Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SGRI-HF: A Semantic Group-Aware Gated Residual Interaction and Heterogeneous Fusion Framework for College Student Mental Health Risk Identification

Fengyue Zhang
Henan University of Urban Construction
Chunlei Shi
201931210011@mail.bnu.edu.cn

Henan University of Urban Construction
Junfeng Zhang
Henan University of Urban Construction

Research Article

Keywords: College student mental health, Semantic group modeling, Gated residual interaction, Heterogeneous ensemble learning

Posted Date: August 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10633968/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SGRI-HF: A Semantic Group-Aware Gated Residual Interaction and Heterogeneous Fusion Framework for College Student Mental Health Risk Identification

Abstract

Early identification of college students with elevated mental health risk is essential for timely psychological intervention. However, questionnaire-derived psychosocial variables are heterogeneous and involve complex relationships across multiple domains, while existing methods often treat individual items as independent predictors and overlook semantic structures and individual differences in domain contributions. To address these limitations, a Semantic Group-Aware Gated Residual Interaction and Heterogeneous Fusion (SGRI-HF) framework is proposed for college student mental health risk identification. The study involved 2,894 students, including 1,394 at-risk and 1,500 not-at-risk individuals, with 51 predictors organized into seven predefined semantic domains. Specifically, SGRI-HF employs domain-specific encoders to learn semantic representations, a sample-adaptive gating mechanism to estimate individualized domain importance, explicit interaction modeling to capture cross-domain dependencies, and a global residual pathway to preserve complementary information. Auxiliary semantic supervision and group-masking consistency regularization are further introduced to enhance representation robustness. The neural predictions are then combined with CatBoost through nonnegative convex weighting optimized exclusively on validation data. Under leakage-controlled stratified five-fold cross-validation, SGRI-HF achieved the best overall performance among the evaluated machine learning and deep learning methods, with an accuracy of 0.6465, a macro-F1 score of 0.6436, and a ROC-AUC of 0.6911. Interpretation results indicated that social support, lifestyle, digital behavior, and stress-related factors contributed substantially to risk identification. These findings demonstrate the effectiveness of semantic domain modeling, individualized interaction learning, and heterogeneous fusion for interpretable questionnaire-based mental health risk identification.

Keywords: College student mental health; Semantic group modeling; Gated residual interaction; Heterogeneous ensemble learning

1. Introduction

Mental health problems among college students have become an important concern in higher education and student well-being research. The transition to adulthood involves substantial changes in academic demands, interpersonal relationships, living environments, career planning, and social expectations. These factors may jointly influence students' psychological states through complex interactions with lifestyle behaviors, coping strategies, social support, and environmental conditions [1-3]. When potential mental health risks are not identified in a timely manner, opportunities for early support and intervention may be delayed, potentially affecting academic performance, social functioning, and long-term development. Therefore, scalable and data-driven approaches that

complement conventional psychological assessment procedures are increasingly needed for early risk identification.

Current mental health assessment mainly relies on standardized questionnaires and professional evaluations. Although questionnaires provide structured information regarding psychological and psychosocial conditions, large-scale implementation is often constrained by limited professional resources and difficulties in continuous monitoring. Moreover, mental health risk is influenced by multiple heterogeneous factors, including demographic characteristics, academic behaviors, lifestyle patterns, social relationships, and environmental stressors, which cannot be fully characterized by isolated indicators. Consequently, machine-learning approaches have been increasingly investigated as complementary tools for learning complex patterns from questionnaire-based data.

Despite recent advances, existing approaches still face several challenges. First, most machine-learning models treat questionnaire items as independent features, ignoring the semantic relationships among related psychosocial variables. Since questionnaire predictors are naturally organized into meaningful domains, neglecting such structures may limit domain-level representation learning and reduce interpretability. Second, existing methods generally estimate global feature importance but overlook individual differences in psychosocial patterns. The contribution of different domains may vary across students, requiring models capable of adaptively learning personalized domain importance and cross-domain interactions. Third, although neural networks and tree-based models capture complementary characteristics in structured data, existing fusion strategies mainly focus on prediction combination and rarely incorporate semantic structures and individual-specific variations during representation learning. Therefore, an effective framework that integrates semantic organization, personalized interaction modeling, and heterogeneous learning is still required.

To address these challenges, this study proposes a Semantic Group-Aware Gated Residual Interaction and Heterogeneous Fusion (SGRI-HF) framework for college student mental health risk identification. The proposed framework organizes questionnaire predictors into seven predefined semantic domains and learns domain-specific representations through a semantic group-aware neural branch. A sample-specific gating mechanism is further introduced to estimate individualized domain contributions, while residual interaction learning captures complex relationships among psychosocial domains. In addition, a heterogeneous fusion strategy combines neural predictions with CatBoost outputs to exploit complementary modeling capabilities. A multidimensional dataset consisting of 2,894 college students and 51 questionnaire predictors was constructed, and comprehensive experiments, including baseline comparison, ablation analysis, and interpretation studies, were conducted to evaluate the proposed framework.

The main contributions of this study are summarized as follows:

(1) A semantic-aware representation framework is proposed for questionnaire-based mental health risk identification, where heterogeneous psychosocial variables are organized into meaningful domains and incorporated into representation learning rather than being treated as independent predictors.

(2) A personalized gated residual interaction mechanism is developed to adaptively model individual domain contributions and capture cross-domain psychosocial relationships while preserving global information.

(3) A heterogeneous neural–tree fusion strategy is introduced to integrate deep representation learning and tree-based nonlinear modeling, improving the robustness and predictive capability of questionnaire-based risk identification.

2. Related Work

2.1 Statistical and Machine Learning Methods for Questionnaire-Based Mental Health Risk Identification

Questionnaire-based mental health research has traditionally relied on psychometric measurements and statistical modeling. Regression-based methods, including logistic regression and generalized linear models, have been widely applied due to their interpretability and ability to identify associations between psychological outcomes and related factors. Fitzmaurice et al. [1] and Alishiri et al. [2] demonstrated the effectiveness of regression-based frameworks for analyzing health-related factors, while studies in college populations further revealed the relationships between mental health status and multidimensional behavioral and social characteristics [3][4]. Beyond regression analysis, latent variable models and structural approaches have been employed to investigate the multidimensional nature of psychological constructs. Massé et al. [5] emphasized that mental health should be considered a multidimensional construct, and factor analysis and structural equation modeling have further explored relationships among observed variables, latent psychological factors, and environmental influences [6-9]. Cen et al. [10] investigated the relationship between mental health and academic performance using psychological assessments and academic records, while Mao and Chen [11] developed an interpretable mental health risk model based on multivariable modeling and nomogram analysis.

Although these statistical approaches provide valuable interpretability, they generally depend on predefined assumptions and manually specified relationships, limiting their ability to capture nonlinear dependencies and complex interactions among heterogeneous psychosocial factors. To overcome these limitations, machine-learning approaches have been increasingly explored for questionnaire-based mental health risk identification. Srividya et al. [12] evaluated classical classifiers for mental health classification, and subsequent studies compared ensemble models and neural approaches for student mental health prediction [13][14]. Among these methods, tree-based ensemble models have received considerable attention due to their strong capability in modeling nonlinear thresholds and conditional relationships in structured data. Baba and Bunji [15] applied LightGBM to longitudinal student survey data, Rahman and Kohli [16] investigated depression identification among international students, and Gohari et al. [17] employed random forests on a large-scale student dataset to identify important correlates of depressive symptoms. Saleh et al. [18] further demonstrated that XGBoost could reveal predictors beyond conventional linear associations. Machine-learning methods have also been applied to other mental health-related data sources, including electronic health records, digital behavior, and neuroimaging [19-23]. However, questionnaire data remain attractive for university settings due to their accessibility and scalability.

However, existing methods still face challenges in generalizability, robustness, and representation capability. Limited sample sizes may cause overfitting and unstable evaluation [24-27], while most approaches treat questionnaire variables as independent predictors and overlook complex psychosocial relationships.Therefore, more effective approaches are needed to incorporate structured semantic information and complex psychosocial interactions into prediction models.

2.2 Deep Learning for Structured Tabular Mental Health Prediction

Unlike image and signal data, questionnaire-based datasets are typically composed of

heterogeneous numerical, categorical, and behavioral variables without explicit spatial or temporal structures. Conventional multilayer perceptrons (MLPs) usually process these variables as unordered feature vectors, which may limit their ability to capture structured dependencies among related factors. To improve deep learning performance on tabular data, several specialized architectures have introduced additional inductive biases. TabNet employs attentive feature selection to identify informative variables, while TabPFN and Transformer-based approaches explore attention-based or pretrained representations for tabular classification tasks [28-30]. In health-related applications, Jui et al. [31] applied TabNet for physiological stress assessment, and Yuan et al. [32] developed a dual-attention TaBERT framework for academic mental health analysis.

Recent studies have further incorporated attention, masking, gating, and auxiliary supervision mechanisms into tabular representation learning. Lee et al. [33] introduced a gated-attention Transformer with masked self-supervised learning for clinical records, while Habiba and Ludwig [34] combined TabTransformer with domain-informed feature engineering and explainability strategies for medical risk prediction. These studies demonstrate that adaptive representation learning can improve the modeling capability of heterogeneous structured data.

However, most existing deep tabular models mainly operate at the feature level and rarely exploit the semantic organization of questionnaire variables. In mental health prediction, questionnaire items are naturally grouped into meaningful psychosocial domains, and directly combining all variables may ignore domain-specific characteristics. Furthermore, deep learning models often show unstable performance on moderate-sized tabular datasets. Shwartz-Ziv and Armon [35] demonstrated that gradient-boosting methods frequently achieve competitive or superior performance compared with deep tabular models, while subsequent analyses suggested that the effectiveness of neural and tree-based models depends on dataset characteristics [36]. These observations motivate the integration of semantic-aware neural representation learning with complementary tree-based modeling.

2.3 Semantic Interaction Learning and Heterogeneous Fusion for Tabular Data

Feature interaction modeling has been extensively investigated in structured prediction tasks. Factorization machine-based methods capture pairwise interactions among high-dimensional features, while neural interaction models further improve interaction learning through attention mechanisms and adaptive selection strategies. Kang et al. [37] introduced a self-attentional neural factorization model with domain interaction learning, and Chen et al. [38] investigated adaptive interaction selection by considering the unequal contributions of different feature combinations. However, these approaches mainly focus on feature-level interactions and rarely consider higher-level semantic relationships among groups of related variables.

