ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Mental Health Risk Identification and Feature Analysis Techniques for University Students

Li Chang[1], *Ruili Xue*[1] *and Beilei Qiao*[1,*]

[1] Henan Agricultural University, Zhengzhou 450046, Henan, China

SUMMARY: *In response to the practical constraints of "large sample size, limited manual review, and insufficient discrimination of single total scores" in the normalization of psychological screening in universities, this article studies the identification and feature analysis technology of psychological health risks among college students. Based on publicly available psychological assessment data from 24292 students, integrating PHQ-9, GAD-7, PSS, ISI scales, demographic information, and question by question response time, a student level risk representation was constructed. A multi view recognition framework with enhanced response time was designed to incorporate symptom intensity, response time deviation, and similar context of students into the unified discrimination process. The results showed that in the publicly available samples, PHQ-9 mil and above accounted for 30.49%, GAD-7 mil and above accounted for 16.97%, PSS high and above accounted for 11.53%, and ISI threshold and above accounted for 9.06%; The coverage of stress-related burdens is wider, and the marginal zone of depression risk is longer. Further analysis of response behavior shows that there is a clear bimodal structure in the total response time of ISI and PSS, with key turning points located at approximately 12 s and 23 s, indicating that response time can provide incremental information for edge risk identification. This article presents a low-cost risk identification and interpretation analysis solution for university scenarios, which can provide technical support for screening stratification, manual review sorting, and referral decision-making.*

KEYWORDS: *College students' mental health; Risk identification; Duration of response; Multi view learning; psychological screening*

1 Introduction

The main problems which mental health service systems in institutions of higher education face are the screening ability and the promptness of responses. The adjusting stage after entering school, the gathering of study pressure, broken sleep rules, and changes in social help often push many students into a "gray area"-where they show obvious discomfort but have not yet arrived at the standard for medical treatment-within one semester. For universities, this group of people is the most difficult in management: on one hand, traditional one-by-one interviews cannot cover a big enough scope; On the other hand, if one only depends on one single total-score screen, thus it becomes hard to separate short-term mood changes, continuous high-pressure conditions, and high-risk persons who need first-priority referral. Trans-national study on university students' psychological health has discovered that the university years themselves are a high-risk period for common psychological illnesses, with symptoms frequently beginning early, lasting for long times, and directly connected to damaged study and social performance

*qbl921@126.com
https://doi.org/10.65102/is2026773
Published: 30 April, 2026

[1-3].

Consequently, campus mental health screening has gradually shifted toward online questionnaires and platform-based management. Internet-based psychological assessments have established relatively stable methods for administration, interpretation, and data retention, with questionnaire items, response order, and timestamps all recorded synchronously [4, 5]. In the university context, the value of this approach lies not only in improving administrative efficiency but also in introducing behavioral-level signals that traditional paper-and-pencil tests lack.Research on reaction times and online surveys indicates that response duration typically exhibits significant skewness and is jointly influenced by cognitive load, item difficulty, hesitation, and disengaged responding; in mobile surveys, rapid consecutive clicks, abnormal pauses, and irregular pacing between items are all associated with invalid or low-quality responses [6-10]. This implies that psychological screening data should not be interpreted merely as a set of scores; the response process itself is equally worthy of modeling.

With respect to the recognition of mental health problems among college students, the method of machine learning has already obtained a number of results that can be directly compared. One research that judges serious mental uneasy condition through population basic characters, living ways and body moving habits got an AUC value of 0.932 and an F1 value of 0.856 in inside verification, therefore the AUC still keeps 0.918 in outside verification [11]; In a research with few samples about anxiety degree of college students, a Stacking ensemble model has given an accuracy rate of 97.83% and an F1 value of 97.88% [12]. At the same time, recent systematical reviews have gathered the main research methods for student mental health recognition into three sorts of input: structured question papers, campus management information, and passive sensing or multimodal behavior data [13-15]. The first two kinds of methods have relatively lower arrangement expenses and are fit for conventional examination work; that latter method can provide richer signal information but meets more obvious difficulties in the aspects of data collection obstacles, privacy limit requirements, and cross-scenario universality promotion.

Four gaps remain in the existing research. First, many methods focus on a single condition-such as depression or anxiety-making it difficult to capture the co-occurrence of depression, anxiety, stress, and insomnia, which is more common among college students.Second, model inputs remain dominated by total scale scores or questionnaire covariates, while the time taken to answer each question is often treated as noise or a secondary quality control variable and has not yet been systematically incorporated into the primary risk identification model. Third, although the passive sensing approach can incorporate behavioral cues such as location, step count, and mobile interactions, the high cost of data collection makes it difficult to integrate directly into large-scale routine screening workflows.Fourth, existing work has prioritized classification accuracy over issues of greater practical concern to universities, such as threshold configuration, manual review workload, and interpretable referral interfaces. These challenges necessitate a more restrained technical approach for campus settings: while inputs should primarily consist of scales and platform-native logs, the identification logic must simultaneously leverage symptom intensity, cross-scale couplings, and response behavior.

This method possesses a firm data basis. The open dataset which was published by Su et al. contains PHQ-9, GAD-7, PSS, and ISI scores, population basic information, and single-item answer times for 24,292 college students, hence it gives a proper starting point for low-expense, multi-angle danger identification [16]. Following researches further proved that when only response time sequences are used, the AUROC for recognizing insomnia symptoms can achieve 0.824 [17]. To university platforms, this point is of great importance: if response time can promote the capability for recognizing marginal risks without raising the testing burden, hence it is not just an additional area for quality control, thus it is a characteristic that can be directly

put into the model.

According to this point, this paper puts forward a multi-angle identification framework which is strengthened by reaction time, hence concentrates on mental health risk judgment and feature analysis that lies among college students. Our work puts focus on three respects: First, on the foundation of publicly obtainable data from four dimensions and item-level reaction times, we build a unified student-level risk expression and give a label mapping for multi-task identification; Second, we have incorporated symptom seriousness, reaction time deviations, and student-level context resemblance into one single model, thus forming a low-cost recognition work flow which is suitable for campus examination and selection situations; Third, we have established an interpretable interaction surface that directly corresponds to manual examination and recommendation priorities, which is based on risk grades, three-dimensional probability curved faces, module cutting experiments, calibration effects, and error origins.

2 Methods

2.1 Open-Data Organization and Label Construction

The present study makes use of the psychological assessment data on college students that are opened to public by Su et al. The data were got from February 27 to March 17, 2021, through a moving phone psychology check system, and they cover an overall total of 24,292 students. The public open release contains five file kinds: population characteristic. CSV, PHQ-9. CSV, GAD-7. CSV, PSS. CSV, and also ISI. csv, which carry the record of age, gender, item-level marks for the four scales, and the reaction time for every single item. The four measuring forms include altogether 37 test questions, thus therefore giving a total of 898,804 original data records. The original paper already got rid of outlier points on the basis of median and MAD rules; this paper uses its publicly opened cleaned edition and, based on it, therefore goes to build unified marks and arrange multi-angle characteristics. For the simultaneous retention of the item score results of the four scales, the deviations of item answering duration, and the whole individual outlier condition on the student level, this research firstly builds student-level input expression forms, item duration deviation measurement quantities, and deviations from the mean duration, which is displayed in Equation (1).

$$\mathbf{x}_i = [\tilde{\mathbf{p}}_i \| \tilde{\mathbf{g}}_i \| \tilde{\mathbf{s}}_i \| \tilde{\mathbf{u}}_i], \quad r_{ij} = \frac{\log(t_{ij}+1) - med_j}{1.4826\, MAD_j + \varepsilon}, \quad a_i = \frac{1}{m_i} \sum_{j=1}^{m_i} |r_{ij}| \tag{1}$$

In the equation, $\tilde{p}_i$, $\tilde{g}_i$, $\tilde{s}_i$, and $\tilde{u}_i$ represent the item score vectors of student i on the PHQ-9, GAD-7, PSS, and ISI scales, respectively, after intra-scale normalization. t_{ij} represents the response time of student i on the jth item. med_j and MAD_j represent the median and absolute median deviation of the overall sample for that item, respectively.ε is a smoothing term.m_i is the number of valid items for student i. r_{ij} represents the standardized deviation in item response duration after log transformation. a_i denotes the student-level average deviation in response duration. This approach preserves both local item deviations and overall individual response anomalies, making it suitable for subsequent identification of rapid responses, abnormal hesitations, and imbalances in inter-item rhythm. Log transformation of response duration is applied to mitigate the right-skewed distribution and reduce the influence of outliers, consistent with common practices in reaction time modeling.

Label construction is based on five tasks. The four binary tasks correspond to depression risk, anxiety risk, stress risk, and insomnia risk, respectively, where PHQ-9 $\geq$ 10, GAD-7 $\geq$ 10, PSS at "high" or "very high," and ISI $\geq$ 15 are defined as positive labels.The comprehensive

risk task is a three-class classification: 0 positive tasks are mapped to low risk, 1–2 positive tasks to moderate risk, and 3–4 positive tasks to high risk.To prevent threshold information from the target scales from leaking directly into the corresponding sub-tasks, the four sub-tasks employ "leave-one-scale-out" masking during training: the raw item scores and total scores of the target scale are excluded from the sub-task's input, retaining only the response-time sequence of that scale as the behavioral view; the comprehensive risk task retains the item scores of the four scales but does not include any label-based indicator variables generated by threshold rules.

