents the distribution of total RT and total scale scores at the level of completing the entire scale, respectively. We utilized the Mann-Whitney *U* test to assess the differences in total RT between individuals with and without insomnia symptoms. The test revealed a significant difference in the distribution between these two independent groups ($U = 65{,}879.5$, $p < 0.001$). Specifically, the median score for the

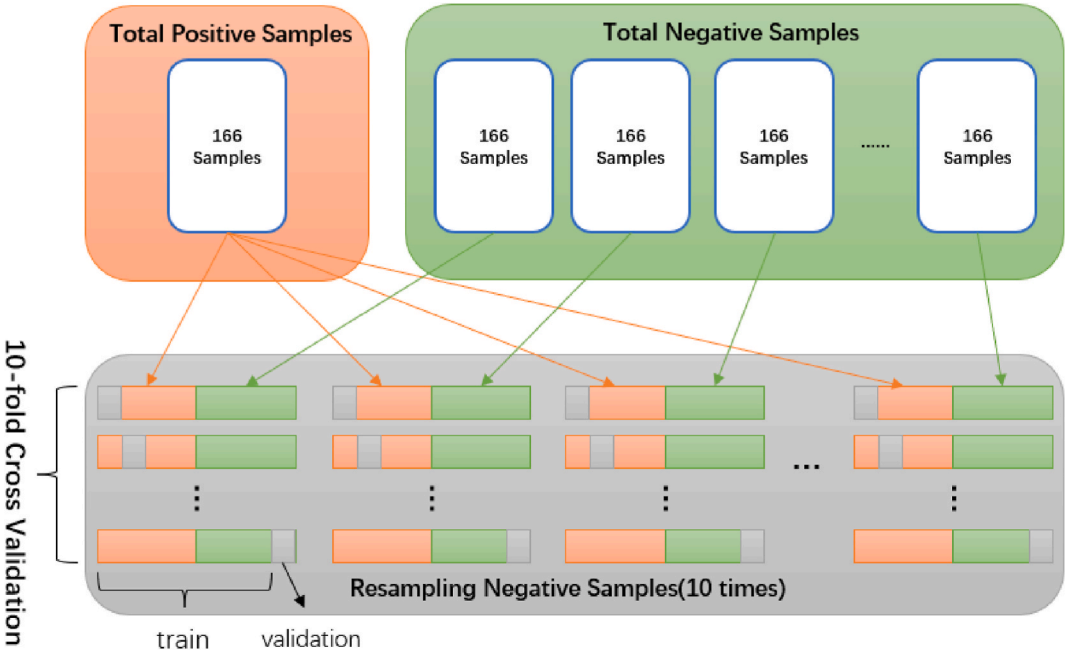


Fig. 2. Data downsampling and 10-fold cross-validation.

Table 1
Number of participants with different labels and compliance with exclusion criteria.

	Total	Included	Excluded
non-insomnia	2171	1710	461
insomnia	149	141	8

Table 2
Demographics of the participants.

	Total(n = 2101)	Insomnia(n = 166)	F or t value	p value
Gender			1.02	0.308
male	869 (46.948 %)	74 (52.482 %)		
female	982 (53.052 %)	67 (47.518 %)		
History psychological illness			−5.702	<0.001
no	1830 (98.865 %)	133 (94.326 %)		
yes	21 (1.135 %)	8 (5.674 %)		
History physical illness			−4.152	<0.001
no	1797 (97.083 %)	131 (92.908 %)		
yes	54 (2.917 %)	10 (7.092 %)		
Smoke			13.05	<0.001
never	1834 (99.082 %)	135 (95.745 %)		
occasional	11 (0.594 %)	2 (1.418 %)		
frequent	4 (0.216 %)	3 (2.128 %)		
former	2 (0.108 %)	1 (0.709 %)		
Drink			12.913	<0.001
never	1416 (76.499 %)	92 (65.248 %)		
occasional	414 (22.366 %)	47 (33.333 %)		
frequent	3 (0.162 %)	1 (0.709 %)		
former	18 (0.972 %)	1 (0.709 %)		

insomnia group (Median = 27, IQR = 9) was statistically significantly higher compared to the non-insomnia group (Median = 21, IQR = 9). Fig. 3e illustrates the difference in the distribution of total RT between these two groups.

We investigated the relationship between participants' scores on a single question, which served as the independent variable, and their RT to that question, which served as the dependent variable. Each question had five options, corresponding to scores from 0 to 4, where higher scores indicate more severe conditions. First, we conducted an analysis of variance to compare RTs among the different options for each of the 7 questions. All questions showed a statistically significant effect, $p < 0.001$. Questions 1, 2, 3 and 4 showed significant effect sizes with F-values ranging from 75.71 to 99.01. Questions 5, 6 and 7 also showed significant effects with F-values of 21.81, 8.73 and 33.56 respectively. Fig. 3f illustrates the differences between the insomnia and non-insomnia groups at the individual question level. All results are summarized in Supplement Table 2.