Beyond feature interaction, recent studies have explored adaptive representation enhancement through feature selection, masking mechanisms, and auxiliary objectives. Liu et al. [39] proposed iTabNet with feature selection constraints to improve feature utilization and interpretability. Xu and Xie [40] demonstrated that symbolization and masking mechanisms could facilitate representation learning for categorical variables, while Guo et al. [41] showed that auxiliary supervision could improve intermediate representations. Nevertheless, these approaches mainly adapt individual features and do not explicitly model semantic domains, which are particularly important for questionnaire-based psychosocial data.

Meanwhile, heterogeneous fusion between neural networks and tree-based models has attracted increasing attention because these models capture complementary characteristics in structured data.

Luo and Xu [42] integrated neural architectures with differentiable decision structures, while Li et al. [43] and Zhu et al. [44] explored ensemble frameworks combining tree-based models and neural networks. However, existing fusion strategies generally rely on generic prediction combination and rarely consider semantic structures or individual-specific variations during representation learning. Therefore, a unified framework that jointly incorporates semantic domain modeling, adaptive interaction learning, and heterogeneous fusion is required for reliable questionnaire-based mental health risk identification.

3. Methods

3.1 Overview of the Proposed SGRI-HF Framework

Questionnaire-based mental health risk identification involves heterogeneous psychosocial variables that are naturally organized into meaningful domains. However, existing models usually treat questionnaire items as independent predictors, which may ignore semantic relationships among related factors and individual differences in domain contributions. Moreover, neural networks and tree-based models capture different characteristics of structured data, indicating the potential of heterogeneous learning for improved risk identification.

To address these limitations, we propose the Semantic Group-Aware Gated Residual Interaction and Heterogeneous Fusion (SGRI-HF) framework, which integrates semantic domain representation learning, personalized interaction modeling, and heterogeneous neural–tree fusion. As illustrated in Fig. 1, SGRI-HF consists of two complementary branches: a semantic group-aware neural branch (SGRI-Net) and a tree-based learning branch based on CatBoost.

The SGRI-Net branch first organizes questionnaire predictors into predefined semantic domains and learns domain-specific representations through independent encoders. A sample-specific gating mechanism is then introduced to adaptively estimate domain importance for individual students, followed by residual interaction learning to capture complex relationships among psychosocial domains while preserving global information. In parallel, the CatBoost branch directly models original questionnaire variables and provides complementary nonlinear decision modeling. Finally, the predictions from both branches are integrated through a validation-guided heterogeneous fusion strategy. The overall framework jointly exploits semantic structures, personalized domain contributions, and complementary learning capabilities for college student mental health risk identification.

Throughout this study, the label coding stored in the analytical dataset was retained during model fitting. The native and risk-oriented labels are defined as

$$y_i^N = \begin{cases} 1, & \text{not at risk,} \\ 0, & \text{at risk,} \end{cases} \qquad y_i^R = 1 - y_i^N ., \tag{1}$$

For $b \in \{S, C, \mathrm{HF}\}$, the corresponding class probabilities are defined as

$$p_i^{N,b} = P(y_i^N = 1 | \ \mathbf{x}_i), \qquad p_i^{R,b} = 1 - p_i^{N,b} ., \tag{2}$$

Here, the superscripts N and R denote the native not-at-risk and complemented at-risk label spaces, respectively, whereas S, C, and HF denote SGRI-Net, CatBoost, and heterogeneous fusion. Unless otherwise specified, model fitting and fusion-weight selection are expressed in the native label space, while final risk-oriented metrics are reported with the at-risk class treated as positive.

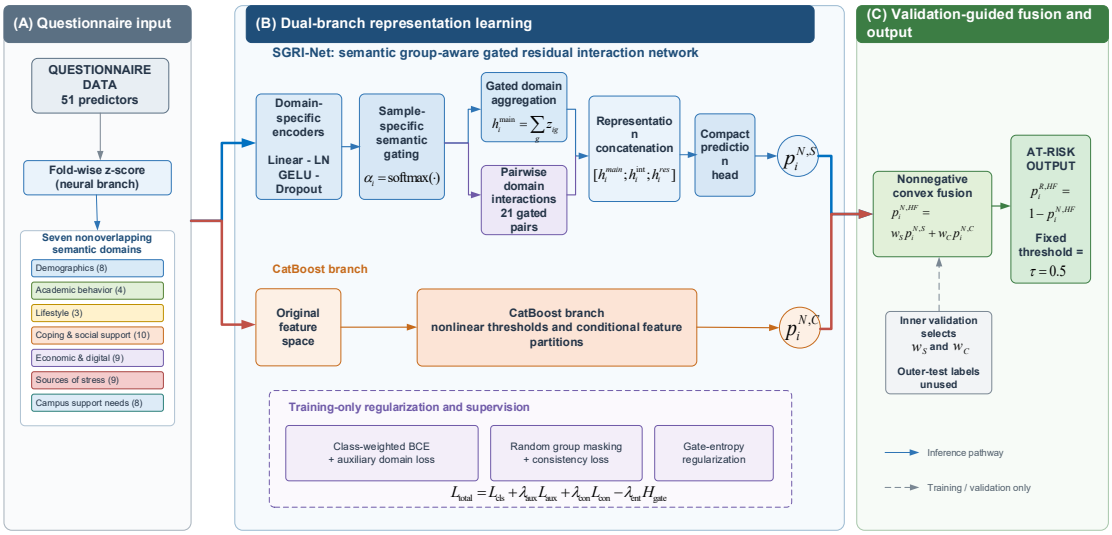


Fig.1 Architecture of SGRI-HF.

3.2 Semantic Group-Aware Representation Learning

Questionnaire variables are not independent observations but are naturally organized into meaningful psychosocial domains. To incorporate this structural prior into representation learning, SGRI-HF introduces a semantic group-aware encoding strategy that explicitly models domain-level information before learning cross-domain relationships.

Let $x_i \in \mathbb{R}^d$ denote the questionnaire feature vector of participant i, where d represents the number of input predictors. The predictors are assigned a priori into $G = 7$ non-overlapping semantic domains, including demographic characteristics, academic behavior, lifestyle, coping and social support, economic and digital behavior, sources of stress, and campus support needs. Let I_g denote the index set corresponding to domain g. The domain-specific input vector is therefore represented as:

$$x_i^{(g)} = x_i[\mathrm{I}_g], g = 1, \ldots, G, \tag{3}$$

where the sets $\{\mathrm{I}_g\}_{g=1}^{G}$ form a partition of the complete predictor space.

Unlike conventional approaches that directly concatenate all questionnaire variables into an unordered feature vector, the proposed strategy introduces semantic partitioning before representation learning. This design enables each domain to develop specialized feature transformations according to its measurement characteristics, while avoiding excessive mixing of heterogeneous psychosocial factors at the initial learning stage.

Each semantic domain is independently projected into a shared latent space through a domain-specific encoder $f_g(\cdot)$. Specifically, each encoder consists of a linear transformation, layer normalization (LN), Gaussian error linear unit (GELU) activation, and dropout operation. The latent representation of domain g for participant i is formulated as:

$$h_{\mathrm{ig}} = \mathrm{Dropout}[\mathrm{GELU}\{\mathrm{LN}(W_g \tilde{x}_i^{(g)} + b_g)\}] \in \mathbb{R}^h, g = 1, \ldots, G., \tag{4}$$

where $\tilde{x}_i^{(g)}$ denotes the normalized domain input, W_g and b_g are learnable parameters of the domain encoder, and h represents the dimension of the latent domain representation.

Independent encoder parameters allow different psychosocial domains to learn domain-specific representations, while the shared latent dimension enables subsequent comparison and interaction modeling across domains. The domain representations are concatenated as:

$$h_{\mathrm{ig}} = \mathrm{Dropout}[\mathrm{GELU}\{\mathrm{LN}(W_g \tilde{x}_i^{(g)} + b_g)\}] \in \mathbb{R}^h, g = 1, \ldots, G., \tag{5}$$

where H_i provides the contextual representation for subsequent personalized domain weighting.

The semantic group-aware representation has two advantages. First, it incorporates prior questionnaire organization into neural representation learning, allowing the model to capture domain-level characteristics rather than isolated feature contributions. Second, it produces interpretable intermediate domain representations, which provide the basis for personalized gating and cross-domain interaction modeling described in the following section.

3.3 Personalized Gated Residual Domain Interaction Learning

Although semantic grouping enables domain-aware representation learning, the contribution of different psychosocial domains may vary among individuals. For example, mental health risks may be dominated by academic stress for some students, whereas social support or lifestyle factors may play a more important role for others. Therefore, SGRI-HF introduces a personalized gating mechanism to adaptively estimate the contribution of each semantic domain based on individual questionnaire profiles.

3.3.1 Sample-specific Semantic Domain Gating

Given the concatenated domain representation H_i obtained from Section 3.2, a gating network is constructed to generate sample-specific domain importance weights. Specifically, a two-layer multilayer perceptron followed by a softmax normalization is used:

$$\alpha_i = \mathrm{softmax}[W_2 \mathrm{GELU}(W_1 H_i + b_1) + b_2] \tag{6}$$

where $\alpha_i = [\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iG}]$ represents the contribution distribution among G semantic domains for participant i, satisfying:

$$\sum_g \alpha_{\mathrm{ig}} = 1, \alpha_{\mathrm{ig}} \geq 0. \tag{7}$$

The gated domain representation is calculated as:

$$z_{\mathrm{ig}} = \alpha_{\mathrm{ig}} h_{\mathrm{ig}} \tag{8}$$

where α_{ig} dynamically adjusts the contribution of domain g according to the complete questionnaire profile of participant i. Unlike conventional feature importance methods that provide population-level interpretations, the proposed gating mechanism enables individualized domain weighting by allowing different participants to obtain distinct domain contribution patterns. The aggregated domain representation is formulated as:

$$h_i^{main} = \sum_g z_{\mathrm{ig}} \tag{9}$$

The learned gate values are retained for subsequent interpretation analysis and are regarded as model-derived contribution weights rather than causal effects.

3.3.2 Residual Cross-domain Interaction Learning

Mental health risk is usually associated with the combined influence of multiple psychosocial conditions rather than isolated domain effects. To explicitly model such dependencies, SGRI-HF introduces a residual cross-domain interaction module based on gated semantic representations.

For each pair of semantic domains, element-wise multiplication is employed to capture multiplicative interactions:

$$u_i = \frac{1}{\binom{G}{2}} \sum_{1 \le g < r \le G} \left(z_{ig} \odot z_{ir} \right), \tag{10}$$

where $\odot$ denotes the Hadamard product. With $G = 7$, this formulation considers $\binom{7}{2} = 21$ pairwise domain interactions while avoiding the excessive parameter growth caused by unrestricted high-order interaction tensors.