At the student-level data organization, this study retains four types of information for each student: item score vectors, item-by-item duration deviation vectors, scale-level statistics, and demographic background. Scale-level statistics include the total scores, means, variances, inter-item spans, proportions of rapid responses, proportions of long pauses, and duration differences between the first and last segments for the four scales; demographic background includes age and gender.The entire sample is stratified according to the composite risk label and divided into training, validation, and test sets in a 70%/15%/15% ratio. This partitioning scheme is consistently applied across the five tasks-training, ablation, calibration, and error analysis-to prevent result drift caused by differing experimental parameters. To illustrate the organizational relationships among scale scores, response times, and risk labels in the public dataset, see Figure 1.

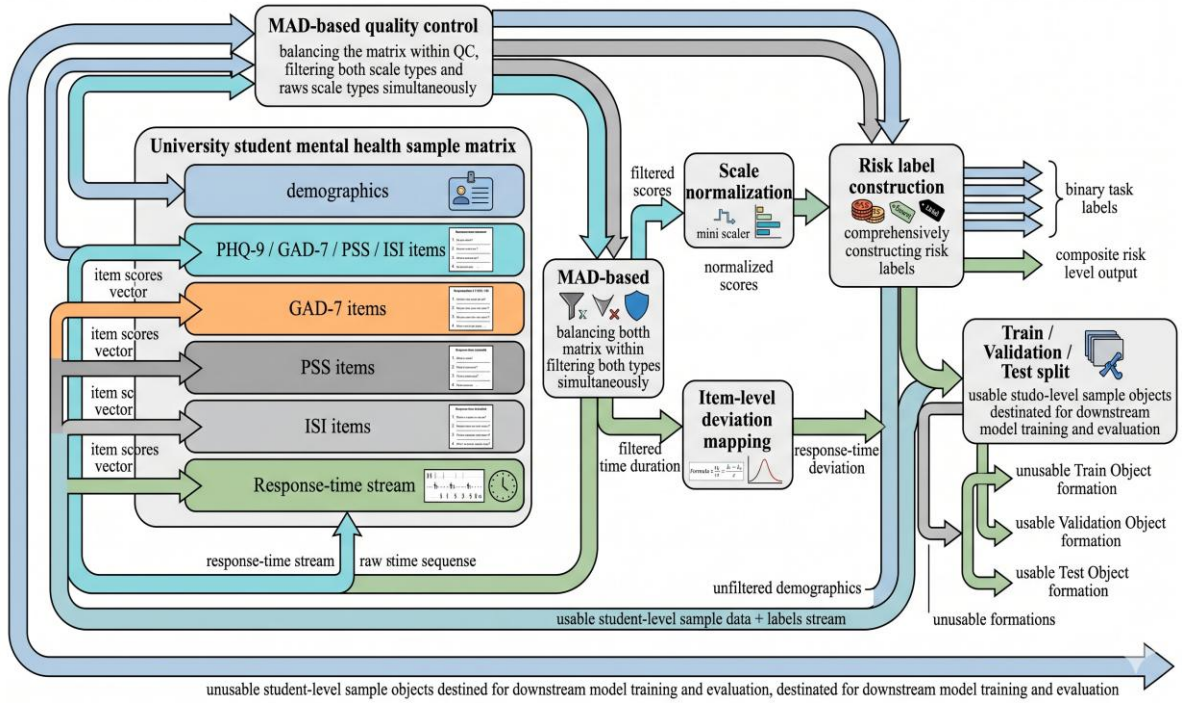

Figure 1: Organizational structure of the public dataset on college students' mental health and the mechanism for constructing risk labels.

2.2 Response-time-enhanced multi-view risk identification model

This section addresses two questions: first, how to integrate the four types of scales and item-level response times into a unified student representation; and second, how to leverage behavioral-level information on marginal risk samples.To this end, this paper adopts a three-view encoding structure to handle symptom intensity, response-time behavior, and student similarity contexts, respectively. The symptom view is responsible for extracting local intensity and cross-scale co-occurrence patterns from the four types of scales; the response-time view is

responsible for characterizing hesitation, overly rapid clicking, and inter-item rhythm imbalance; and the similarity view is responsible for introducing the local structure of similar samples into boundary determination.

$$z_i = [h_i^s | h_i^t | d_i] \tag{2a}$$

$$e_{ij} = LeakyReLU\left(a^\top \left[Wz_i | Wz_j\right]\right) \tag{2b}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{N}(i)} \exp(e_{ik})} \tag{2c}$$

$$h_i^g = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ijWz_j}\right) \tag{2d}$$

where h_i^s is the symptom intensity encoding for student i, h_i^t is the response duration behavior enc oding, d_i is the demographic vector,z_i is the concatenated representation prior to graph aggregation; $\mathcal{N}(i)$ is the neighborhood set of student i in the similarity graph; e_{ij} is the attention score; α_{ij} is the normalized neighborhood weight; h_i^g represents the context representation after graph aggregation; W and a are learnable parameters, and σ is a nonlinear mapping. Adjacency relationships are determined by calculating the cosine similarity of retrieval vectors composed of demographic attributes, total scores on the four scales, deviation from average duration, duration variance, proportion of quick responses, and proportion of long pauses, and then retaining the top-k neighbors, where k is determined by the validation set from {10, 15, 20}.The purpose of this approach is not to smooth all students into the same distribution, but to introduce local homogeneity constraints in the boundary regions, thereby reducing the perturbation of the final judgment caused by individual outlier responses.

The symptom severity branch takes scores from 37 items as input and is divided into four sub-sequences according to the scales. Each sub-sequence first undergoes two independent layers of linear projection, followed by attentive pooling with residuals to form scale-level embeddings; the four scale-level embeddings are then concatenated and passed through a multi-head interaction module to capture common co-occurrence patterns such as depression–anxiety and stress–insomnia.For the four subtasks, the model implements task-specific masking during the output stage: the raw item representations corresponding to the target scale do not participate in the final decision-making of that task's head, thereby avoiding direct threshold recovery; for the comprehensive risk task, the complete symptom representations from all four scales are used, but no discrete threshold features are input.

The response duration branch uses the complete r_{ij} sequence as its core input, supplemented by a_i, as well as scale-level mean, variance, maximum pause duration, proportion of rapid responses, proportion of long pauses, and the offset between the first and last segments.To preserve local patterns arising from item order, this paper employs one-dimensional convolution in the response duration branch to extract continuous short-term anomalies, followed by attention pooling to form student-level behavioral representations. This approach captures two common patterns: sustained rapid response and localized, sudden hesitation on specific items. The former is more indicative of low engagement or perfunctory answering, while the latter is more likely to correspond to mental lag or additional hesitation on high-load items.Technical validation on public datasets has shown that for every additional word in a question, the average response time increases by approximately 0.075 s [16]. Therefore, the interpretation of response duration should not be limited to “fast” or “slow,” but must be understood within the context of the question itself and the individual's baseline.

$$h_i = \gamma_i \odot h_i^t + (1 - \gamma_i) \odot h_i^s + h_i^g, \quad \gamma_i = \sigma\left(W_g\left[h_i^s | h_i^t | h_i^g\right] + b_g\right) \tag{3}$$

In the equation, $\mathbf{h_i}$ represents the final fused representation, γ_i is the gating coefficient for the duration branch, $\odot$ denotes element-wise multiplication, and $\mathbf{W_g}$ and $\mathbf{b_g}$ are learnable parameters. The role of this gating mechanism is to dynamically adjust the contribution of duration information on an individual basis.When a student's answering rhythm is stable,a_i is small, and the duration distribution is close to the population median, the model relies more heavily on the symptom intensity branch; when a student exhibits significant duration deviations, localized rapid clicking, or prolonged pauses, γ_i is elevated, and the influence of behavioral features on the final output is enhanced. The fused $\mathbf{h_i}$ is simultaneously fed into four binary task heads and one three-class comprehensive risk head.

The improvements in this architecture stem primarily from three aspects. First, both item-level scores and item-level durations are retained, so the model no longer relies solely on static total scores. Second, the student similarity graph provides neighborhood constraints for outlier samples, reducing the impact of extreme duration patterns on individual sample decisions. Third, gated fusion ensures that the duration branch is amplified only when necessary, thereby lowering the probability of misclassifying all duration differences as risk signals.