We investigated the relationship between the variables single question score (denoted as x) and single question RT (denoted as y) by fitting a quadratic regression model to the data. Furthermore, it is important to note that the single question score can also be viewed as an indicator of the severity of specific insomnia aspects. The model was specified as follows: $y = \beta 0 + \beta 1 \times x + \beta 2 \times x^2 + \varepsilon$. Our sample consisted of 1851 observations. Focusing on the first item as an example, the overall model was significant ($F = 171.1$, $p < 0.001$) and accounted for approximately 15.6 % of the variance in the dependent variable y ($R^2 = 0.156$, adjusted $R^2 = 0.155$). The linear term (x) was statistically significant with a coefficient of 1.0251 ($t = 10.805$, $p < 0.001$, 95 % CI [0.839, 1.211]), indicating a positive relationship between x and y. The quadratic term (x^2) was also statistically significant with a coefficient of −0.1143 ($t = -2.825$, $p = 0.005$, 95 % CI [−0.194, −0.035]) indicating a U-shaped relationship between score and RT. Fig. 4a–g illustrate the relationships between question scores and RTs for each of the 7 questions, respectively.

3.3. Statistical analysis of features

In the initial step, we set a threshold for the correlation coefficient at 0.8, excluding features with a correlation above this threshold to mitigate the impact of collinearity, resulting in 28 key features listed in Supplement Table 3. To compare the groups with and without insomnia, we applied t-tests to each of the 28 features. The results revealed significant differences between the two groups on several features. For example, individuals with insomnia exhibited higher levels of mean of RT values and lower levels of coefficient of variation (CV) of RT values compared to those without insomnia. Finally, we performed correlation analyses between each of the 28 features and the severity of insomnia (the scores on the ISI scale). The results indicate that several features were significantly correlated with the severity of insomnia which stand by ISI scores. All the results are shown in Supplement Table 3.

3.4. Classification results

Fig. 5 illustrates the performance evaluations of various models used in our study, including logistic regression, decision tree, SVM,