The interaction representation is further transformed through a nonlinear projection:

$$h_i^{\text{int}} = \text{Dropout}\left[\text{GELU}\left(W_{\text{int}} u_i + b_{\text{int}} \right) \right] \tag{11}$$

Compared with implicit interactions learned only through fully connected layers, this explicit interaction mechanism introduces domain-level multiplicative dependencies into the representation learning process, improving the ability to capture complex psychosocial relationships.

To avoid losing complementary information caused by predefined semantic partitioning, a global residual pathway is introduced:

$$h_i^{res} = \text{GELU}\left\{ \text{LN}\left(W_{\text{res}} \tilde{x}_i + b_{\text{res}} \right) \right\} \tag{12}$$

where the original feature representation provides additional information that may not be fully preserved by domain-specific encoding and interaction aggregation.

Finally, the neural representation is obtained by concatenating the three components:

$$q_i = \left[h_i^{main} \,\|\, h_i^{\text{int}} \,\|\, h_i^{res} \right] \in \mathbb{R}^{3h} \tag{13}$$

A two-layer prediction head maps q_i to the neural logit s_i^S and probability $p_i^{N,S} = \sigma\left(s_i^S \right)$, where $\sigma(\cdot)$ is the sigmoid function.

3.4 Heterogeneous Neural–Tree Fusion Strategy

Although SGRI-Net is capable of learning semantic domain representations and personalized psychosocial interactions, neural models may smooth discrete decision boundaries in structured questionnaire data. In contrast, tree-based ensemble models are effective at capturing nonlinear thresholds and conditional relationships among heterogeneous variables. Therefore, SGRI-HF introduces a heterogeneous fusion strategy that combines the complementary characteristics of neural representation learning and tree-based decision modeling.

Specifically, a CatBoost branch is constructed to directly process the original questionnaire

variables. Unlike SGRI-Net, which learns distributed latent representations from semantic domains, CatBoost operates in the original feature space and exploits gradient-boosted decision trees to model nonlinear feature partitions. These two branches therefore provide complementary predictive information from different representation perspectives.

Using the notation introduced in Section 3.1, the fused native not-at-risk probability is obtained through nonnegative convex fusion:

$$p_i^{N,\mathrm{HF}} = w_S p_i^{N,S} + w_C p_i^{N,C}, w_S \geq 0, w_C \geq 0, w_S + w_C = 1. \quad (14)$$

The convex constraint ensures that the fused prediction remains within the valid probability range while allowing adaptive contribution from both branches. The fusion weights are determined using validation data only, preventing information leakage from the held-out test samples.

In implementation, the two endpoint solutions were excluded, and candidate weights were restricted to $w_S \in \mathrm{W} = \{0.1, 0.2, \ldots, 0.9\}$, with $w_C = 1 - w_S$, thereby ensuring a nonzero contribution from both branches. For each $w_S \in \mathrm{W}$, the fused native not-at-risk probability was computed as $p_{i,\mathrm{val}}^{N,\mathrm{HF}} = w_S p_{i,\mathrm{val}}^{N,S} + w_C p_{i,\mathrm{val}}^{N,C}$. The candidate pair maximizing $J = 0.35\mathrm{AUC}^N + 0.25\mathrm{AP}^N + 0.20\mathrm{BAcc} + 0.20(\mathrm{MCC}+1)/2$ was retained and applied unchanged to the corresponding outer-test fold. Here, AUC^N and AP^N denote ROC-AUC and average precision in the native not-at-risk label space, whereas BAcc and MCC denote balanced accuracy and the Matthews correlation coefficient evaluated at $\tau = 0.5$. Fusion-weight selection was therefore conducted in the native label space used for model fitting, whereas final risk-oriented evaluation was conducted in the complemented at-risk label space.

$$p_i^{R,\mathrm{HF}} = 1 - p_i^{N,\mathrm{HF}}, \qquad \hat{y}_i^R = \mathrm{I}[p_i^{R,\mathrm{HF}} \geq \tau], \qquad \tau = 0.5. \quad (15)$$

Here, $p_i^{R,\mathrm{HF}}$ denotes the final at-risk probability obtained by complementing the fused native not-at-risk probability. The binary prediction $\hat{y}_i^R$ equals 1 when the estimated at-risk probability reaches the prespecified threshold $\tau = 0.5$, and 0 otherwise, where $\mathrm{I}[\cdot]$ denotes the indicator function. The threshold was fixed before outer-test evaluation and was not optimized using outer-test labels.

Through this heterogeneous fusion strategy, SGRI-HF integrates two complementary learning paradigms: semantic-aware neural representation learning for capturing latent psychosocial relationships and tree-based modeling for exploiting nonlinear structures in the original questionnaire space. This design improves the robustness of risk identification by avoiding dependence on a single modeling assumption.

3.5 Optimization Objective

The proposed SGRI-Net is optimized by jointly considering prediction accuracy, domain-specific supervision, representation consistency, and personalized domain weighting. Since the native questionnaire labels are used for neural branch optimization, a class-weighted binary cross-entropy loss is adopted to alleviate the influence of class imbalance.

Let $\rho = n_R / n_N$ denote the class-weighting coefficient applied to the native-positive not-at-risk term within the current inner-training subset, where n_R and n_N are the numbers of at-risk (native-negative) and not-at-risk (native-positive) participants, respectively.

$$L_{\text{cls}} = -\frac{1}{B}\sum_{i=1}^{B}\left[\rho y_i^N \log p_i^{N,S} + (1-y_i^N)\log(1-p_i^{N,S})\right] \tag{16}$$

where B denotes the mini-batch size.

To encourage each semantic domain encoder to preserve risk-related information, an auxiliary classifier is attached to each domain representation h_{ig}. The corresponding auxiliary losses are calculated using the same prediction target and averaged across all semantic domains:

$$L_{aux} = \frac{1}{G}\sum_{g=1}^{G} L_{BCE}^{g} \tag{17}$$

where L_{BCE}^{g} represents the binary cross-entropy loss generated from the auxiliary classifier of domain g. This auxiliary supervision prevents the representation learning process from being dominated by only a small subset of semantic domains.

Furthermore, a group-masking consistency regularization strategy is introduced to improve representation robustness under incomplete semantic information. During training, semantic domains are randomly masked, and the prediction from the masked representation is encouraged to remain consistent with the complete-input prediction:

$$L_{\text{con}} = B^{-1}\sum_{i=1}^{B}\left[\tilde{p}_i^{N,S} - \text{stopgrad}\left(p_i^{N,S}\right)\right]^2 \tag{16}$$

where $\tilde{p}_i^{N,S}$ denotes the prediction obtained from the masked forward process, and $stopgrad(\cdot)$ prevents the complete-input prediction from being altered by the consistency constraint.

To maintain the flexibility of personalized domain weighting, a weak entropy regularization is introduced on the gating distribution:

$$H_{\text{gate}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{g=1}^{G}\alpha_{ig}\log(\alpha_{ig}+\varepsilon) \tag{18}$$

where ε is a small constant for numerical stability. Maximizing gate entropy prevents premature concentration on a single semantic domain while preserving the ability of the gating mechanism to generate individualized weighting patterns.

The overall optimization objective of SGRI-Net is formulated as:

$$L_{\text{total}} = L_{\text{cls}} + \lambda_{\text{aux}}L_{\text{aux}} + \lambda_{\text{con}}L_{\text{con}} - \lambda_{\text{ent}}H_{\text{gate}} \tag{19}$$

where λ_{aux}, λ_{con}, λ_{ent} control the contributions of auxiliary supervision, consistency regularization, and entropy regularization, respectively.

The proposed objective jointly optimizes predictive discrimination, semantic domain representation quality, robustness against incomplete domain information, and individualized domain contribution learning, thereby enabling SGRI-Net to effectively model complex psychosocial risk patterns.

Algorithm 1 gives the steps of the proposed method.

Algorithm 1. Fold-wise training and inference for SGRI-HF

Input: Questionnaire dataset $\mathrm{D} = \{(\mathbf{x}_i, y_i^N)\}_{i=1}^{n}$, where $y_i^N = 1$ denotes not at risk and $y_i^N = 0$ denotes at risk; seven semantic groups $I = \{\mathrm{I}_g\}_{g=1}^{7}$; number of outer folds $K = 5$; SGRI-Net

hyperparameters H_S; CatBoost hyperparameters H_C; loss coefficients λ_{aux}, λ_{con}, λ_{ent}; fusion grid $\mathrm{W} = \{0.1, 0.2, \ldots, 0.9\}$; fixed risk threshold $\tau = 0.5$.

Output: Out-of-fold (OOF) SGRI-Net probabilities $\mathbf{p}^{N,S}$, CatBoost probabilities $\mathbf{p}^{N,C}$, fused at-risk probabilities $\mathbf{p}^{R,\mathrm{HF}}$, predicted at-risk labels $\mathbf{y}^{R}$, fold-specific fusion weights, and semantic gate weights $\mathbf{A} = \{\alpha_{ig}\}_{i=1,g=1}^{n,G}$.

BEGIN

Construct five stratified outer folds using seed 2026 and initialize empty OOF containers.

For $k = 1, \ldots, K$ **do**

Assign fold k to the outer-test set and the remaining folds to the outer-development set.

Stratify the outer-development set into inner training (80%) and inner validation (20%) using seed 2026 + k.

Fit the neural standardizer on inner training only; apply it unchanged to inner training, validation, and outer test.

Initialize the fold-specific SGRI-Net $f_S^{(k)}$.

For epoch t = 1, ..., 160 **do**

Compute seven domain representations, sample-specific gates, gated aggregation, 21 pairwise interactions, the global residual representation, and $p_i^{N,S}$.

Compute $L_{\text{total}} = L_{\text{cls}} + \lambda_{\text{aux}} L_{\text{aux}} + \lambda_{\text{con}} L_{\text{con}} - \lambda_{\text{ent}} H_{\text{gate}}$

Update the SGRI-Net parameters using AdamW and gradient clipping.

Retain the lowest-loss parameter state and stop if validation loss does not improve by at least 10^{-4} for 18 epochs.

End For

Restore the best-performing SGRI-Net parameter state.

Fit CatBoost $f_C^{(k)}$ on the original integer-coded inner-training variables without neural standardization.

Generate SGRI-Net and CatBoost probabilities on the inner-validation subset.

For each $w_S \in \mathrm{W}$, $w_C = 1 - w_S$ set and compute

$$p_{i,\text{val}}^{N,\mathrm{HF}} = w_S\, p_{i,\text{val}}^{N,S} + w_C\, p_{i,\text{val}}^{N,C}$$

followed by $J(w_S, w_C) = 0.35\,\mathrm{AUC}^N + 0.25\,\mathrm{AP}^N + 0.20\,\mathrm{BAcc} + 0.20\left(\dfrac{\mathrm{MCC}+1}{2}\right)$

End For

$$w_S^{(k)} = \arg\max_{w_S \in \mathrm{W}} J(w_S, 1 - w_S), \qquad w_C^{(k)} = 1 - w_S^{(k)}.$$

Generate SGRI-Net and CatBoost probabilities for the held-out outer-test subset exactly once.