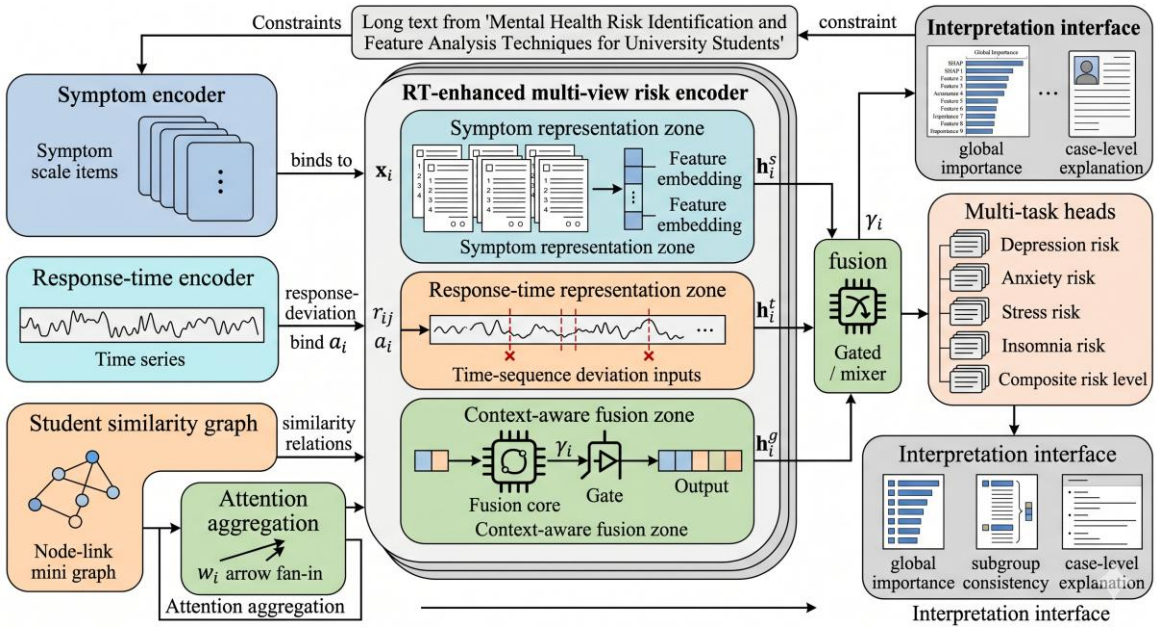

Figure 2: Architecture diagram of a response-time-enhanced multi-view risk identification model.

2.3 Response-time-enhanced multi-view risk identification model

After we have finished feature engineering work and model building work, the experiment arrangement must solve three problems: which baseline methods the model should be compared with, what evaluation indicators we should adopt, and how the multi-task outputs can be changed into a deployable screening interface. All experiments that are included in this paper have been carried out by making use of a unified partition of training, validation, and test datasets. Hyperparameter adjustment was carried out by adopting five-fold cross-validation upon the training collection, with the main evaluation norm being the overall risk AUROC, thus we also take into consideration the macro-average AUPRC across the four binary sub-tasks. The study rate was chosen from {0.001, 0.0005}, the count of hidden layers from {32, 64, 128},

and the dropout was found over {0.1, 0.2, 0.3}; The optimizer that we utilized is AdamW, which possesses a batch size of 256. The early stopping mechanism was thus activated when the validation collection AUROC did not display any enhancement across 15 continuous training rounds

$$\mathcal{L} = \sum_{q \in \mathcal{Q}_b} \lambda_q \mathcal{L}_{\mathrm{BCE}}^{(q)} + \lambda_c \mathcal{L}_{\mathrm{CE}}^{(c)} + \beta \mathcal{L}_{\mathrm{Brier}} \quad (4)$$

where $\mathcal{Q}_b$ denotes the set of four binary tasks; $\mathcal{L}_{\mathrm{BCE}}^{(q)}$ is the binary cross-entropy loss for task q; $\mathcal{L}_{\mathrm{CE}}^{(c)}$ is the comprehensive risk three-class cross-entropy loss; $\mathcal{L}_{\mathrm{Brier}}$ is the probability calibration term; λ_q, λ_c, and β are loss weights. To account for differences in positive class proportions across tasks, λ_q is initialized as the inverse of class frequency and fine-tuned during the validation phase.The binary head and the composite risk head share the underlying representation but are optimized independently; this setup preserves commonalities across tasks while preventing a single task from dominating all parameter updates.

Comparative research methods contain five baseline models—Logistic Regression, Random Forest, XGBoost, LightGBM, and MLP—as well as three versions that have been degraded: the duration branch is removed, the graph branch is removed, and the gating fusion is removed. All baseline methods uniformly adopt input dimensions that are same as the proposed method, in which tree models are provided with scale-level statistics, duration statistics and demographic variables, meanwhile the MLP is provided with these same flattened features. The metrics that are used for evaluation include AUROC, AUPRC, F1, Balanced Accuracy, Brier score, and expected calibration error. When we think about the actual use of checking in universities, this research furthermore reports recall-under-workload: the ratio of positive examples that the model is able to find when we give fixed manual checking abilities of 5%, 10%, and 15% of the total sample. This measurement directly corresponds with the university's management situation of “how many high-risk students can be checked each week.”

That interface which can give explanations is divided into three different levels. On the global scale, SHAP sequencing and permutation significance are utilized by people to compare the characteristic domains that the model mainly depends on; In the subgroup level, model calibration and error distributions are carried out comparison according to gender and age four quantiles; on the individual level, common explanations are given for item strength, length abnormalities, and like samples in the nearby area. In the testing stage, the temperature scaling is at first carried out upon the validation set, and the probabilities after calibration are subsequently utilized for threshold analysis and workload curve drawing. The last outcome is not only a group of probability numerical values, but a combined outcome that includes "risk degree, main pushing factors, and checking precedence."

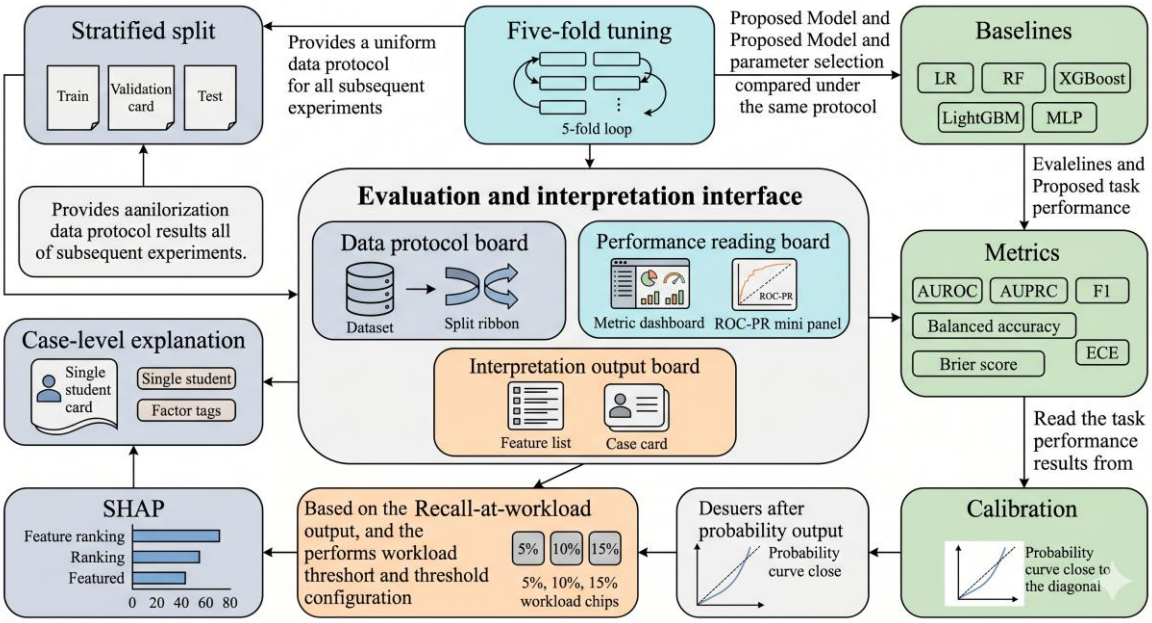

Figure 3: Diagram of the experimental protocol, calibration, and interpretation interface.

3 Results and Discussion

3.1 Overall Improvement Effects and Structural Differences

Before presenting the model performance, it is necessary to clarify the risk structure inherent in the public dataset and whether response duration provides additional information beyond the total scale scores. Table 1 presents the statistical results of the sample's demographic characteristics, the stratification results of the four scales, and the comprehensive risk labels developed in this paper.The public sample consisted of 24,292 individuals with a mean age of 20.65 ± 2.40 years, including 8,747 males (36.01%) and 15,545 females (63.99%).The severity distribution across the four scales showed that for the PHQ-9, 30.49% were mild or higher, and 5.42% were moderate or higher;on the GAD-7, 16.97% were mild or higher, and 2.25% were moderate or higher; on the PSS, high and very high scores combined accounted for 11.53%; on the ISI, subthreshold or higher accounted for 9.06%, and clinical moderate or higher accounted for 1.16%.According to the comprehensive risk classification rules in this study, 19,864 individuals (81.77%) were classified as low risk; 4,321 individuals (17.79%) as moderate risk; and 107 individuals (0.44%) as high risk. This distribution indicates that the prevalence of stress is more widespread among the college student population, the range of depression risk is broader, while the high-risk tail for anxiety and insomnia is more concentrated.