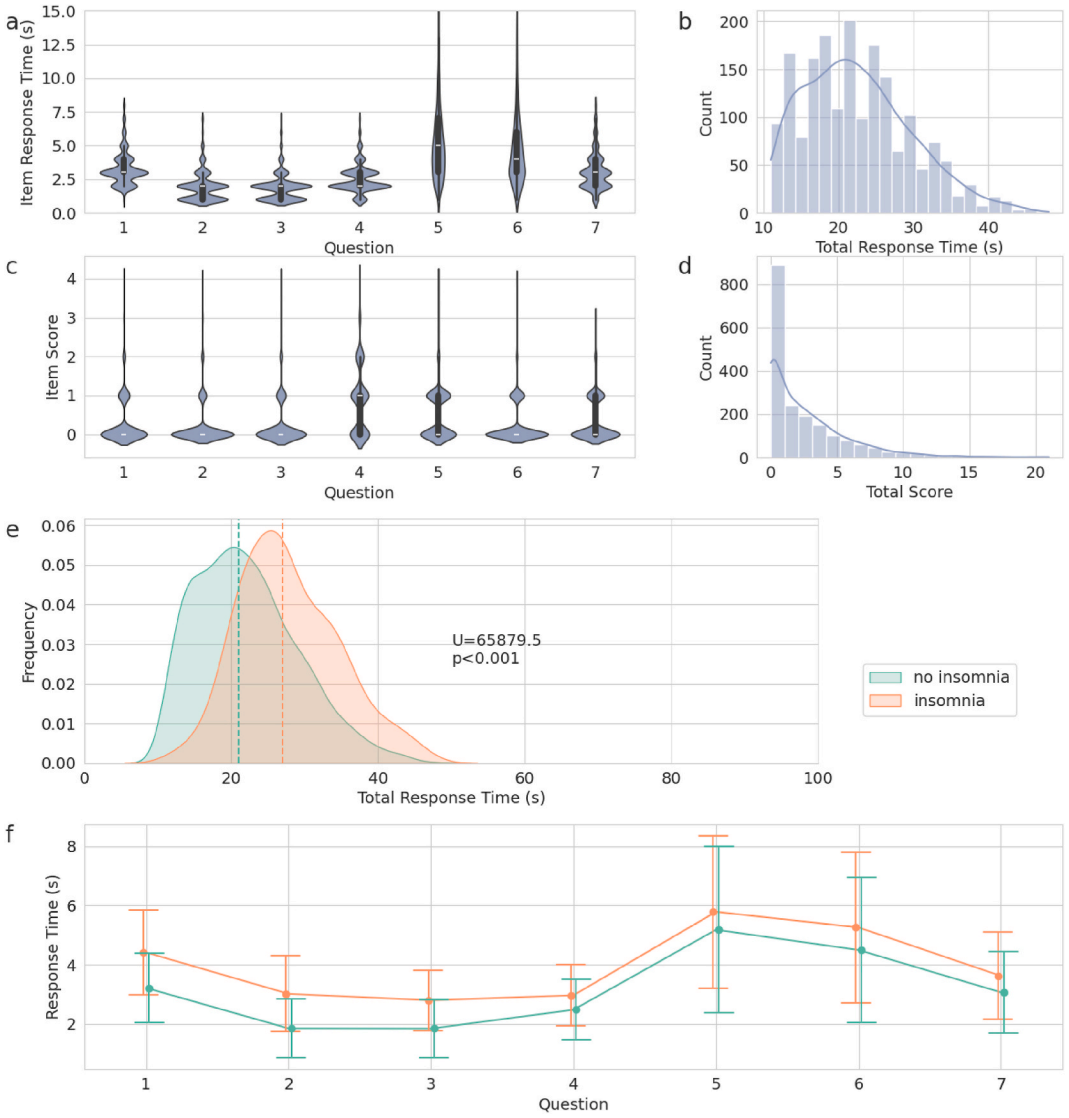


Fig. 3. Comparative Analysis of RTs and Scores on ISI for Groups with and Without Symptoms of Insomnia. a) Violin plot representing the distribution of RTs for each of the seven questions on the ISI scale. Each "violin" represents a different question, with the width of the plot at a given point indicating the density of RTs at that value. **b)** Histogram showing the overall distribution of RTs for finish the total ISI scale. The x-axis represents the total time taken to answer all seven questions, while the y-axis represents the frequency of each RT. **c)** Violin plot representing the distribution of scores for each of the seven questions on the ISI scale. Similar to a), each "violin" represents a different question, but here, the width of the plot at a given point indicates the density of scores at that value. **d)** Histogram showing the overall distribution of total scores on the ISI scale. The x-axis represents the score of ISI, and the y-axis indicates the frequency of each total score. **e)** Distribution plot comparing the ISI score (y-axis) versus RT distribution (x-axis) for two groups: those with symptoms of insomnia (marked in orange) and those without symptoms (marked in a green). The median values for each group are highlighted. **f)** Showing the differences in RTs for individual questions between the two groups (with and without symptoms of insomnia). Each pair of bars represents a different question from the ISI scale, with the point of the bar indicating the average RT. The error bars represent the standard deviation of the RTs for each question in each group.

KNN, and MLP. Participants scoring 8 or above were identified as exhibiting symptoms of insomnia and were labeled '1', while others were labeled '0'. Prior to model training, the dataset underwent 10 iterations of random undersampling to ensure a balance between positive and negative samples (141 individuals with insomnia and 141 individuals without insomnia). Following this, a 10-fold cross-validation was conducted on the undersampled data. The results reported in Fig. 5 represent the mean outcomes of these ten sampling iterations.

The evaluation was done using Receiver Operating Characteristic (ROC) curves for two categories of models: feature-based and raw data-based. In the feature-based category, the logistic regression model demonstrated the most proficient performance. The logistic regression model achieved an AUROC of 0.814, suggesting a high true positive rate relative to the false positive rate, thus underscoring the model's excellent ability in distinguishing between the classes. The MLP model also presented a commendable performance with

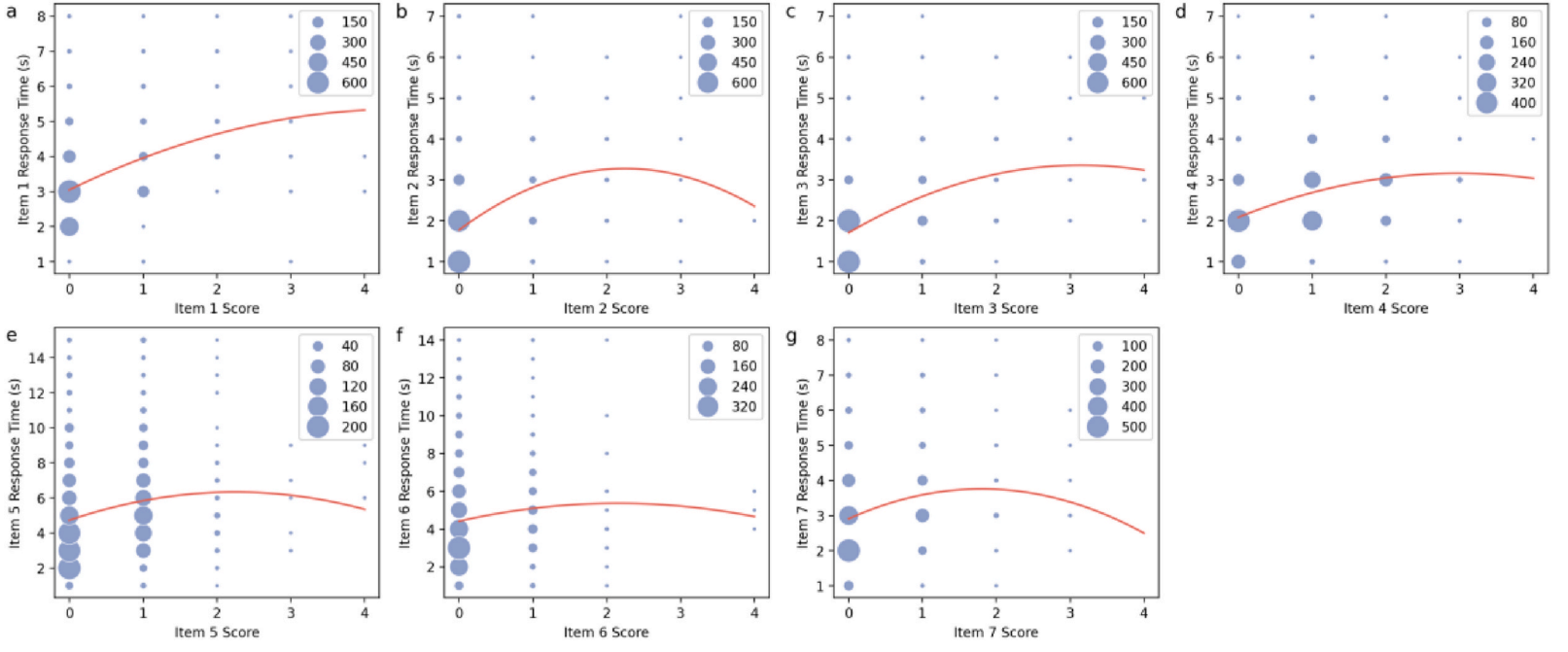


Fig. 4. Results of Regression Analysis for Item Score and Item RT. a) For question 1, the model was significant, explaining 15.6 % of the variance in y. The coefficients were 1.0251 for x and −0.1143 for x^2. **b)** For question 2, the model explained 14.4 % of the variance, with coefficients of 1.3394 for x and −0.2984 for x^2. **c)** Question 3's model explained 17.6 % of the variance, with coefficients of 1.0496 and −0.167. **d)** The model for question 4 explained 14 % of the variance, with coefficients of 0.7239 and −0.1218. **e)** For question 5, 4.3 % of the variance was explained, with coefficients of 1.433 and −0.3194. **f)** Question 6's model explained 1.3 % of the variance, with coefficients of 0.8849 and −0.2060. **g)** For question 7, the model explained 5.1 % of the variance in y, with coefficients of 0.9439 for x and −0.2617 for x^2.