Compute $p_{i,\text{test}}^{N,\mathrm{HF}} = w_S^{(k)}\, p_{i,\text{test}}^{N,S} + w_C^{(k)}\, p_{i,\text{test}}^{N,C}$.

Predict $\hat{y}_i^R = \mathrm{I}[p_i^{R,\mathrm{HF}} \geq \tau]$

Store branch probabilities, fused risk probabilities, predicted labels, selected weights, and gate weights in the OOF containers.

End For

Aggregate the five held-out outputs so that each participant contributes exactly one OOF prediction.

Return the OOF branch probabilities, fused at-risk probabilities and labels, fold-specific fusion weights, and semantic gate weights.

END

4. Experiments

4.1 Study Cohort and Questionnaire Dataset

The analytical cohort consisted of undergraduate students recruited from a university in Henan Province, China, between October and December 2024. Cluster sampling was conducted across

different grades and academic units. Among approximately 2,950 distributed questionnaires, 2,894 valid responses were retained, resulting in an effective response rate of approximately 98.10%. Participation was voluntary and anonymous, and informed consent was obtained before questionnaire completion. The study protocol was approved by the institutional ethics committee and conducted according to relevant ethical requirements.

The final cohort included 1,677 male students (57.95%) and 1,217 female students (42.05%). First-, second-, third-, and fourth-year students accounted for 47.62%, 25.78%, 22.49%, and 4.11% of the cohort, respectively. According to the questionnaire-based mental health evaluation, 1,394 students (48.17%) were classified as at risk and 1,500 students (51.83%) were classified as not at risk.

The questionnaire consisted of 24 thematic questions covering demographic characteristics, academic behavior, lifestyle, coping and social support, economic and digital behavior, sources of stress, and campus support needs. After feature encoding, multiple-choice responses were expanded into binary indicators, resulting in 51 predictors, including 34 binary variables, 16 five-level ordinal variables, and one four-level ordinal variable (Table 1). Based on questionnaire semantics, these predictors were predefined into seven mutually exclusive domains, which were consistently used for semantic representation learning in SGRI-HF.

Table 1 Semantic organization of the questionnaire predictors.

Semantic domain	No. of predictors	Primary data type	Content represented
Demographic characteristics	8	Binary and ordinal	Sex, grade, place of origin, and student-leadership indicators
Academic behavior	4	Five-level ordinal	Attendance, assignment completion, study planning, and failed/retaken courses
Lifestyle	3	Five-level ordinal	Sleep duration, sleep quality, and regular exercise
Coping and social support	10	Binary and ordinal	Coping strategies, peer/family support, family communication, and campus participation
Economic and digital behavior	9	Binary and ordinal	Disposable funds, economic-pressure responses, screen use, late-night use, and phone dependence
Sources of stress	9	Binary	Academic, interpersonal, financial, family, appearance, and future-related stressors
Campus support needs	8	Binary	Counseling, career, financial, academic, health, dormitory, and related support needs

4.2 Competing Methods

To evaluate the effectiveness of SGRI-HF, representative machine-learning and deep-learning methods were selected for comparison. The comparison methods covered linear models, probabilistic classifiers, kernel-based models, tree-based ensembles, and neural tabular models.

The compared baselines included conventional machine-learning models and deep tabular learning models. Conventional models consisted of logistic regression (LR), Gaussian naive Bayes (GNB), radial basis function support vector machine (RBF-SVM), random forest (RF), extremely randomized trees (Extra Trees), histogram-based gradient boosting (HistGradientBoosting), and CatBoost. Deep tabular models included multilayer perceptron (MLP), TabTransformer, and a compact TabNet-style model.

In addition, SGRI-Net was evaluated as the standalone semantic neural branch before heterogeneous fusion, while SGRI-HF denotes the complete framework integrating SGRI-Net and CatBoost. All models were evaluated under the same data partitions and experimental protocol. The main configurations of all comparison methods are summarized in Table 2.

Table 2 Comparator models and principal configurations.

Model	Principal configuration
Logistic regression	L2-regularized logistic regression; $C = 1.0$; balanced class weights; maximum 3,000 iterations; fold-wise standardization

Gaussian naive Bayes	Variance-smoothing coefficient 10^{-8}; fold-wise standardization
RBF-SVM	Radial-basis-function kernel; $C = 1.0$; gamma set to scale; probability estimation enabled; balanced class weights; fold-wise standardization
Random forest	500 trees; square-root feature sampling; minimum leaf size 4; balanced-subsample class weighting
Extra Trees	500 trees; square-root feature sampling; minimum leaf size 3; balanced class weighting
HistGradientBoosting	Learning rate 0.05; maximum 300 iterations; maximum 15 leaf nodes; L2 regularization 1.0
CatBoost(standalone comparator)	450 boosting iterations; tree depth 5; learning rate 0.03; L2 leaf regularization 5.0; binary log-loss
MLP	Hidden dimensions 64 and 32; LayerNorm-GELU; dropout 0.25/0.20; standardized inputs
TabTransformer	16-dimensional feature tokens; two encoder layers; four heads; feed-forward dimension 64; dropout 0.20
Compact TabNet-style	Four attentive steps; decision/attention dimensions 16/16; sparsemax masks; relaxation factor 1.5
SGRI-Net	Complete semantic neural branch evaluated before heterogeneous fusion; architecture and objective defined in Section 3

4.3 Experimental Setup

All experiments were conducted using leakage-controlled stratified five-fold cross-validation with a fixed random seed of 2026. The same outer-fold partitions were applied to SGRI-HF and all comparison models. In each fold, the data were divided into development and independent outer-test subsets, while the development subset was further split into inner-training and inner-validation subsets for model fitting, early stopping, and fusion-weight selection. The outer-test subset was only used after all fold-specific decisions were completed.

All preprocessing procedures were performed within the corresponding training folds to prevent information leakage. Models requiring standardized inputs used normalization parameters estimated from the inner-training subset only, whereas tree-based models directly used the original encoded questionnaire variables. No resampling strategies or post hoc sample removal were applied.

SGRI-Net was optimized using AdamW with a learning rate of 8×10^{-4}, weight decay of 2×10^{-4}, batch size of 64, maximum 160 epochs, hidden dimension of 24, and dropout rate of 0.20. Early stopping was performed based on inner-validation loss. The standalone CatBoost comparator used the configuration reported in Table 2. Within SGRI-HF, the CatBoost branch used 500 boosting iterations, a tree depth of 5, a learning rate of 0.025, L2 leaf regularization of 6.0, Bayesian bootstrap, and binary log-loss. For outer fold k, the fold-specific random seed was set to $2026+k$. The CatBoost branch was fitted exclusively on the inner-training subset, whereas the fusion weights were selected exclusively on the inner-validation subset; neither procedure used the corresponding outer-test data.

Ablation experiments were conducted to evaluate the contribution of individual SGRI-Net components. All variants used identical data partitions, preprocessing procedures, optimization settings, and decision thresholds as the complete model. The ablation configurations are summarized in Table 3.

Table 3 SGRI-Net ablation variants.

Variant	Modification relative to the complete SGRI-Net
Complete SGRI-Net	All semantic encoders, sample-specific gating, pairwise interaction learning, global residual pathway, auxiliary supervision, and consistency regularization retained
Without auxiliary supervision	Group-specific auxiliary classification loss removed
Without consistency regularization	Random semantic-group masking and prediction-consistency loss disabled
Without semantic gating	Learned participant-specific gates replaced by uniform weights of $\frac{1}{7}$
Without pairwise interaction	Explicit inter-domain interaction representation replaced with a zero vector of equal dimension
Without global residual pathway	Global residual representation replaced with a zero vector of equal dimension

4.4 Performance Metrics and Statistical Analysis

The performance of all models was evaluated using the outer-test predictions generated from the leakage-controlled five-fold cross-validation procedure. The at-risk class was treated as the positive class for all risk-oriented analyses. Threshold-dependent metrics included accuracy, balanced accuracy, macro-F1, Matthews correlation coefficient (MCC), sensitivity, specificity, precision, and at-risk F1-score. Threshold-independent discrimination was assessed using ROC-AUC and at-risk average precision (AP^R), while probability calibration was evaluated using the Brier score and binary log loss.

For the at-risk class, sensitivity, specificity, precision, and F1-score were calculated as:

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},$$

$$F_1^R = \frac{2 \cdot \mathrm{Precision}^R \cdot \mathrm{Sensitivity}^R}{\mathrm{Precision}^R + \mathrm{Sensitivity}^R}.$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. MCC was used as an additional balanced evaluation metric:

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}.$$

Metrics were calculated independently across the five outer-test folds and reported as mean $\pm$ standard deviation. Pooled out-of-fold predictions were used for ROC curves, precision-recall curves, and calibration analysis. A fixed threshold of $\tau = 0.5$ was applied to all models without test-set optimization.

Paired statistical comparisons between SGRI-HF and competing models were performed using participant-level bootstrap resampling based on pooled out-of-fold predictions. Confidence intervals of metric differences were estimated using percentile bootstrap intervals, and multiple comparisons were corrected using the Holm procedure. Statistical significance was determined using adjusted two-sided p-values below 0.05.

5. Results and Discussion

5.1 Cohort Characteristics and Multidomain Risk Patterns

The final analytical cohort consisted of 2,894 undergraduate students, including 1,394 students classified as at risk and 1,500 students classified as not at risk. The relatively balanced distribution between the two groups provided a suitable basis for comparative analysis and model evaluation. Although demographic characteristics provide fundamental background information, mental health risk is unlikely to be determined by a single population attribute. Therefore, we further investigated whether risk-related patterns were distributed across multiple behavioral, psychosocial, and environmental dimensions.

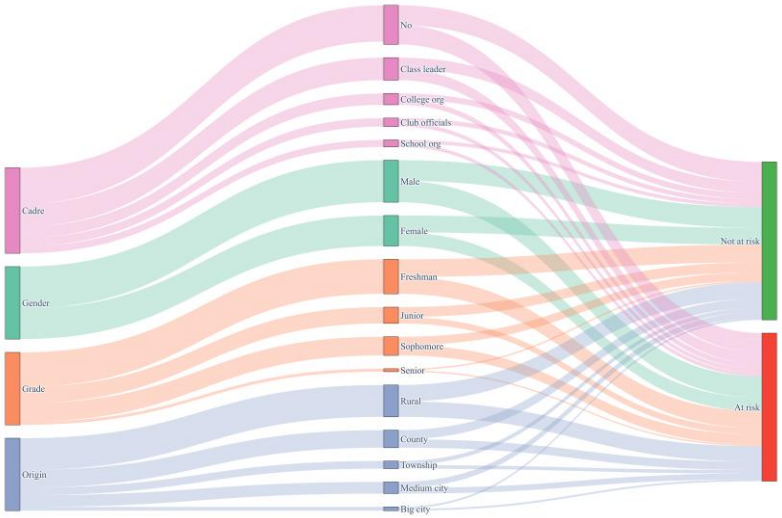


Fig.2 Distribution of demographic characteristics in the analytical cohort.