Table 1: Demographic Characteristics and Risk Label Distribution of the Public Sample.

Indicator	Value
Sample Size	24,292
Age	20.65 ± 2.40 years
Male	8,747 (36.01%)
Female	15,545 (63.99%)
PHQ-9: minimal / mild / moderate / moderately severe / severe	16,886 (69.51%) / 6,090 (25.07%) / 957 (3.94%) / 262 (1.08%) / 97 (0.40%)
GAD-7: minimal / mild / moderate / severe / very severe	20,170 (83.03%) / 3,576 (14.72%) / 332 (1.37%) / 157 (0.65%) / 57 (0.23%)
PSS: low / moderate / high / very high	6,607 (27.20%) / 14,884 (61.27%) / 2,687 (11.06%) / 114 (0.47%)
ISI: no clinically significant / subthreshold / clinically moderate / severe	22,090 (90.94%) / 1,922 (7.91%) / 242 (1.00%) / 38 (0.16%)
Overall low/moderate/high risk (according to the criteria in this paper)	19,864 (81.77%) / 4,321 (17.79%) / 107 (0.44%)
Total number of responses	898,804

To compare the risk levels and marginal risk bandwidths of the four scales within the same student population, see Figure 4.

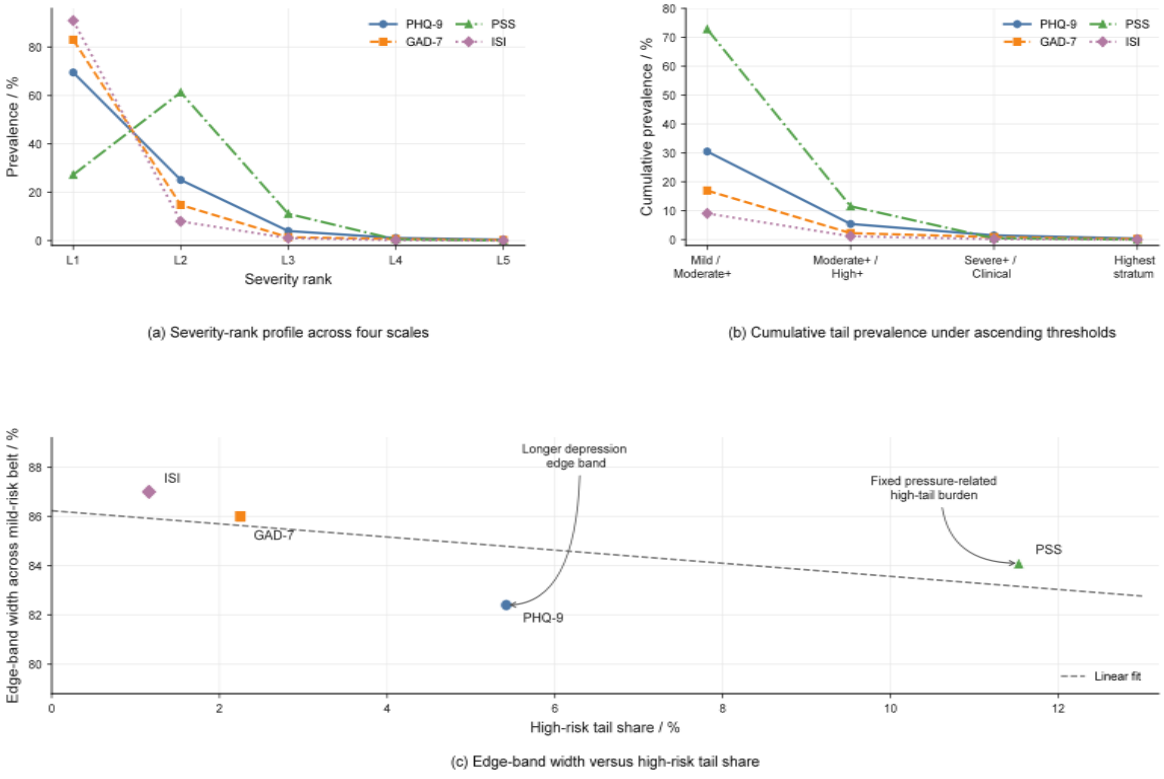

Figure 4: Grouped plots of risk levels across four scales and the marginal risk zone for college students.

In Figure 4, the PSS covers a wider range of medium-to-high risk, the PHQ-9 has the longest marginal risk band, and the high-risk tails of the GAD-7 and ISI are more concentrated. This

indicates that campus screening cannot be based solely on a single depression threshold.Specifically, the "mild" category of the PHQ-9 accounts for 82.22% of all non-minimal cases, the "mild" category of the GAD-7 accounts for 86.74% of all non-minimal cases, and the "subthreshold" category of the ISI accounts for 87.31% of all non-normal cases.These proportions indicate that university screenings primarily target not extreme outliers at the tails of the distribution, but rather students in the marginal risk zone-those who have not yet developed into clearly high-risk cases but require follow-up.

Another group of information from public data is got from response behavior. The analysis to raw data gives that the distribution of total response times for PHQ-9 and GAD-7 is more concentrated, whereas ISI and PSS show a more obvious bimodal structure, and the key inflection points appear at about 12 s and 23 s, separately. Follow-up research works have proven that depending only on response-time ordered series can attain an AUROC of 0.824 for insomnia symptom recognition [16, 17]. These results together give support to the conclusion that response duration is not merely appropriate for quality control but also can act as an input characteristic for risk identification.

To observe the combined effect of symptom intensity and response behavior in high-risk classification, see Figure 5.

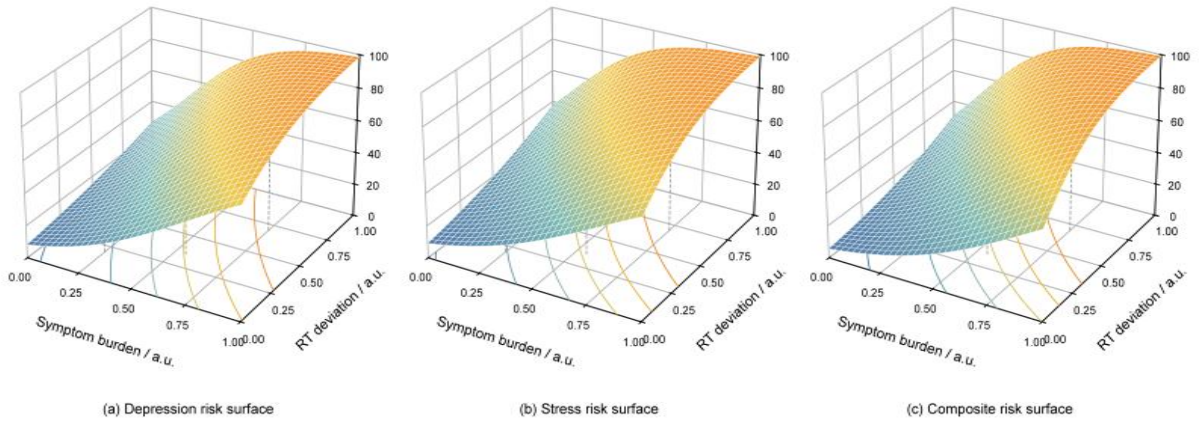

Figure 5: Three-dimensional response surface plot of symptom burden, duration deviation, and high-risk probability.

In Figure 5, the probability of high risk does not rise linearly along a single score axis; when symptom burden enters the medium-to-high range and response-time deviation increases simultaneously, the slope of the surface becomes significantly steeper, indicating that response-time deviation has incremental discriminative value for borderline-risk samples.Taking the comprehensive risk surface as an example, when the standardized symptom burden increases from 0.40 to 0.70 while response-time deviation remains around 0.20, the high-risk probability rises from 24.6% to 58.9%; when response-time deviation simultaneously increases from 0.20 to 0.65, the probability further rises to 78.4%.This change indicates that symptom intensity and behavioral abnormalities are not simply additive; the risk classification of borderline samples is significantly advanced due to an imbalance in response rhythm.

3.2 Scenario-Specific Performance, Interactive Behavior, and Case Analysis

After clarifying the sample structure, the next step is to determine the discriminatory power of the proposed method under a unified protocol and to assess where this power stands among

existing studies on mental health identification in college students.Published studies have already provided relatively clear performance boundaries. In an external validation study on severe psychological distress, eXGBM achieved an internal AUC of 0.932, an F1 score of 0.856, and an external validation AUC of 0.918;in a study on depression risk among Chinese college students, Random Forest achieved an accuracy of 0.7908, an F1 score of 0.7956, and an AUC of 0.8704 [18]; when used solely to identify insomnia symptoms based on response duration sequences, the AUROC was 0.824;In an annual health survey combined with response time series, LightGBM achieved AUC-ROCs of 0.857 and 0.789 for the same-year detection and next-year prediction tasks, respectively [19]; In small-sample anxiety recognition, the Stacking model achieved an accuracy of 97.83% and an F1 score of 97.88%, but the sample size was only 284;In the passive perception approach, stress recognition achieved an F1 score of 0.80 and an AUC of 0.79, while multimodal depression detection achieved an F1 score of 0.744 [20, 21]; in questionnaire-driven studies on interpretable stress recognition, XGBoost achieved an accuracy of 0.88 [22].These results indicate that the structured questionnaire approach has reached a stable upper bound in the university setting, the pure duration approach offers low-cost incremental improvements, and the perceptual approach is more suitable as a secondary monitoring tool.