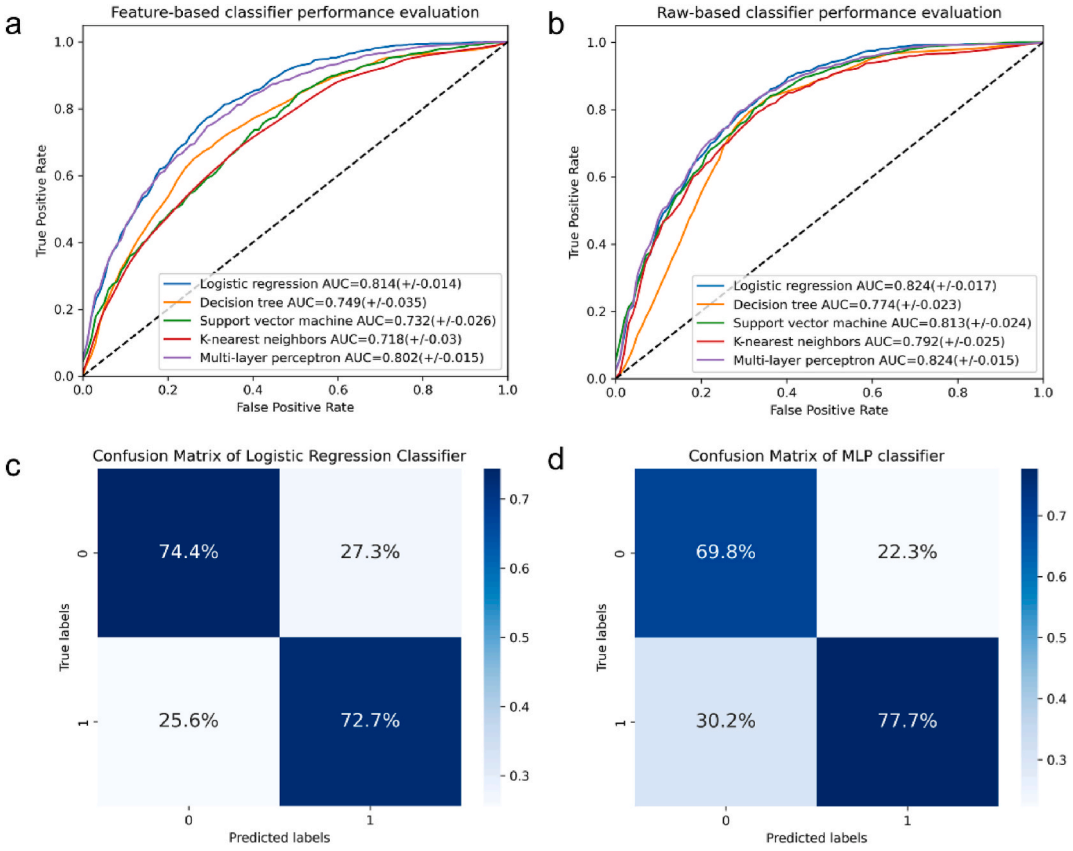


Fig. 5. Performance Comparison of Various Machine Learning Models on Processed and Raw RT Data. a) ROC curves of Logistic Regression, Decision Tree, Support Vector Machine, K-nearest Neighbors, and Multi-layer Perceptron models applied to features derived from RT data. **b)** ROC curves of the same models applied directly to raw RT data. **c)** Confusion matrix of the best performing model (Logistic Regression) trained on the feature data. **d)** Confusion matrix of the best performing model (Multi-layer Perceptron) trained on raw data. All plots were obtained using 10-fold cross-validation, with the process repeated 10 times for resampling purposes.

an AUROC of 0.802. For the raw data-based models, the performances were similar across the various models. Notably, the logistic regression and MLP model came out on top, with an AUROC of 0.824, indicating an excellent balance between sensitivity and specificity. Fig. 5a and c illustrate the comparison of ROC curves for different feature-based models and the confusion matrix of the best-performing model, respectively. Similarly, Fig. 5b and d presents the comparison of ROC curves for different raw data-based models and the confusion matrix of the best-performing model, respectively.

In feature prediction (Supplement Table 4), logistic regression emerged as the most effective model, offering superior performance in accuracy (0.735), precision (0.704), and F1 score (0.717). It also demonstrated competitive recall (0.745). These results suggest that logistic regression excelled at accurate classification and balancing precision and recall.

However, with raw data prediction (Supplement Table 5), the MLP model achieved the highest accuracy (0.743) and precision (0.74). Logistic regression maintained high accuracy (0.741) and precision (0.733), while the SVM model excelled in recall (0.806). Interestingly, the SVM model showed substantial improvement in recall when employing raw data (0.806) compared to feature prediction data (0.551). This suggests that the SVM model was particularly effective at identifying true positives when raw data was utilized.