Figure 2 illustrates the distribution of demographic characteristics between the at-risk and not-at-risk groups. The observed differences across demographic variables indicate that risk status was not completely independent of student background characteristics. However, these variables alone provided limited information for distinguishing risk status, suggesting that additional behavioral and psychosocial factors are required for comprehensive risk representation.

Beyond demographic characteristics, academic behaviors showed noticeable differences between the two groups. Students in the at-risk group exhibited less favorable academic adaptation patterns, particularly in study planning, attendance-related behaviors, and assignment completion. These findings indicate that academic functioning may reflect important behavioral signals associated with mental health risk.

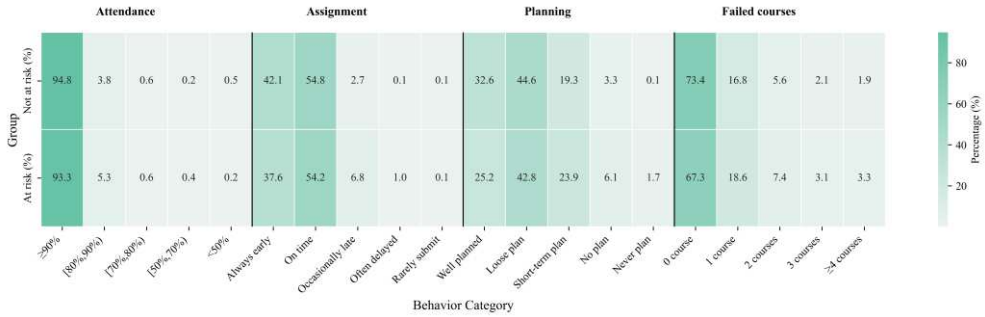
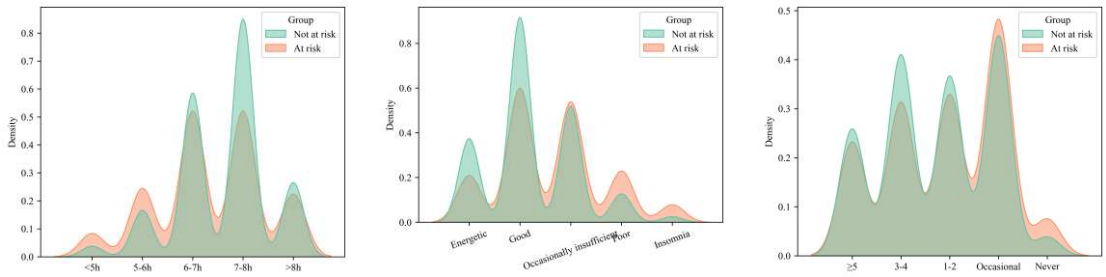


Fig.3 Distribution of academic behavior characteristics across at-risk and not-at-risk groups.

Figure 3 presents the distribution of academic behavior characteristics across the two groups. The heterogeneous patterns observed in academic-related variables suggest that risk identification requires modeling multiple related behaviors rather than relying on individual academic indicators.

Lifestyle-related factors also demonstrated clear differences between groups. Sleep duration, sleep quality, and daily activity patterns showed consistent associations with risk status, indicating that lifestyle behaviors represent another important dimension of student mental health risk.



(a) Sleep duration (b) Sleep quality (c) Physical activity

Fig.4 Distribution of lifestyle characteristics between at-risk and not-at-risk groups.

Figure 4 summarizes the differences in lifestyle characteristics between at-risk and not-at-risk students. The results highlight that daily behavioral regulation, particularly sleep-related factors, provides meaningful information for identifying students with elevated mental health risk.

In addition to academic and lifestyle factors, economic pressure and digital behaviors exhibited distinct patterns between groups. Variables related to mobile-phone dependence, screen exposure, and financial-related stress responses showed noticeable differences, suggesting that contemporary digital environments and economic conditions may contribute additional information for risk characterization.

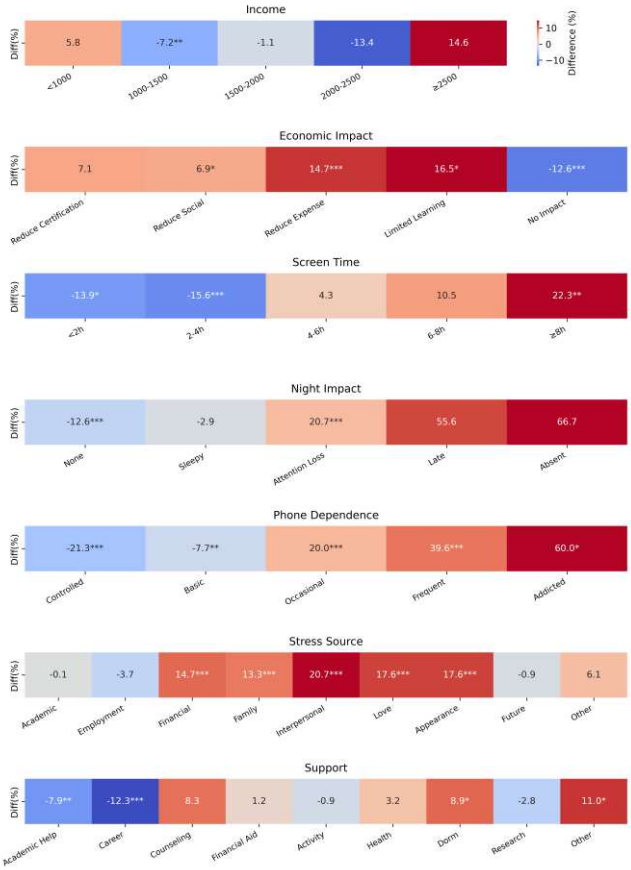


Fig.5 Distribution differences of economic and digital-behavior-related factors between at-risk and not-at-risk groups.

Figure 5 shows the distribution of economic and digital behavior characteristics across the two groups. These findings further support the necessity of integrating heterogeneous psychosocial

information from multiple domains.

To quantify the relationship between individual predictors and mental health risk, chi-square tests followed by Benjamini–Hochberg correction were conducted. Among the 51 predictors, 32 variables remained statistically significant after multiple-testing correction ($q < 0.05$). The strongest associations were observed for peer support availability (Cramér's V = 0.206), self-rated sleep quality (0.200), perceived family understanding (0.189), family communication frequency (0.173), mobile-phone dependence (0.167), campus participation (0.162), and sleep duration (0.161).

Table 4 Fifteen strongest predictor-outcome associations ranked by Cramér's V.

Predictor	Chi-square	df	BH-adjusted q	Cramér's V
Peer support availability	123.09	4	5.97e-24	0.206
Self-rated sleep quality	115.90	4	1.02e-22	0.200
Perceived family understanding	103.04	4	3.77e-20	0.189
Frequency of family communication	86.42	4	9.67e-17	0.173
Mobile-phone dependence	81.09	4	1.05e-15	0.167
Campus activity participation	75.70	4	1.21e-14	0.162
Average daily sleep duration	74.92	4	1.51e-14	0.161
Appearance/body-image anxiety	55.72	1	5.32e-13	0.139
Study-planning habits	54.83	4	1.80e-10	0.138
Interpersonal conflict	51.44	1	4.18e-12	0.133
Daily entertainment screen time	43.99	4	2.53e-08	0.123
Assignment completion	43.42	4	3.08e-08	0.122
Family-expectation pressure	41.55	1	5.32e-10	0.120
Reduced food expenditure due to financial pressure	39.29	1	1.55e-09	0.117
Grade	37.16	3	1.36e-07	0.113

Table 4 lists the fifteen strongest predictor–outcome associations ranked by Cramér's V. Although several variables showed significant associations with mental health risk, their individual effect sizes remained moderate. Importantly, the significant predictors were distributed across multiple semantic domains rather than concentrated within a single category. This distributed association pattern suggests that mental health risk may emerge from coordinated changes across psychosocial domains, providing the rationale for domain-level representation learning and cross-domain interaction modeling in SGRI-HF.



Fig.6 Three-layer network visualization of psychosocial factors and mental health risk status.

The multidomain association network in Figure 6 further demonstrates the relationships among psychosocial factors and risk status. Factors related to social support, lifestyle, academic behavior, stress sources, and digital behavior jointly contributed to the observed risk patterns.

5.2 Comparative Performance and Statistical Evidence

Under the leakage-controlled stratified five-fold cross-validation protocol, SGRI-HF achieved the most favorable overall performance profile among the evaluated models (Table 5). The proposed framework achieved an accuracy of 0.6465 ± 0.0058, macro-F1 of 0.6436 ± 0.0057, and ROC-AUC of 0.6911 ± 0.0242, together with balanced accuracy of 0.6441 ± 0.0056 and MCC of 0.2911 ± 0.0115. For the at-risk class, SGRI-HF achieved a precision of 0.6493 ± 0.0102, sensitivity of 0.5796 ± 0.0194, and F1-score of 0.6122 ± 0.0090. The relatively small standard deviations of most threshold-dependent metrics indicate stable classification behavior across outer folds, although the larger variation in ROC-AUC suggests that ranking performance was more sensitive to data partitioning.

Table 5 Predictive performance under stratified five-fold cross-validation.