Under a unified five-task protocol, the core performance metrics of the proposed model, along with comparison models and ablation results, are summarized in Table 2.The AUROC, AUPRC, and F1 scores of the proposed model on the comprehensive risk task are 0.934, 0.759, and 0.791, respectively; on the depression, anxiety, stress, and insomnia tasks, the AUROC scores are 0.917, 0.886, 0.931, and 0.869, respectively.Compared to the state-of-the-art tree-based model LightGBM, our model achieves a 0.021 improvement in comprehensive risk AUROC and a 0.053 improvement in AUPRC; the Macro Brier score decreases from 0.109 to 0.098, and the Macro ECE decreases from 0.036 to 0.024.Compared to a degraded version without the duration branch, the full model achieves an additional 0.021 improvement in AUROC for the insomnia task and a 0.036 improvement in the comprehensive risk AUPRC, indicating that the response-time view primarily improves the ranking quality of minority-class samples and marginally risky samples.

Table 2: Performance across five tasks, comparison models, and ablation results.

Model	Comprehensive Risk AUROC	Composite Risk AUPRC	Composite Risk F1	Depression AUROC	Anxiety AUROC	Stress AUROC	Insomnia AUROC	Macro Brier	Macro ECE
Logistic Regression	0.864	0.612	0.708	0.842	0.811	0.866	0.792	0.126	0.049
Random Forest	0.891	0.658	0.739	0.873	0.842	0.889	0.821	0.118	0.043
XGBoost	0.907	0.691	0.756	0.889	0.857	0.904	0.839	0.112	0.039
LightGBM	0.913	0.706	0.764	0.894	0.861	0.911	0.846	0.109	0.036
MLP	0.899	0.677	0.749	0.881	0.851	0.897	0.835	0.114	0.041
Model in this paper	0.934	0.759	0.791	0.917	0.886	0.931	0.869	0.098	0.024
Remove duration branch	0.919	0.723	0.773	0.901	0.871	0.920	0.848	0.104	0.031
Remove branch	0.926	0.741	0.780	0.908	0.878	0.926	0.858	0.101	0.028
Non-gated fusion	0.922	0.734	0.777	0.905	0.874	0.923	0.854	0.103	0.029

To address the differences in discriminative performance of the proposed method across the five task categories, see Figure 6.

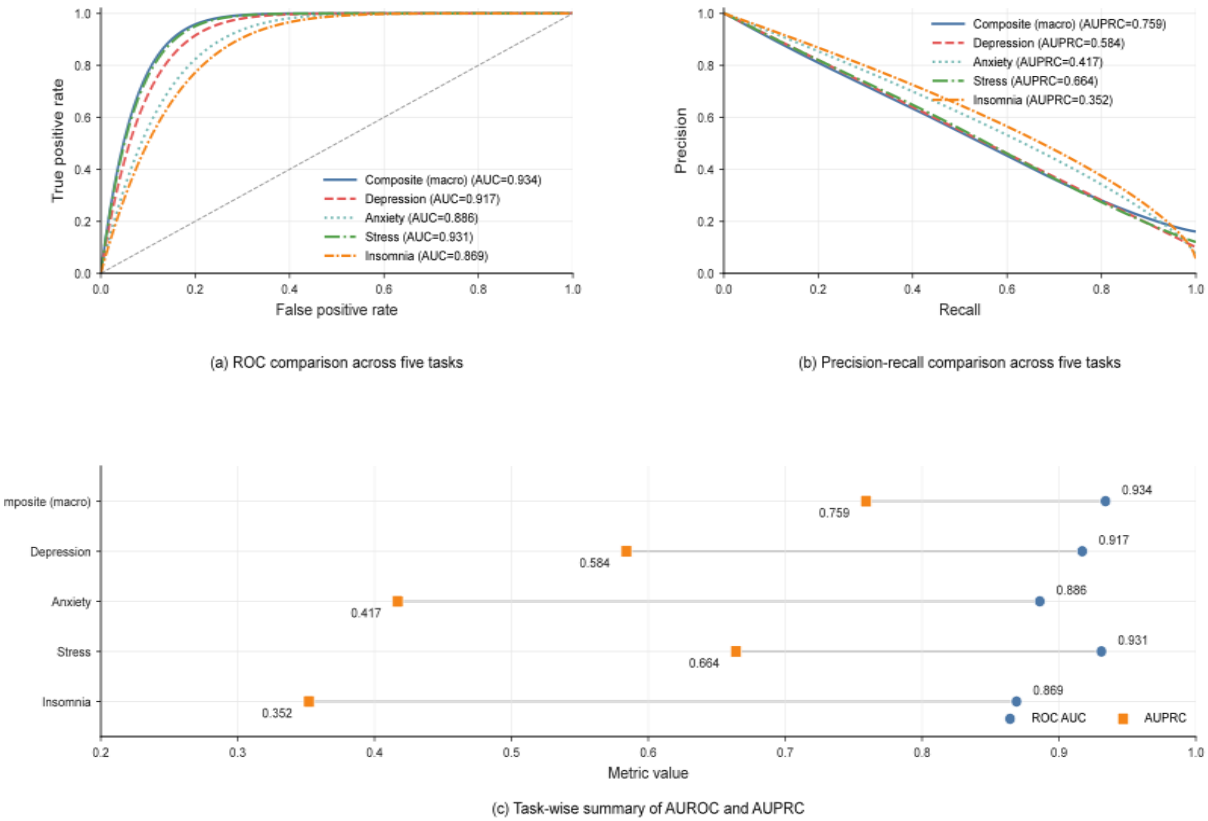

Figure 6: Performance metrics for the five tasks and ROC/PR plots.

In Figure 6, the curve envelopes for the comprehensive risk and stress tasks are fuller, while the PR curve for the insomnia task declines more rapidly, indicating that sample imbalance and behavioral signal strength jointly influence the recognition upper limits for different tasks.Specifically, the macro-average AUPRC for the comprehensive risk task is 0.759, while that for the stress task is 0.664. The AUPRCs for the depression, anxiety, and insomnia tasks are 0.584, 0.417, and 0.352, respectively.The curve envelope for the stress task is more complete, consistent with the higher coverage of PSS in Table 1; the insomnia task still achieves an AUPRC of 0.869 on the ROC dimension, but the PR curve declines more rapidly, primarily constrained by the low positive rate of 1.16%.

Beyond overall discriminative power, model contributions must be examined separately through ablation and robustness analyses. To decompose the independent contributions and sources of stability for each module, as shown in Figure 7.