3.5. Feature importance

In this study, we utilized Shapley Additive Explanations (SHAP) values to evaluate the contribution of each feature in the model [42]. SHAP values help in understanding how each feature impacts model predictions. Fig. 6 illustrates the contribution rates of 28 selected features by correlation using a logistic regression model. The top five features with the highest contribution rates are pca_3, mean, freq_2, RT6_trans and RT4_trans. These findings suggest that several features in the RT sequence are effective representations for insomnia symptoms.

The third principal component, pca_3, accounted for approximately 12.2 % of the total variance in the data, representing a significant portion of the information in the dataset. Specifically, the pca_3 feature is primarily composed of the first, second, and third questions' RT in ISI scale with corresponding loadings of 0.7105, 0.3003, and 0.3046, respectively. These loadings suggest that these

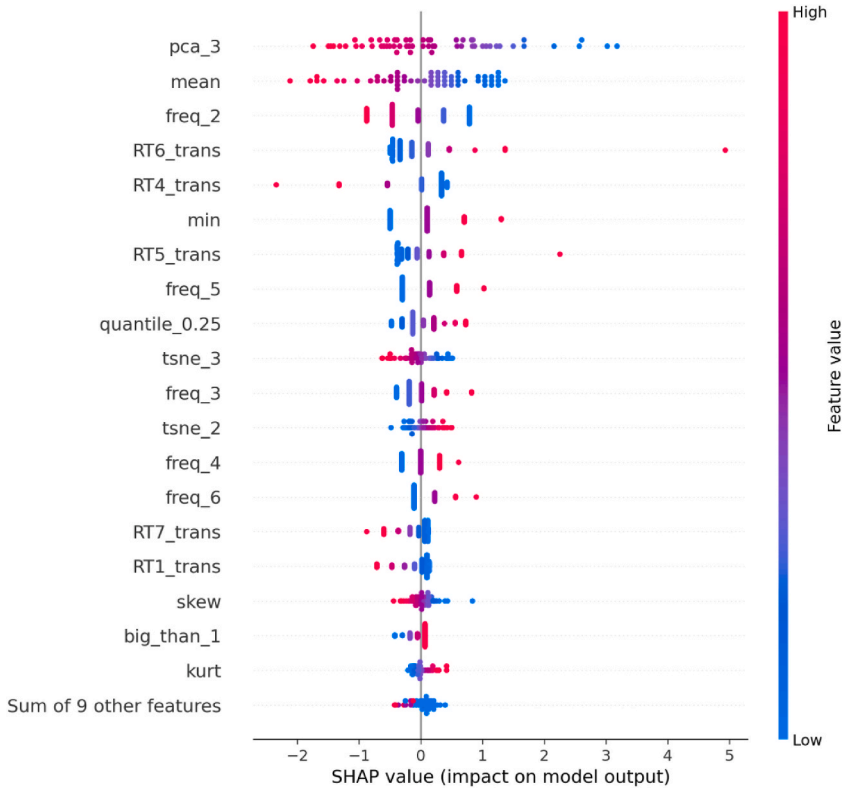


Fig. 6. SHAP values of top features of best model.

features are the most relevant for pca_3, given their relatively high positive correlation. Interestingly, the fifth and sixth question RT appear to contribute negatively to pca_3, with loadings of −0.1752 and −0.4141, respectively. The mean of RT emerged as one of the most important features, indicating that the overall average speed of a participant's response plays a crucial role in the predictive model. The proportion of RT sequence within the 2–3 s range, also emerged as highly significant. This result underscores the relevance of the timely reaction category of 2–3 s, suggesting that responses within this time frame carry substantial predictive weight. Additionally, regarding the sixth question on "How WORRIED/DISTRESSED are you about your current sleep problem? " and fourth question on "How SATISFIED/DISSATISFIED are you with your CURRENT sleep pattern?", this implies that those specific questions on the ISI scale hold significant predictive value. The particular importance of this feature suggests a profound link between insomnia severity and the variable our model predicts.

4. Discussion

The key finding of this study lies in the utility of user behavior during the process of responding to a scale, specifically, the RT, as an effective predictor of ISI scale results. The machine learning classifier model constructed achieved a remarkable accuracy of up to 74 % and a recall of 80 %. Our preliminary statistical test and regression analysis have identified a quadratic relationship pattern between RT and option choices, consistent with previous literature [8]. By exploring the interpretability of machine learning models, we found that features related to the length of RT, such as mean, tend to have high weights. This observation corroborates prior studies suggesting insomnia often results in a decline in daytime cognitive abilities [12].

While our study demonstrates the utility of RT data in predicting insomnia symptoms as defined by ISI scores, future research should investigate whether RT data can outperform ISI scores when insomnia diagnosis is assigned through psychiatric interviews. Psychiatric interviews are considered the gold standard for diagnosing insomnia, as they provide a more comprehensive assessment of an individual's sleep patterns and related symptoms. By comparing the predictive power of RT data against expert-assigned diagnoses, we can gain a clearer understanding of its potential as an objective marker for insomnia.