Model	Accuracy	Balanced accuracy	MCC	At-risk F1	Macro-F1	ROC-AUC
SGRI-HF	**0.6465 ± 0.0058**	**0.6441 ± 0.0056**	**0.2911 ± 0.0115**	0.6122 ± 0.0090	**0.6436 ± 0.0057**	**0.6911 ± 0.0242**
Gaussian naive Bayes	0.6382 ± 0.0275	0.6359 ± 0.0276	0.2744 ± 0.0557	0.6039 ± 0.0332	0.6353 ± 0.0280	0.6755 ± 0.0324
Random forest	0.6372 ± 0.0238	0.6354 ± 0.0240	0.2724 ± 0.0480	0.6094 ± 0.0291	0.6352 ± 0.0241	0.6872 ± 0.0298
CatBoost	0.6375 ± 0.0122	0.6352 ± 0.0121	0.2729 ± 0.0249	0.6024 ± 0.0135	0.6346 ± 0.0120	0.6846 ± 0.0193
RBF-SVM	0.6358 ± 0.0232	0.6347 ± 0.0235	0.2700 ± 0.0469	**0.6151 ± 0.0282**	0.6347 ± 0.0235	0.6714 ± 0.0332

Model	Accuracy	Balanced accuracy	MCC	At-risk F1	Macro-F1	ROC-AUC
SGRI-Net	0.6296 ± 0.0300	0.6289 ± 0.0297	0.2582 ± 0.0600	0.6133 ± 0.0293	0.6288 ± 0.0298	0.6756 ± 0.0389
Extra Trees	0.6275 ± 0.0155	0.6258 ± 0.0155	0.2529 ± 0.0312	0.5995 ± 0.0168	0.6256 ± 0.0154	0.6745 ± 0.0261
HistGradientBoosting	0.6254 ± 0.0139	0.6237 ± 0.0136	0.2487 ± 0.0280	0.5974 ± 0.0119	0.6236 ± 0.0135	0.6694 ± 0.0178
Logistic regression	0.6234 ± 0.0249	0.6230 ± 0.0250	0.2461 ± 0.0501	0.6109 ± 0.0274	0.6229 ± 0.0250	0.6676 ± 0.0291
MLP	0.6237 ± 0.0285	0.6226 ± 0.0281	0.2459 ± 0.0570	0.6026 ± 0.0259	0.6225 ± 0.0281	0.6690 ± 0.0336
TabTransformer	0.6189 ± 0.0298	0.6171 ± 0.0315	0.2369 ± 0.0616	0.5858 ± 0.0569	0.6152 ± 0.0324	0.6615 ± 0.0358
Compact TabNet-style	0.5802 ± 0.0285	0.5797 ± 0.0271	0.1612 ± 0.0563	0.5651 ± 0.0263	0.5781 ± 0.0261	0.6067 ± 0.0351

Compared with individual baselines, SGRI-HF consistently improved the overall balance between discrimination and classification performance. Relative to CatBoost, the strongest tree-based comparator, SGRI-HF improved balanced accuracy, macro-F1, and ROC-AUC by approximately 0.9, 0.9, and 0.7 percentage points, respectively. Random forest achieved a comparable ROC-AUC (0.6872), while RBF-SVM and SGRI-Net obtained slightly higher at-risk F1 values than SGRI-HF. Therefore, the advantage of SGRI-HF should be interpreted as improved overall balance across multiple metrics rather than universal superiority on every individual endpoint.

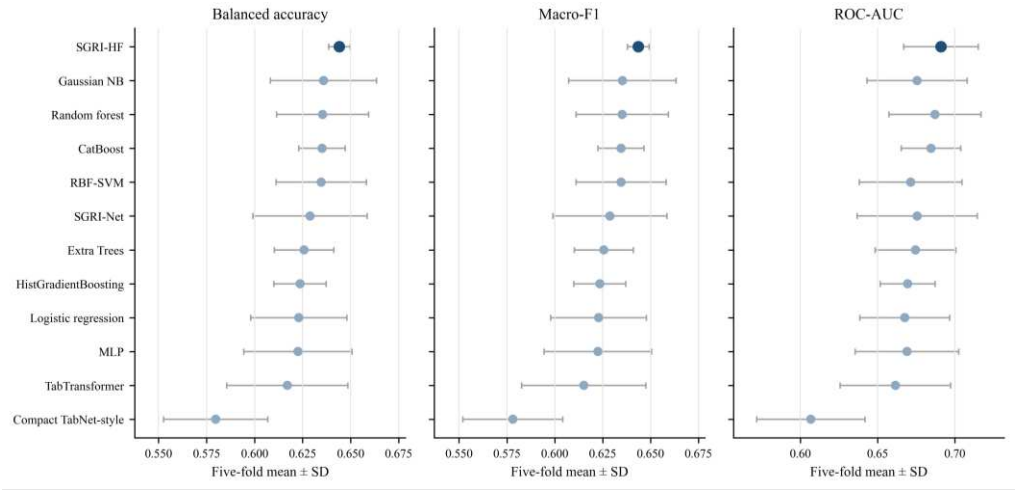


Fig.7 Fold-level balanced accuracy, macro-F1, and ROC-AUC across the unified model comparison.

Tree-based models remained competitive on this medium-scale questionnaire dataset, whereas generic neural tabular baselines did not surpass the strongest tree ensembles. The improvement of SGRI-HF over CatBoost suggests that semantic representation learning and tree-based decision modeling capture complementary information rather than competing alternatives.

To further assess the robustness of performance differences, paired participant-level bootstrap comparisons were conducted using pooled out-of-fold predictions. Compared with CatBoost, SGRI-HF showed positive trends in balanced accuracy, macro-F1, and ROC-AUC, although the confidence intervals included zero after Holm correction. In contrast, SGRI-HF significantly improved ROC-AUC over SGRI-Net (difference = 0.0161, 95% CI: 0.0064 – 0.0259, adjusted p = 0.005), confirming that heterogeneous neural – tree fusion provided additional discriminative information beyond the semantic neural branch alone. This finding supports the hypothesis that semantic representation learning and

tree-based decision modeling capture partially complementary patterns in questionnaire data.

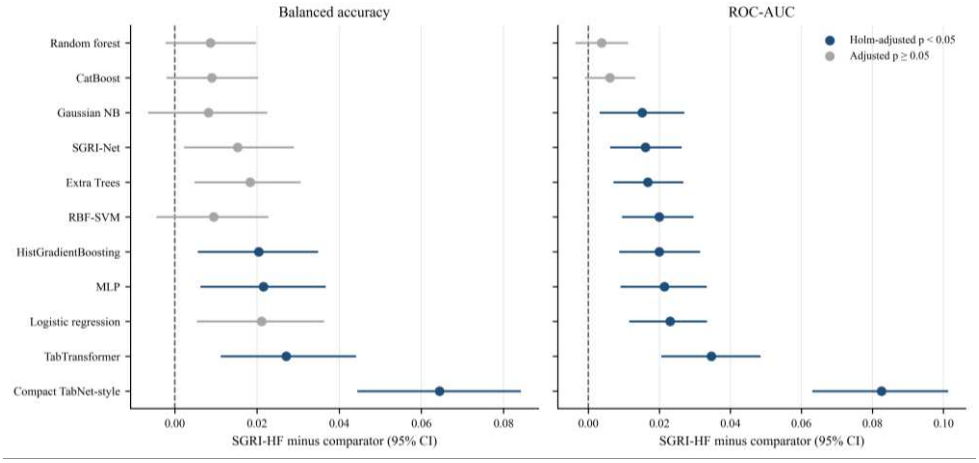


Fig.8 Paired bootstrap differences in balanced accuracy and ROC-AUC.

5.3 Architectural Contributions and Heterogeneous Fusion

The ablation study was conducted to investigate the contribution of individual components within SGRI-Net under controlled experimental settings. All variants were evaluated using identical data partitions, preprocessing procedures, and decision thresholds. The complete SGRI-Net achieved a balanced accuracy of 0.6289, macro-F1 of 0.6289, ROC-AUC of 0.6737, and Brier score of 0.2270 (Table 6).

Table 6 Pooled out-of-fold performance of the SGRI-Net ablation variants.

Variant	Balanced accuracy	MCC	At-risk F1	Macro-F1	ROC-AUC	Brier score
Complete SGRI-Net	0.6289	0.2579	0.6133	0.6289	0.6737	0.2270
Without auxiliary supervision	0.6282	0.2565	0.6123	0.6282	0.6737	0.2270
Without consistency regularization	0.6269	0.2538	0.6117	0.6269	0.6740	0.2270
Without semantic gating	0.6276	0.2553	0.6135	0.6277	0.6726	0.2274
Without pairwise interaction	0.6278	0.2556	0.6154	0.6278	0.6728	0.2276
Without global residual pathway	0.6275	0.2562	0.6007	0.6273	0.6637	0.2304

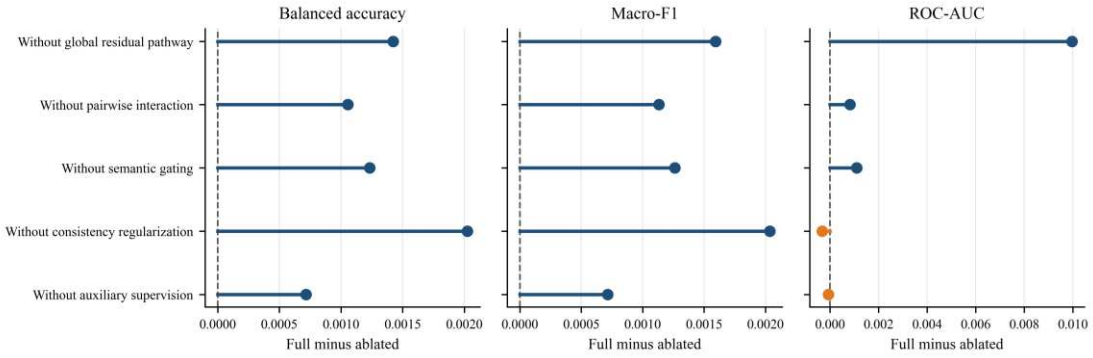


Fig.9 Changes induced by removing individual SGRI-Net components.

Removing the global residual pathway resulted in the largest performance degradation among all ablation variants. Specifically, ROC-AUC decreased from 0.6737 to 0.6637, while at-risk F1 decreased from 0.6133 to 0.6007 and the Brier score increased from 0.2270 to 0.2304. This finding indicates that the residual pathway provides complementary global information that may be partially lost during

semantic-domain aggregation and interaction learning.

The remaining components showed relatively smaller but interpretable effects. Removing semantic gating caused a slight decrease in ROC-AUC, indicating that sample-specific domain weighting provides adaptive information beyond uniform aggregation. The pairwise interaction module also contributed directionally to representation learning, whereas auxiliary supervision and consistency regularization mainly served as representation constraints rather than directly improving every metric. These findings suggest that the components provide complementary functions within SGRI-Net, with residual learning preserving global information and gating enabling personalized domain allocation.

Heterogeneous fusion provided the most evident improvement at the framework level. Unlike component-level regularization, heterogeneous fusion directly expanded the representation space by combining complementary inductive biases. Compared with SGRI-Net alone, SGRI-HF improved pooled ROC-AUC by 0.0161 (95% CI: 0.0064 – 0.0259, adjusted p = 0.005). This result indicates that the neural and tree branches captured complementary predictive patterns: SGRI-Net learned semantically organized psychosocial interactions, while CatBoost captured nonlinear threshold relationships from original questionnaire variables. Their combination therefore enhanced the representation of heterogeneous questionnaire information.