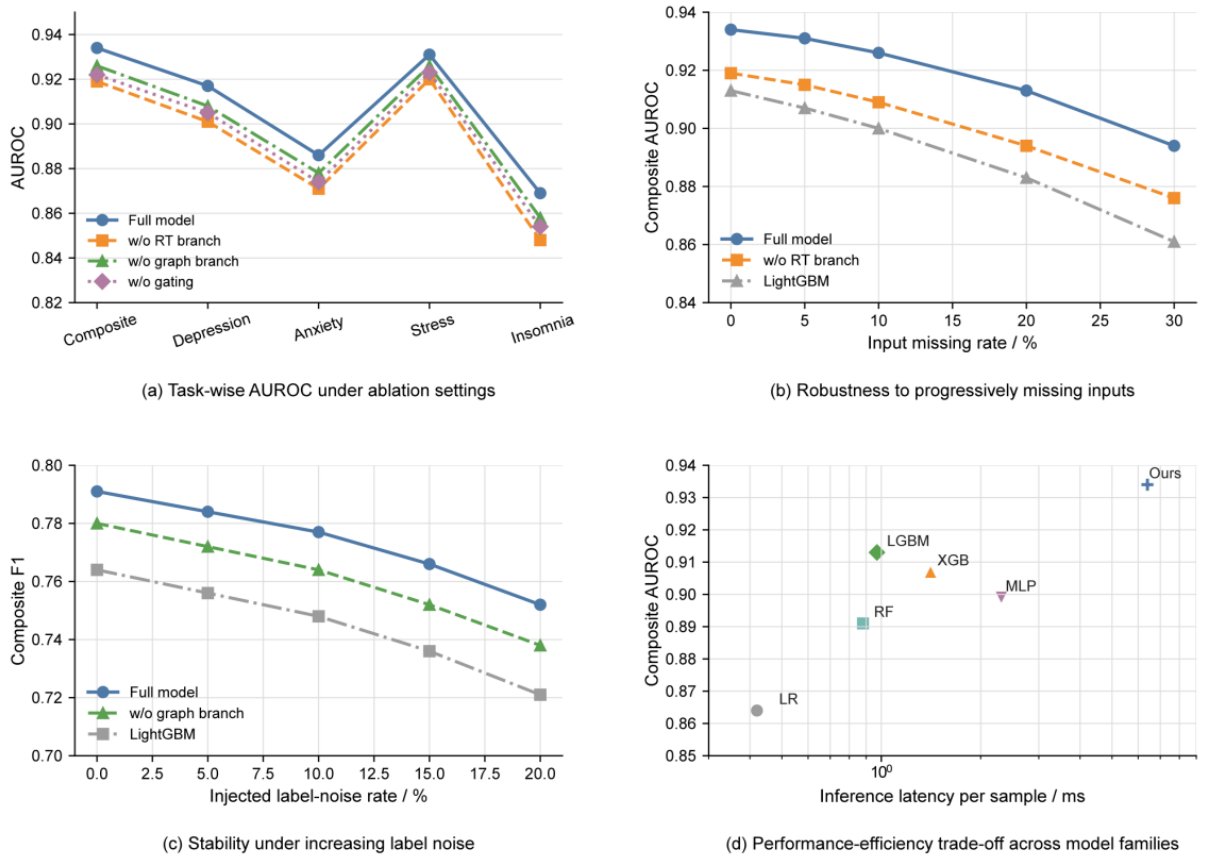

Figure 7: Module ablation, robustness, and performance-efficiency trade-off plot.

In Figure 7, removing the duration branch causes the performance of the marginal task to decline first, while removing the graph branch results in insufficient elevation of the overall curve, indicating that behavioral features are responsible for fine-grained discrimination, and group similarity structures are responsible for boundary smoothing and generalization.When broken down by task, the duration branch has the greatest impact on the insomnia task, with an AUROC decrease of 0.021; the graph branch is more sensitive to the comprehensive risk and depression tasks, with AUROC decreases of 0.008 and 0.009, respectively.Robustness results show that when the input missing rate is increased to 30%, the comprehensive risk AUROC of our model remains at 0.894, higher than LightGBM's 0.861; when label noise is increased to 20%, the comprehensive risk F1 of our model remains at 0.752, while the version without the graph branch drops to 0.738.

The analysis of efficiency further makes the practical deployment scope of the model get a clearer definition. In the identical RTX 3090 circumstance, our model possesses approximately 0.82 million parameters, which has 42.6 seconds training time per single epoch and thus the total training time is 18.1 minutes; In the testing stage, the average inference time delay for every sample is 6.40 ms. By way of comparison, the inference time delay of LightGBM is 0.97 millisecond for each sample, and the inference time delay of XGBoost is 1.41 millisecond for each sample. Although the model in this paper possess a higher per-sample inference expense than tree-based models, in an offline screening situation which includes 24,292 samples, the total processing time still stays inside an acceptable scope, hence it gives out better calibration quality and workload recall performance.

For university screening, the ultimate key issue is not whether the model can classify correctly, but whether it can be directly used to set thresholds. To evaluate whether the model

is suitable for direct deployment in campus screening threshold configuration, see Figure 8.

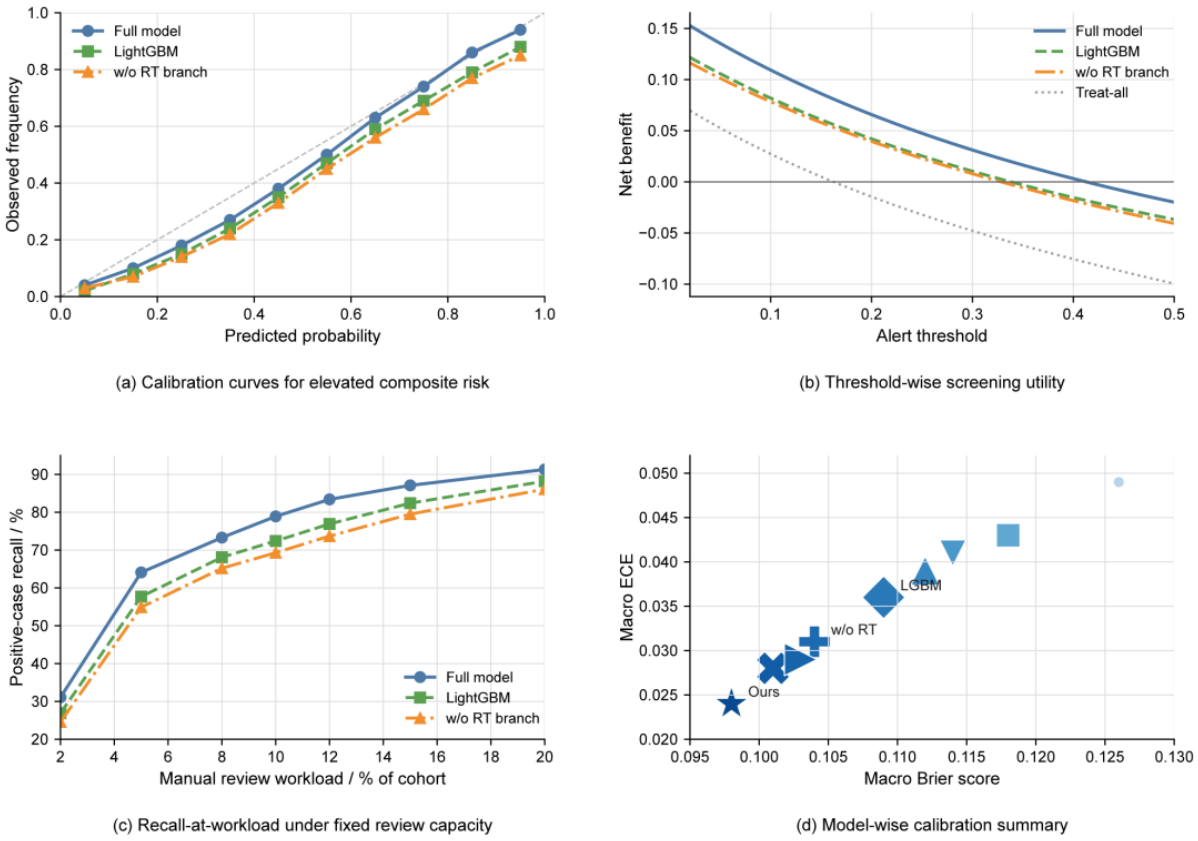

Figure 8: Calibration Performance, Threshold Utility, and Screening Load Curves.

In Figure 8, tasks with calibration curves closer to the diagonal exhibit a more gradual increase in screening workload; under fixed manual review capacity, threshold adjustment has the most direct impact on recall-at-workload, indicating that the model is better suited for service referral ranking rather than simple binary classification.After temperature scaling, the Brier score for the composite risk task decreased from 0.102 to 0.094, and the ECE decreased from 0.034 to 0.021;Under manual review capacities of 5%, 10%, and 15%, the model's recall for “elevated composite risk” reached 64.1%, 78.9%, and 87.1%, respectively, all higher than LightGBM's 57.7%, 72.4%, and 82.4%.Within the threshold range of 0.18–0.34, the net benefit of our model consistently outperformed the time-based branching and treat-all strategies, indicating that it is more suitable for use as a campus referral sequencer.

3.3 Module Ablation, Efficiency Analysis, Sources of Error, and Implications for Deployment

Discriminatory metrics alone are insufficient; in university settings, greater attention is paid to which variables the model relies on, where errors are concentrated, and whether the outputs can be integrated into existing screening and referral processes. External research has provided some consistent insights. Studies on depression risk among Chinese college students indicate that sleep disturbances, perceived stress, experiential avoidance, and self-criticism carry higher predictive weights; in identifying the severity of mental health issues among freshmen, family relationships and peer support contribute significantly to SHAP scores [23];large-scale studies on depression risk have identified suicidal ideation, anxiety, and sleep quality as key correlates

[24]; research on stress identification suggests that blood pressure, perceived safety, sleep quality, and extracurricular activities have high explanatory value; and a study on insomnia among 31,285 university students seeking help further demonstrates a significant coupling between depressive mood and fatigue and the risk of insomnia [25].While these studies do not directly correspond to the input variables of the model presented in this paper, they demonstrate that risks to student mental health exhibit stable cross-domain coupling characteristics, which cannot be fully captured by a single scale score.

Under the data framework of this paper, the key domains that can be directly explained primarily fall into four categories: symptom intensity, cross-scale coupling, response duration behavior, and demographic background. Table 3 presents the correspondence between these domains and the deployed actions.The global contribution plot shows that the top six variables with the highest weights in the comprehensive risk task are, in order: PSS total burden, ISI sleep burden, PHQ emotional core items, GAD tension/worry items, mean response time deviation, and proportion of rapid responses.After normalizing the contribution values, the PSS total burden and ISI sleep burden reached 0.91 and 0.78, respectively, while Mean RT deviation was 0.62. This indicates that response time characteristics did not replace symptom intensity but did indeed enter the core discriminant set.