In machine learning experiments, we observed that models trained on raw RT interval series slightly outperformed feature-based models. Specifically, both the Logistic Regression and MLP models achieved an AUROC of 0.824 when using raw data. In contrast, the best-performing feature-based model, Logistic Regression, attained an AUROC of 0.814. This observation suggests that, despite the superior fitting capacity of neural networks and their ability to effectively represent non-linear relationships, simple models can sometimes yield unexpected results, and additionally provides enhanced interpretability. The computation of statistical features, distribution features, and manually designed features of the RT interval series revealed a linear relationship with the labels. This made

the machine learning models simpler and more effective, often leading to greater reliability in practical applications.

From a medical perspective, our findings suggest that RT data may have potential as a complementary tool for assessing symptoms of insomnia, leveraging existing data collected during the administration of self-report scales. However, further research is needed to validate the clinical utility of this approach and determine its role in relation to established diagnostic methods used by healthcare professionals. In the context of psychological assessment scales, our results advocate for the use of RT as a guiding tool in scale design. Methodologically, the collection of participant behavior data during scale responses provides a complementary viewpoint, enhancing the precision of scale evaluations and reducing subjectivity, offering potential improvements to current assessment systems.

There are some notable limitations to our approach that should be considered. Firstly, our study relied on ISI scores to identify participants with insomnia symptoms, rather than clinical interviews, which are considered the gold standard for insomnia diagnosis. This reliance on self-reported data poses a constraint, as such data can be prone to bias and may not faithfully represent an individual's sleep patterns. Future research could benefit from using structured clinical interviews to establish insomnia diagnoses, which would provide a more robust ground truth for evaluating the predictive utility of RT data. Furthermore, since our measurements are based on the environment in which respondents complete these scales is uncontrolled, environmental factors could skew the results. Secondly, the limited size of our sample within different age groups means our findings may not extend to larger populations. Another potential limitation of our study is that our model relies solely on RT as a predictor of insomnia severity. Future studies could incorporate additional behavioral data, such as screen scrolling speed, response revision frequency, and touch pressure, to further enhance the predictive power of the model and provide a more comprehensive understanding of the relationship between digital behaviors and insomnia severity. Lastly, considering the potential network latency issues that may arise from online data collection, even within a university campus setting, a future direction for research could involve redeveloping the data collection program to function entirely offline. This would allow for the RT data to be saved directly on the device and subsequently uploaded, ensuring that any network-related delays do not influence the measurements. Despite these limitations, our study lays a foundational groundwork for future investigations, facilitating the development of our understanding of the relationship between RT and insomnia severity.

5. Conclusions

In conclusion, our study underscores the potential utility of RT as an effective indicator of insomnia severity. The findings presented herein suggest that machine learning techniques have the potential to contribute to a more efficient and straightforward interpretation of RT data, thereby complementing and potentially enhancing the implications of ISI scale results. The integration of behavioral data, such as response time, with machine learning algorithms shows promise in enhancing the assessment of psychological and behavioral aspects of insomnia. We anticipate that as technology continues to advance, more personalized and data-driven approaches to insomnia assessment and treatment will emerge.

Ethics statement

This study was reviewed and approved by the Ethics Committee of Nanjing Medical University with the approval number: No. (2022)793, dated 2022-07-12. Informed consent was obtained from all participants involved in this study. The consent process was conducted online. Specifically, all participants were required to read and agree to an informed consent form at the initial page of the application before proceeding to answer any scales.

Funding statement

This study is funded by National Natural Science Foundation of China (62176129 to Xizhe Zhang), National Science Fund for Distinguished Young Scholars (81725005 to Fei Wang), the National Natural Science Foundation Regional Innovation and Development Joint Fund (U20A6005 to Fei Wang), Jiangsu Provincial Key Research and Development Program, China (BE2021617 to Fei Wang), Henan Provincial Research and Practice Project for Higher Education Teaching Reform (2021SJGLX189Y to Rongxun Liu).

Data availability statement

The data supporting the results in this study are available within the paper and its Supplementary Information. The dataset used in this study is available upon reasonable request and subject to privacy considerations.

CRediT authorship contribution statement

Zhao Su: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Rongxun Liu:** Writing – review & editing. **Keyin Zhou:** Writing – review & editing. **Xinru Wei:** Writing – review & editing, Methodology. **Ning Wang:** Writing – review & editing. **Zexin Lin:** Writing – review & editing. **Yuanchen Xie:** Writing – review & editing. **Jie Wang:** Writing – review & editing. **Fei Wang:** Writing – review & editing, Resources, Conceptualization. **Shenzhong Zhang:** Writing – review & editing. **Xizhe Zhang:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e33485.