5.4 Probability Calibration and Model Interpretation

Beyond discrimination performance, the reliability of predicted probabilities is important for questionnaire-based mental health risk identification because practical screening decisions often require prioritizing follow-up resources. Based on pooled out-of-fold predictions, SGRI-HF achieved a ROC-AUC of 0.6898 and an at-risk average precision of 0.6774 (Figure 10A-B). The precision-recall curve remained above the empirical prevalence level over a substantial recall range, indicating that the model provided meaningful ranking ability under the observed risk distribution. However, the close curves among SGRI-HF, random forest, CatBoost, and SGRI-Net further confirm that the discrimination improvement was incremental rather than substantial.

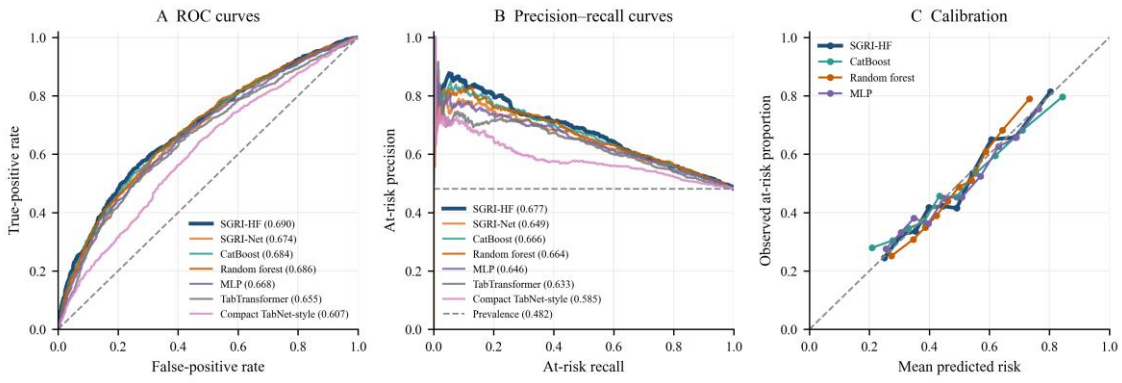


Fig.10 Pooled out-of-fold discrimination and calibration. (A) ROC curves. (B) At-risk precision-recall curves with empirical prevalence as the no-skill reference. (C) Quantile-based calibration curves.

Probability calibration provided additional evidence supporting the heterogeneous fusion strategy. SGRI-HF achieved the lowest pooled Brier score (0.2215) and log loss (0.6335) among all evaluated models. The calibration curve showed relatively good agreement between predicted and observed probabilities in the central probability range, with larger deviations occurring near the extreme regions. These results suggest that the fused model provided more reliable probability estimation within the evaluated cohort. Nevertheless, calibration performance remains dataset-dependent and should be

reassessed before deployment in other universities or populations.

The two branches of SGRI-HF provided complementary perspectives for model interpretation. Within SGRI-Net, semantic-domain gate weights reflected participant-specific allocation among psychosocial domains rather than fixed global importance. The highest average weights were assigned to coping and social support (0.1997 ± 0.0791), followed by lifestyle (0.1461 ± 0.0454), economic and digital behavior (0.1453 ± 0.0354), and sources of stress (0.1422 ± 0.0389) (Table 7). The relatively large variation in coping and social support weights further indicates that different students may exhibit distinct combinations of risk-related factors.

Table 7 Out-of-fold semantic-domain gate weights from the SGRI-Net branch.

Semantic domain	Mean	SD	Median
Coping and social support	0.1997	0.0791	0.1865
Lifestyle	0.1461	0.0454	0.1396
Economic and digital behavior	0.1453	0.0354	0.1423
Sources of stress	0.1422	0.0389	0.1380
Campus support needs	0.1330	0.0364	0.1323
Demographic characteristics	0.1195	0.0367	0.1125
Academic behavior	0.1142	0.0356	0.1119

At the feature level, the CatBoost branch provided a complementary interpretation through SHAP analysis. Peer support showed the largest mean absolute SHAP magnitude on the at-risk log-odds scale (0.1498), followed by campus participation, grade, sleep duration, mobile-phone dependence, entertainment screen time, family communication frequency, family support, and sleep quality (Figure 11). The consistency among SHAP results, semantic-domain gate patterns, and univariable associations suggests that SGRI-HF relied on distributed information from social support, lifestyle, digital behavior, and academic context rather than a single dominant predictor.

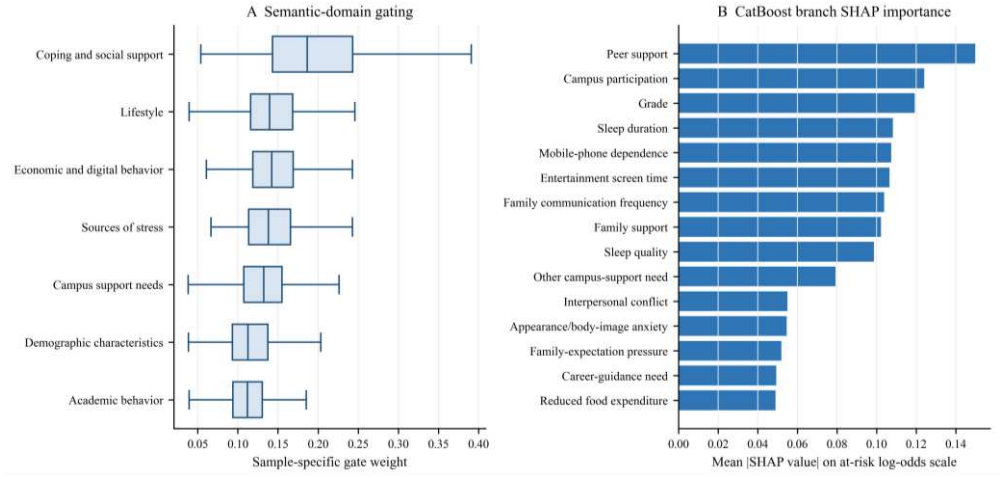


Fig.11 Dual-view interpretation of the heterogeneous framework. (A) Participant-level distributions of semantic-domain gate weights from SGRI-Net. (B) Fifteen largest mean absolute OOF SHAP values from the CatBoost branch on the at-risk log-odds scale.

The interpretation results should be understood as predictive explanations rather than causal evidence. Gate weights represent adaptive domain allocation within the neural branch, whereas SHAP values quantify feature contributions within the tree branch; neither directly decomposes the final fused probability into causal effects. Therefore, these analyses are primarily useful for identifying recurrent risk-related patterns and generating hypotheses for future longitudinal validation or intervention studies.

5.5 Error Characterization, Subgroup Analysis, and Threshold Sensitivity

At the prespecified decision threshold of 0.5, SGRI-HF correctly classified 1,063 of 1,500 not-at-risk students and 808 of 1,394 at-risk students. The pooled out-of-fold confusion matrix contained 437 false positives and 586 false negatives, corresponding to a specificity of 0.7087 and an at-risk sensitivity of 0.5796 (Figure 12A). The class-conditional probability distributions showed substantial overlap around the decision boundary (Figure 12B), suggesting that classification errors were mainly related to intrinsically ambiguous questionnaire patterns rather than an inappropriate threshold alone.

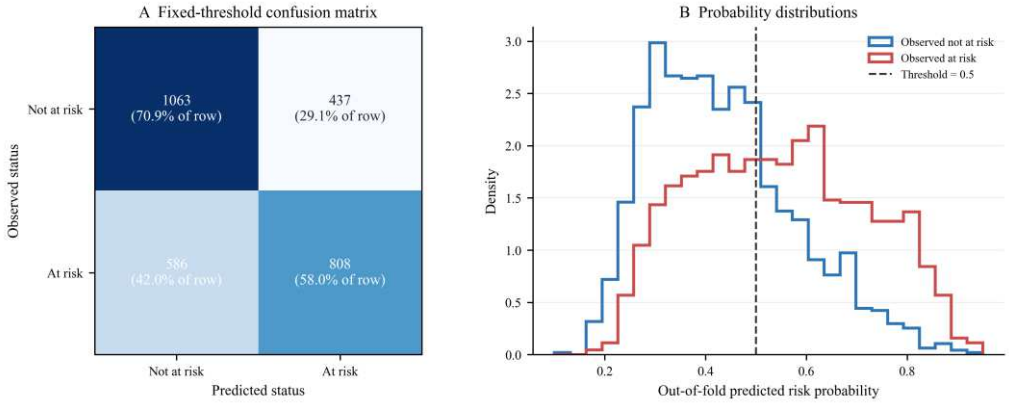


Fig.12 Fixed-threshold error profile of SGRI-HF. (A) Pooled OOF confusion matrix at the prespecified threshold of 0.5. (B) Class-conditional distributions of OOF risk probabilitie.

Exploratory subgroup analysis showed comparable discrimination between male and female students, with overlapping ROC-AUC confidence intervals (0.6844 vs. 0.6935). Grade-level performance showed larger variability, although the limited number of senior students resulted in wider uncertainty intervals. These findings should be interpreted as exploratory rather than evidence of subgroup differences.

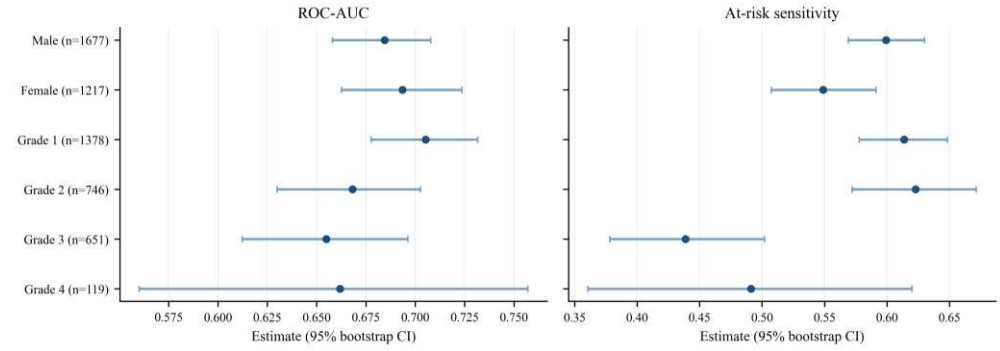


Fig.13 Exploratory sex- and grade-stratified SGRI-HF performance.

To further characterize false-negative cases, missed-risk analysis was conducted among the 1,394 students classified as at risk. At the fixed threshold, 586 students (42.0%) were incorrectly classified as not at risk. The strongest differences between correctly identified and missed at-risk cases were observed for peer support, sleep quality, family support, mobile-phone dependence, and family communication frequency (Figure 14). These associations describe the model's error structure rather

than causal mechanisms. In particular, some missed cases may present relatively favorable responses in highly weighted domains while still meeting the questionnaire-defined risk criterion through other combinations of factors.