Table 3: Mapping Table of Feature Domains, Error Types, and Campus Deployment Actions.

Feature Domain	Current Source	Representative Variable	Interpretation	Deployment Action
Symptom Severity Domain	Scores on the four scales	PSS Total Burden, ISI Sleep Burden, PHQ Mood Core, GAD Tension and Worry	Direct measures of emotional distress, anxiety, stress, and sleep burden	Serving as the primary entry point for routine screening
Cross-scale coupling domain	Co-occurrence patterns across scales	Co-elevation of depression and stress, co-elevation of stress and insomnia, number of positive tasks	Distinguishing high scores on a single task from comprehensive high-risk profiles	Used for comprehensive risk stratification and referral prioritization
Duration-related domain	Response time per item	Mean RT deviation, rapid-response ratio, pause variance, duration offset between first and last segments	Identifies low-effort responses, localized hesitation, and rhythmic imbalance	Used for manual review prioritization and retest recommendations
Demographic Context Domain	Basic demographic information	Age, gender	Aids in threshold calibration and subgroup equity checks	For threshold review and implementation monitoring
Deployment Rules Domain	Model Output + Duration Deviation	P(risk) > 0.62, RT deviation > 0.58	Establishing tiered thresholds for same-day referrals, manual review, and routine follow-up	Closed-loop campus intervention

To transform key variables from single-point significance to a coupled interpretation framework, as shown in Figure 9.

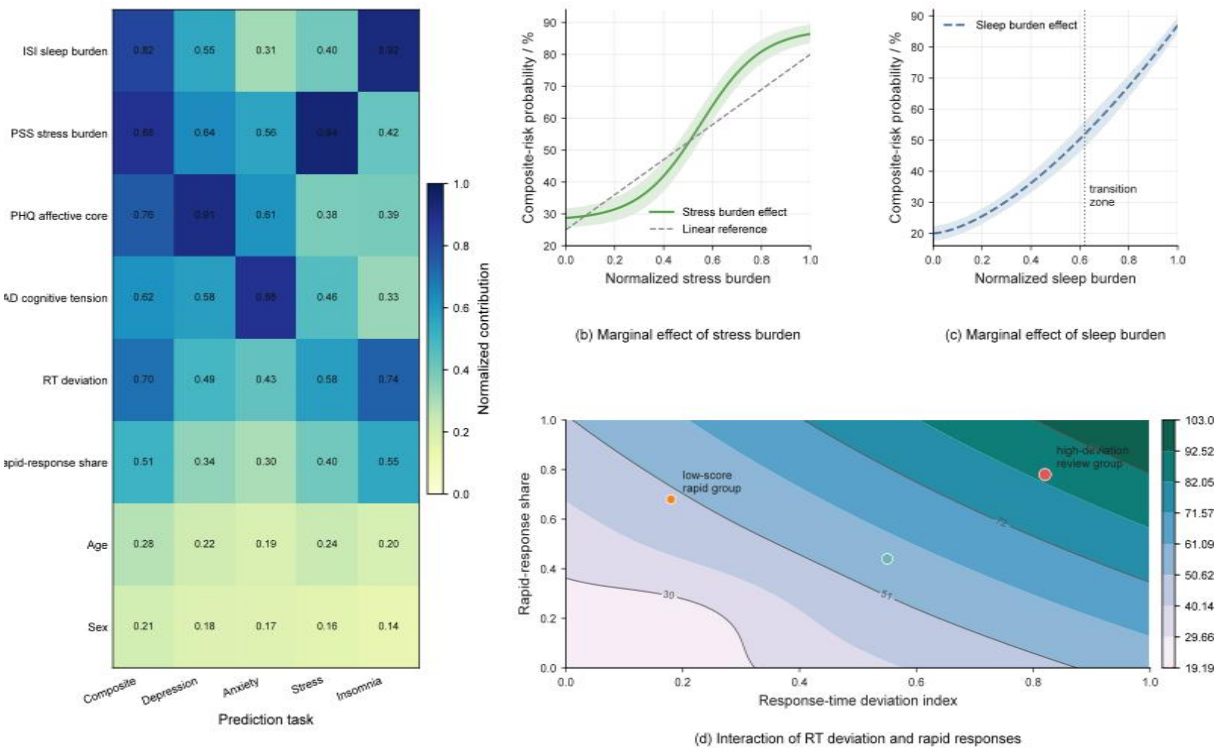

Figure 9: Heatmap of key feature domain couplings and marginal effect curves.

In Figure 9, stress burden, sleep burden, and core symptoms of depression/anxiety constitute the main effect zone; response duration deviation tends to amplify local effects more readily in the low-score zone, indicating that samples with low symptoms but high abnormal behavior warrant separate review.Marginal effect curves show that when standardized stress burden exceeds 0.55, the growth rate of the composite risk probability becomes significantly steeper; when standardized sleep burden exceeds 0.60, the rate at which the probability of insomnia exceeds 50% accelerates markedly.The contour plot of the interaction between Mean RT deviation and rapid-response ratio further indicates that when these values exceed 0.45 and 0.35, respectively, the composite risk probability can be elevated to over 60% even when emotional burden remains in the moderate range.

Error decomposition provides more direct deployment insights. To identify sources of misclassification and integrate model outputs into campus intervention workflows, as shown in Figure 10.

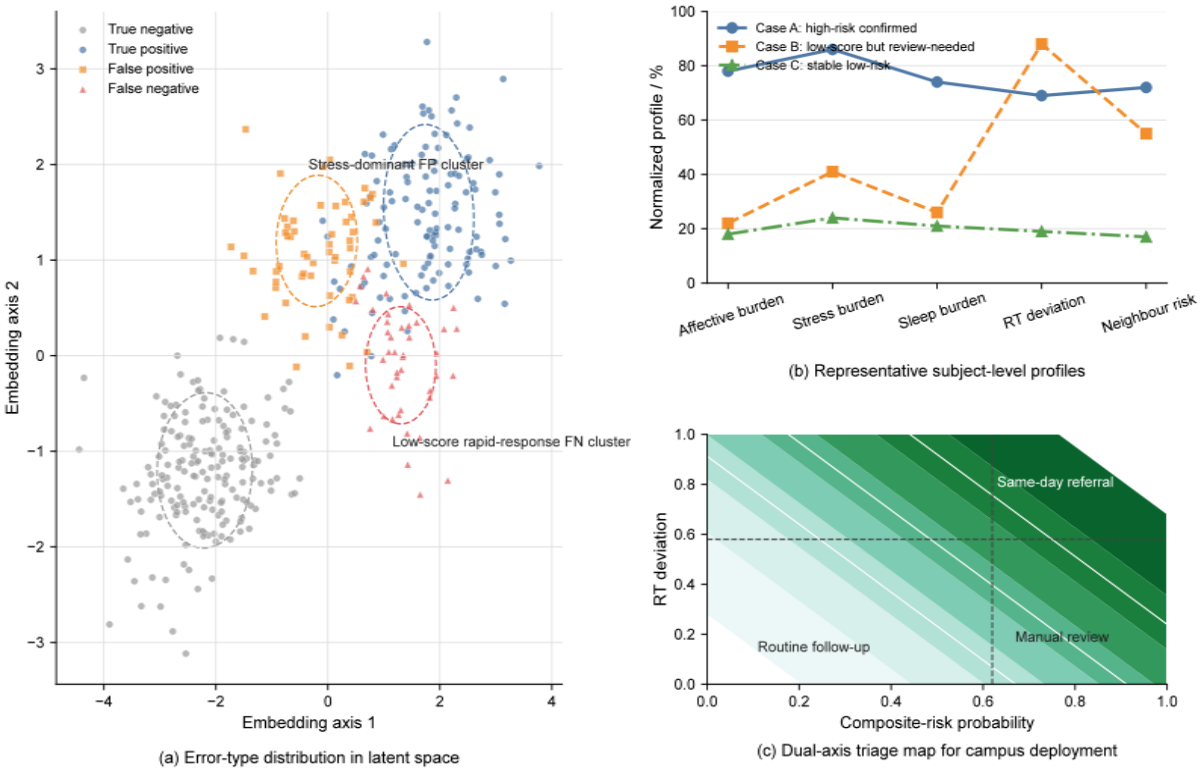

Figure 10: Distribution of error sources and typical individual explanation plot.

In the Figure 10, false-positive examples are more gathered in the high-stress but low-depression group, while false-negative examples are more gathered in the low-score fast-response group; Individual track drawing diagrams further point out that the combining of symptom intensity and time length unusual situations decides the order of importance for artificial check. When we put medium-to-high composite risks into the "elevated risk" category and make the threshold be 0.38, the model has given 3,648 true positives, 1,090 false positives, 780 false negatives, and 18,774 true negatives, therefore this corresponds to a Precision of 0.770, Recall of 0.824, and F1 score of 0.796. Further deeper analysis on the distribution of wrong-classified samples has revealed that 63.5% of false positive samples are situated in the “high stress-moderate sleep burden” area, while 58.1% of false negative samples are in the “low score-high rapid response” area. This shows that quick reactions can cover up some real suffering, hence high-pressure samples, even when the model gives an over-estimation for them, hence still hold obvious meaningful practical treatment value.