References

[1] M.L. Perlis, D. Posner, D. Riemann, C.H. Bastien, J. Teel, M. Thase, Insomnia, Lancet 400 (2022) 1047–1060, https://doi.org/10.1016/S0140-6736(22)00879-0.
[2] D. Riemann, F. Benz, R.J. Dressle, C.A. Espie, A.F. Johann, T.F. Blanken, J. Leerssen, R. Wassing, A.L. Henry, S.D. Kyle, K. Spiegelhalder, E.J.W. Van Someren, Insomnia disorder: state of the science and challenges for the future, J. Sleep Res. 31 (2022) e13604, https://doi.org/10.1111/jsr.13604.
[3] D. Riemann, C.A. Espie, E. Altena, E.S. Arnardottir, C. Baglioni, C.L.A. Bassetti, C. Bastien, N. Berzina, B. Bjorvatn, D. Dikeos, L. Dolenc Groselj, J.G. Ellis, D. Garcia-Borreguero, P.A. Geoffroy, M. Gjerstad, M. Gonçalves, E. Hertenstein, K. Hoedlmoser, T. Hion, B. Holzinger, K. Janku, M. Jansson-Fröjmark, H. Järnefelt, S. Jernelöv, P.J. Jennum, S. Khachatryan, L. Krone, S.D. Kyle, J. Lancee, D. Leger, A. Lupusor, D.R. Marques, C. Nissen, L. Palagini, T. Paunio, L. Perogamvros, D. Pevernagie, M. Schabus, T. Shochat, A. Szentkiralyi, E. Van Someren, A. van Straten, A. Wichniak, J. Verbraecken, K. Spiegelhalder, The European Insomnia Guideline: an update on the diagnosis and treatment of insomnia 2023, J. Sleep Res. 32 (2023) e14035, https://doi.org/10.1111/jsr.14035.
[4] A.G. Horwitz, Z. Zhao, S. Sen, Peak-end bias in retrospective recall of depressive symptoms on the PHQ-9, Psychol. Assess. 35 (2023) 378–381, https://doi.org/10.1037/pas0001219.
[5] D. Watkins, S. Cheung, Culture, gender, and response bias: an analysis of responses to the self-description questionnaire, J. Cross Cult. Psychol. 26 (1995) 490–504, https://doi.org/10.1177/0022022195265003.
[6] P. Lugtig, V. Toepoel, The use of PCs, smartphones, and tablets in a probability-based panel survey: effects on survey measurement error, Soc. Sci. Comput. Rev. 34 (2016) 78–94, https://doi.org/10.1177/0894439315574248.
[7] R. Tourangeau, H. Sun, T. Yan, A. Maitland, G. Rivero, D. Williams, Web surveys by smartphones and tablets: effects on data quality, Soc. Sci. Comput. Rev. 36 (2018) 542–556, https://doi.org/10.1177/0894439317719438.
[8] K. Chung, J.Y. Park, D. Joung, K. Jhung, Response time as an implicit self-schema indicator for depression among undergraduate students: preliminary findings from a mobile app–based depression assessment, JMIR Mhealth Uhealth 7 (2019) e14657, https://doi.org/10.2196/14657.
[9] C.R. Gale, A. Harris, I.J. Deary, Reaction time and onset of psychological distress: the UK Health and Lifestyle Survey, J. Epidemiol. Community Health 70 (2016) 813–817, https://doi.org/10.1136/jech-2015-206479.
[10] R.J. Kosinski, A Literature Review on Reaction Time, 10, Clemson University, 2008, pp. 337–344.
[11] R. Pachella, The Use of Reaction Time Measures in Information Processing Research, Human Information Processing, Erlbaum, Hillsdale, NJ, 1974.
[12] É. Fortier-Brochu, S. Beaulieu-Bonneau, H. Ivers, C.M. Morin, Insomnia and daytime cognitive performance: a meta-analysis, Sleep Med. Rev. 16 (2012) 83–94, https://doi.org/10.1016/j.smrv.2011.03.008.
[13] M. Saint Martin, E. Sforza, J.C. Barthélémy, C. Thomas-Anterion, F. Roche, Does subjective sleep affect cognitive function in healthy elderly subjects? The Proof cohort, Sleep Med. 13 (2012) 1146–1152, https://doi.org/10.1016/j.sleep.2012.06.021.
[14] S.D. Kyle, C.E. Sexton, B. Feige, A.I. Luik, J. Lane, R. Saxena, S.G. Anderson, D.A. Bechtold, W. Dixon, M.A. Little, D. Ray, D. Riemann, C.A. Espie, M.K. Rutter, K. Spiegelhalder, Sleep and cognitive performance: cross-sectional associations in the UK Biobank, Sleep Med. 38 (2017) 85–91, https://doi.org/10.1016/j.sleep.2017.07.001.
[15] T. Blackwell, K. Yaffe, S. Ancoli-Israel, S. Redline, K.E. Ensrud, M.L. Stefanick, A. Laffan, K.L. Stone, For the osteoporotic fractures in men (MrOS) study group, association of sleep characteristics and cognition in older community-dwelling men: the MrOS sleep study, Sleep 34 (2011) 1347–1356, https://doi.org/10.5665/SLEEP.1276.
[16] J.A. Shekleton, E.E. Flynn-Evans, B. Miller, L.J. Epstein, D. Kirsch, L.A. Brogna, L.M. Burke, E. Bremer, J.M. Murray, P. Gehrman, S.W. Lockley, S.M. W. Rajaratnam, Neurobehavioral performance impairment in insomnia: relationships with self-reported sleep and daytime functioning, Sleep 37 (2014) 107–116, https://doi.org/10.5665/sleep.3318.
[17] O. Grau-Rivera, G. Operto, C. Falcón, G. Sánchez-Benavides, R. Cacciaglia, A. Brugulat-Serrat, N. Gramunt, G. Salvadó, M. Suárez-Calvet, C. Minguillon, Á. Iranzo, J.D. Gispert, J.L. Molinuevo, J. Camí, M. Crous-Bou, C. Deulofeu, R. Dominguez, X. Gotsens, L. Hernández, G. Huesa, J.M. González-de-Echávarri, J. Huguet, M. León, P. Marne, E.M. Arenaza-Urquijo, T. Menchón, M. Milà, M. Pascual, A. Polo, S. Pradas, A. Sala-Vila, S. Segundo, M. Shekari, A. Soteras, L. Tenas, M. Vilanova, N. Vilor-Tejedor, For the ALFA Study, Association between insomnia and cognitive performance, gray matter volume, and white matter microstructure in cognitively unimpaired adults, Alzheimer's Res. Ther. 12 (2020) 4, https://doi.org/10.1186/s13195-019-0547-3.
[18] P.C. Kyllonen, R.E. Christal, Reasoning ability is (little more than) working-memory capacity? Intelligence 14 (1990) 389–433, https://doi.org/10.1016/S0160-2896(05)80012-1.
[19] R. Johnson, J. Barnhardt, J. Zhu, Differential effects of practice on the executive processes used for truthful and deceptive responses: an event-related brain potential study, Cognit. Brain Res. 24 (2005) 386–404, https://doi.org/10.1016/j.cogbrainres.2005.02.011.
[20] S. Thomée, Mobile phone use and mental health. A review of the research that takes a psychological perspective on exposure, Int. J. Environ. Res. Publ. Health 15 (2018) 2692, https://doi.org/10.3390/ijerph15122692.
[21] S. Thomée, A. Härenstam, M. Hagberg, Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults - a prospective cohort study, BMC Publ. Health 11 (2011) 66, https://doi.org/10.1186/1471-2458-11-66.
[22] T. Börger, Are fast responses more random? Testing the effect of response time on scale in an online choice experiment, Environ. Resour. Econ. 65 (2016) 389–413, https://doi.org/10.1007/s10640-015-9905-1.
[23] K. Bunji, K. Okada, Item response and response time model for personality assessment via linear ballistic accumulation, Jpn J Stat Data Sci 2 (2019) 263–297, https://doi.org/10.1007/s42081-019-00040-4.
[24] M. Gogami, Y. Matsuda, Y. Arakawa, K. Yasumoto, Detection of careless responses in online surveys using answering behavior on smartphone, IEEE Access 9 (2021) 53205–53218, https://doi.org/10.1109/ACCESS.2021.3069049.
[25] E. Ulitzsch, S. Pohl, L. Khorramdel, U. Kroehne, M. von Davier, A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data, Psychometrika 87 (2022) 593–619, https://doi.org/10.1007/s11336-021-09817-7.
[26] P.J. Ferrando, U. Lorenzo-Seva, A measurement model for likert responses that incorporates response time, Multivariate Behav. Res. 42 (2007) 675–706, https://doi.org/10.1080/00273170701710247.
[27] N. Iwata, K. Kikuchi, Y. Fujihara, The usability of CAT system for assessing the depressive level of Japanese-A study on psychometric properties and response behavior, Int. J. Behav. Med. 23 (2016) 427–437, https://doi.org/10.1007/s12529-015-9503-1.
[28] M.K. Nock, M.R. Banaji, Prediction of suicide ideation and attempts among adolescents using a brief performance-based test, J. Consult. Clin. Psychol. 75 (2007) 707–715, https://doi.org/10.1037/0022-006X.75.5.707.