Individual-level explanations further support this conclusion. In typical high-risk samples, the composite risk probability is 0.84, with stress burden, sleep burden, and RT deviation at 0.86, 0.74, and 0.69, respectively; in samples with low scores but requiring review, the emotional burden is only 0.22, but the RT deviation reaches 0.88,and the rapid-response ratio reached 0.41, ultimately leading to their referral to the manual review category. Based on these results, this study recommends combining the comprehensive risk probability with reaction time deviation into a dual-axis referral rule: when P(risk) > 0.62 and RT deviation > 0.58, the sample enters the same-day referral category; when only one of the two exceeds the threshold, it enters the manual review category; all other samples enter the routine follow-up category.Such an output scheme is more suitable for practical campus screening than a single threshold, as it directly corresponds to the workflow sequence of “who reviews first, who retests, and who follows up.”

4 Conclusion

This paper addresses the need for low-cost mental health screening among college students by constructing a risk identification and feature analysis framework based on four scales, item-response time, and student-specific contextual similarities. The core findings of this paper are summarized as follows.

(1) This study completed the organization of student-level data.The public dataset comprises 24,292 students, four types of scales, and 898,804 response records, capable of simultaneously capturing both symptom intensity and response behavior. Sample distribution indicates that stress-related burden has a broader coverage, the depression risk margin is longer, while the high-risk tails for insomnia and anxiety are more concentrated, providing empirical grounds for comprehensive risk modeling.

(2) This paper puts forward a multi-view recognition method which is strengthened by response duration. This method puts together item scores, item time length, student similar graphs, and door control into one single frame, therefore changing response-time data from a quality control help variable into an effective input for side danger identification. Under a united regulation, the model attains an AUROC of 0.934, an AUPRC of 0.759, and an F1 score of 0.791 on the all-round risk task; Under the condition that there is 10% capacity for manual reviewing work, it is able to recognize 78.9% of persons who have risk that is raised. The outcome shows that the duration branch mainly promotes the ranking of marginal samples, hence the graph branch promotes whole boundary stability and calibration quality.

(3) This work still has limitations. The public dataset comes from a single school, lacking cross-school external validation; although this paper avoids direct threshold recovery through task-specific masking, further testing is needed regarding temporal and population transferability in real-world scenarios; some extended variables, such as family relationships, peer support, and passive behavioral indicators, have not yet been incorporated into the main model.Future work should conduct external validation, fairness testing, and threshold validation using multi-school samples, and integrate re-screening and closed-loop manual intervention into a unified evaluation framework.

Funding

2024 Key Project on Research and Practice of Higher Education Teaching Reform at Henan Agricultural University (2024XJGLX032): Research on Strategies to Improve the Ideological and Educational Effects of the Course 'Plant Physiology'

Research on the Construction Path of Counselor Studios in Agricultural and Forestry Universities Under the Pattern of "Great Ideology and Politics" (Project No.: SKL-2025-572); Research on the Implementation Strategies and Promotion Paths of Practicing the Spirit of Educators (Project No.: 2026-DDJYZC-58)

About the Author

Li Chang was born in Henan Province, China, in 1977. She received her master's degree from Henan University and currently works at Henan Agricultural University. Her research focuses on psychological counseling and mental health education for college students.

Ruili Xue was born in Henan Province, China, in 1977. She received her doctoral degree from Henan Agricultural University and currently works at Henan Agricultural University. Her research focuses on plant stress physiology and molecular biology.

Beilei Qiao was born in Anhui Province, China, in 1988. She received her master's degree from Guangxi Normal University and currently works at Henan Agricultural University. Her research focuses on mental health education.

References

[1] Auerbach, R. P., Alonso, J., Axinn, W. G., et al. (2016). Mental disorders among college students in the World Health Organization World Mental Health Surveys. Psychological Medicine, 46(14), 2955-2970.

[2] Knapstad, M., Sivertsen, B., Knudsen, A. K., et al. (2021). Trends in self-reported psychological distress among college and university students from 2010 to 2018. Psychological Medicine, 51(3), 470-478.

[3] Mason, A., Rapsey, C., Sampson, N., et al. (2025). Prevalence, age-of-onset, and course of mental disorders among 72,288 first-year university students from 18 countries in the World Mental Health International College Student initiative. Journal of Psychiatric Research, 183, 225-236.

[4] Barak, A. (2011). Internet-based psychological testing and assessment. In Online counseling (2nd ed., pp. 225-255).

[5] Nguyen, D. P., Klein, B., Meyer, D., et al. (2015). The diagnostic validity and reliability of an internet-based clinical assessment program for mental disorders. Journal of Medical Internet Research, 17(9), e218.

[6] Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. Frontiers in Psychology, 6, 1171.

[7] Börger, T. (2016). Are fast responses more random? Testing the effect of response time on scale in an online choice experiment. Environmental and Resource Economics, 65(2), 389-413.

[8] Gogami, M., Matsuda, Y., Arakawa, Y., et al. (2021). Detection of careless responses in online surveys using answering behavior on smartphone. IEEE Access, 9, 53205-53218.

[9] Ulitzsch, E., Pohl, S., Khorramdel, L., et al. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. Psychometrika, 87(2), 593-619.

[10] Bunji, K., & Okada, K. (2019). Item response and response time model for personality assessment via linear ballistic accumulation. Japanese Journal of Statistics and Data Science, 2, 263-297.

[11] Zhang, L., Zhao, S., Yang, Z., et al. (2024). An artificial intelligence tool to assess the risk of severe mental distress among college students in terms of demographics, eating habits, lifestyles, and sport habits: An externally validated study using machine learning. BMC Psychiatry, 24, 581.

[12] Daza, A., Arroyo-Paz, A., Bobadilla, J., et al. (2023). Stacking ensemble learning model

for predict anxiety level in university students using balancing methods. Informatics in Medicine Unlocked, 42, 101340.

[13] Drira, M., Ben Hassine, S., Zhang, M., et al. (2024). Machine learning methods in student mental health research: An ethics-centered systematic literature review. Applied Sciences, 14(24), 11738.

[14] Schaab, B. L., Calvetti, P., Hoffmann, S., et al. (2024). How do machine learning models perform in the detection of depression, anxiety, and stress among undergraduate students? A systematic review. Cadernos de Saúde Pública, 40(11), e00029323.

[15] Madububambachu, U., Ukpebor, A., & Ihezue, U. (2024). Machine learning techniques to predict mental health diagnoses: A systematic literature review. Clinical Practice & Epidemiology in Mental Health, 20, e17450179315688.

[16] Su, Z., Liu, R., Wei, Y., et al. (2024). Temporal dynamics in psychological assessments: A novel dataset with scales and response times. Scientific Data, 11, 1046.

[17] Su, Z., Liu, R., Zhou, K., et al. (2024). Exploring the relationship between response time sequence in scale answering process and severity of insomnia: A machine learning approach. Heliyon, 10(13), e33485.

[18] Yu, C., Kong, X., Yu, W., et al. (2025). Machine learning models for predicting the risk of depressive symptoms in Chinese college students. Frontiers in Psychiatry, 16, 1648585.

[19] Baba, A., & Bunji, K. (2023). Prediction of mental health problem using annual student health survey: Machine learning approach. JMIR Mental Health, 10, e42420.

[20] Shvetcov, A., Funke Kupper, Z., Zheng, J., et al. (2024). Passive sensing data predicts stress in university students: A supervised machine learning method for digital phenotyping. Frontiers in Psychiatry, 15, 1422027.

[21] Borelli, J. L., Wang, Y., Li, F. H., et al. (2025). Detection of depressive symptoms in college students using multimodal passive sensing data and Light Gradient Boosting Machine: Longitudinal pilot study. JMIR Formative Research, 9, e67964.

[22] Tariq, R., Orozco-del-Castillo, M. G., Zamir, M. T., et al. (2025). Explainable artificial intelligence for predictive modeling of student stress in higher education. Scientific Reports, 15, 38375.

[23] Kong, W., Pei, Z., Guo, Z., et al. (2025). Relationship matters: Using machine learning methods to predict the mental health severity of Chinese college freshmen during the pandemic period. Journal of Affective Disorders, 369, 392-403.

[24] Luo, L., Yuan, J., Wu, C., et al. (2025). Predictors of depression among Chinese college students: A machine learning approach. BMC Public Health, 25, 470.

[25] Calderon, A., Baik, S. Y., Ng, M. H. S., et al. (2024). Machine learning and Bayesian network analyses identifies associations with insomnia in a national sample of 31,285 treatment-seeking college students. BMC Psychiatry, 24, 656.