[29] C. Wang, D.J. Weiss, S. Su, Modeling response time and responses in multidimensional health measurement, Front. Psychol. 10 (2019), https://doi.org/10.3389/fpsyg.2019.00051.
[30] A. Baba, K. Bunji, Prediction of mental health problem using annual student health survey: machine learning approach, JMIR Mental Health 10 (2023) e42420, https://doi.org/10.2196/42420.
[31] C.H. Bastien, A. Vallières, C.M. Morin, Validation of the Insomnia Severity Index as an outcome measure for insomnia research, Sleep Med. 2 (2001) 297–307, https://doi.org/10.1016/S1389-9457(00)00065-4.
[32] D.S.F. Yu, Insomnia Severity Index: psychometric properties with Chinese community-dwelling older people, J. Adv. Nurs. 66 (2010) 2350–2359, https://doi.org/10.1111/j.1365-2648.2010.05394.x.
[33] N.A. Bowling, J.L. Huang, C.B. Bragg, S. Khazon, M. Liu, C.E. Blackmore, Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality, J. Pers. Soc. Psychol. 111 (2016) 218–229, https://doi.org/10.1037/pspp0000085.
[34] S. Seabold, J. Perktold, Statsmodels: Econometric and Statistical Modeling with Python, 2010, pp. 92–96, https://doi.org/10.25080/Majora-92bf1922-011. Austin, Texas.
[35] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos Trans A Math Phys Eng Sci 374 (2016) 20150202, https://doi.org/10.1098/rsta.2015.0202.
[36] E. Schubert, M. Gertz, Intrinsic t-stochastic neighbor embedding for visualization and outlier detection, in: C. Beecks, F. Borutta, P. Kröger, T. Seidl (Eds.), Similarity Search and Applications, Springer International Publishing, Cham, 2017, pp. 188–203, https://doi.org/10.1007/978-3-319-68474-1_13.
[37] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: a next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA, 2019, pp. 2623–2631, https://doi.org/10.1145/3292500.3330701.
[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine Learning in Python, MACHINE LEARNING IN PYTHON (n.d.). .
[39] G. Lemaıtre, F. Nogueira, Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, (n.d.).
[40] S. Raschka, MLxtend: providing machine learning and data science utilities and extensions to Python's scientific computing stack, JOSS 3 (2018) 638, https://doi.org/10.21105/joss.00638.
[41] D.W. Aha, R.L. Bankert, A comparative evaluation of sequential feature selection algorithms, in: Pre-Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics, PMLR, 1995, pp. 1–7, in: https://proceedings.mlr.press/r0/aha95a.html. (Accessed 25 May 2023).
[42] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017, in: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html. (Accessed 29 April 2023).