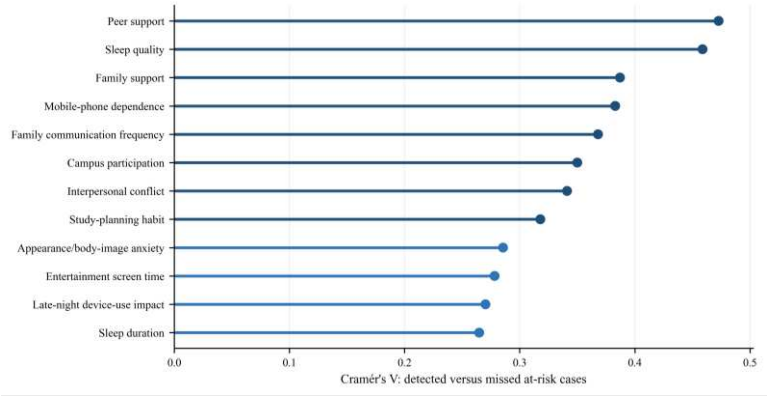


Fig.14 Effect-size profile of the twelve features most strongly differentiating detected and missed at-risk cases at the fixed 0.5 threshold.

Finally, threshold sensitivity analysis was performed using pooled out-of-fold probabilities to examine alternative operating points. Lowering the threshold from 0.500 to 0.430 increased sensitivity from 0.580 to 0.712 and reduced false negatives from 586 to 402, but decreased specificity from 0.709 to 0.538. Further threshold reductions produced additional sensitivity gains at the cost of substantially increased false positives (Table 8 and Figure 15). These results demonstrate the expected trade-off between identifying more potential at-risk students and increasing follow-up burden. Therefore, the operating threshold should be selected according to the intended screening objective and validated in the target population rather than optimized solely on the internal dataset.

Table 8 Exploratory SGRI-HF operating points derived from pooled OOF probabilities.

Threshold	Precision	Sensitivity	Specificity	At-risk F1	Accuracy	FN	FP
0.500	0.649	0.580	0.709	0.612	0.647	586	437
0.430	0.589	0.712	0.538	0.644	0.622	402	693
0.410	0.577	0.750	0.489	0.652	0.615	348	767
0.375	0.558	0.814	0.401	0.662	0.600	259	898
0.350	0.541	0.852	0.327	0.662	0.580	206	1009

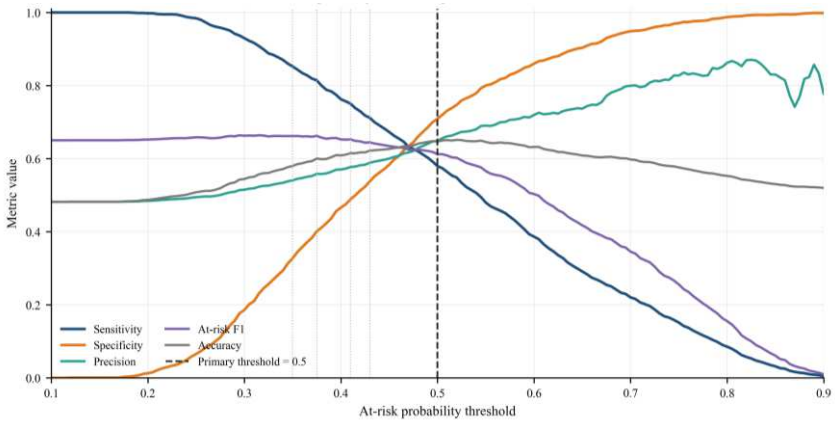


Fig.15 Exploratory threshold–performance trade-off for pooled SGRI-HF OOF probabilities.

Taken together, the experimental analyses demonstrate that mental health risk exhibits

multidomain patterns, SGRI-HF benefits from integrating semantic neural representation with tree-based modeling, and the resulting framework provides interpretable probability estimation for auxiliary risk identification.

6. Conclusions

This study proposed SGRI-HF, a Semantic Group-Aware Gated Residual Interaction and Heterogeneous Fusion framework, for questionnaire-based college student mental health risk identification. By incorporating semantic-domain representation learning, adaptive interaction modeling, and heterogeneous neural–tree fusion, the proposed framework provides a structured approach for modeling complex psychosocial information in questionnaire data.

Under a leakage-controlled evaluation protocol, SGRI-HF achieved a favorable overall predictive profile among the evaluated approaches, demonstrating that explicitly modeling semantic structures and combining complementary learning paradigms can improve risk identification from questionnaire data. The experimental analyses further revealed that mental health risk patterns were distributed across multiple psychosocial domains, particularly social support, lifestyle, digital behavior, academic characteristics, and stress-related factors, rather than being dominated by individual predictors. These findings support the importance of domain-aware representation and interaction learning for complex mental health-related prediction tasks. Nevertheless, SGRI-HF should be considered an auxiliary risk identification approach rather than a replacement for professional psychological assessment. The remaining challenge of balancing missed-risk cases and false-positive burden highlights the need for application-specific threshold selection. Moreover, the current study was limited by its cross-sectional design, questionnaire-based outcome definition, and validation within a single institutional cohort. External validation across diverse universities and populations is required to further assess model generalizability.

Future research should focus on longitudinal modeling of mental health trajectories, multi-source data integration, external calibration, and sensitivity-oriented optimization. Such extensions may further improve the reliability and practical utility of SGRI-HF in supporting early identification and targeted mental health interventions among college students.

Data Availability

The data that support the findings of this study are available on request from the corresponding author.

Declaration of competing interest

The authors declare that they have no competing interests.

References

[1] G. M. Fitzmaurice, N. M. Laird, G. E. P. Zahner, and C. Daskalakis, “Bivariate Logistic Regression Analysis of Childhood Psychopathology Ratings using Multiple Informants,” Am. J. Epidemiol., vol. 142, no. 11, pp. 1194–1203, Dec. 1995, doi: 10.1093/oxfordjournals.aje.a117578.

[2] G. H. Alishiri, N. Bayat, A. F. Ashtiani, S. A. Tavallaii, S. Assari, and Y. Moharamzad, “Logistic regression models for predicting physical and mental health-related quality of life in rheumatoid arthritis patients,” Mod. Rheumatol., vol. 18, no. 6, pp. 601–608, 2008, doi: 10.3109/s10165-008-0092-6.

[3] D. A. Adler, F. Wang, D. C. Mohr, and T. Choudhury, “Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies,” PLoS One, vol. 17, no. 4, pp. e0266516-, Apr. 2022, [Online]. Available: https://doi.org/10.1371/journal.pone.0266516

[4] H. Oh et al., “Food insecurity and mental health among young adult college students in the United States,” J. Affect. Disord., vol. 303, pp. 359–363, 2022, doi: https://doi.org/10.1016/j.jad.2022.02.009.

[5] R. Massé, C. Poulin, C. Dassa, J. Lambert, S. Bélair, and A. Battaglini, “The Structure of Mental Health: Higher-Order Confirmatory Factor Analyses of Psychological Distress and Well-Being Measures,” Soc. Indic. Res., vol. 45, no. 1, pp. 475–504, 1998, doi: 10.1023/A:1006992032387.

[6] S. Manne and R. Schnoll, “Measuring cancer patients’ psychological distress and well-being: A factor analytic assessment of the mental health inventory,” Psychol. Assess., vol. 13, no. 1, pp. 99–109, 2001, doi: 10.1037/1040-3590.13.1.99.

[7] Robert F Dedrick and Paul E Greenbaum, “Multilevel Confirmatory Factor Analysis of a Scale Measuring Interagency Collaboration of Children’s Mental Health Agencies,” J. Emot. Behav. Disord., vol. 19, no. 1, pp. 27–40, Mar. 2011, doi: 10.1177/1063426610365879.

[8] A. R. Stiffman, E. Hadley-Ives, D. Elze, S. Johnson, and P. Doré, “IMPACT OF ENVIRONMENT ON ADOLESCENT MENTAL HEALTH AND BEHAVIOR: Structural Equation Modeling,” American Journal of Orthopsychiatry, vol. 69, no. 1, pp. 73–86, Jan. 1999, doi: https://doi.org/10.1037/h0080383.

[9] Y. Xiao, S. Miao, C. Sarkar, H. Geng, and Y. Lu, “Exploring the Impacts of Housing Condition on Migrants’ Mental Health in Nanxiang, Shanghai: A Structural Equation Modelling Approach,” Int. J. Environ. Res. Public Health, vol. 15, no. 2, 2018, doi: 10.3390/ijerph15020225.

[10] S. Cen, M. Zhao, F. Wang, and L. Tang, “Gender differences in the relationship between mental health and academic performance among undergraduate students at a medical school in Shanghai: a cross-sectional study,” BMC Public Health, vol. 25, no. 1, p. 731, 2025.

[11] X.-L. Mao and H.-M. Chen, “Investigation of contemporary college students’ mental health status and construction of a risk prediction model,” World J. Psychiatry, vol. 13, no. 8, p. 573, 2023.

[12] M. Srividya, S. Mohanavalli, and N. Bhalaji, “Behavioral Modeling for Mental Health using Machine Learning Algorithms,” J. Med. Syst., vol. 42, no. 5, p. 88, 2018, doi: 10.1007/s10916-018-0934-5.

[13] A. E. Tate, R. C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, and R. Kuja-Halkola, “Predicting mental health problems in adolescence using machine learning techniques,” PLoS One, vol. 15, no. 4, pp. e0230389-, Apr. 2020, [Online]. Available: https://doi.org/10.1371/journal.pone.0230389

[14] J. Chung and J. Teo, “Single classifier vs. ensemble machine learning approaches for mental health prediction,” Brain Inform., vol. 10, no. 1, p. 1, 2023, doi: 10.1186/s40708-022-00180-6.

[15] A. Baba and K. Bunji, “Prediction of Mental Health Problem Using Annual Student Health Survey: Machine Learning Approach,” JMIR Ment Health, vol. 10, p. e42420, 2023, doi: 10.2196/42420.

[16] M. A. Rahman and T. Kohli, “Mental health analysis of international students using machine learning techniques,” PLoS One, vol. 19, no. 6, pp. e0304132-, Jun. 2024, [Online]. Available: https://doi.org/10.1371/journal.pone.0304132

[17] M. R. Gohari et al., “Using random forest to identify correlates of depression symptoms among adolescents,” Soc. Psychiatry Psychiatr. Epidemiol., vol. 59, no. 11, pp. 2063–2071, 2024.

[18] F. Aliverdi et al., “Social networks and internet emotional relationships on mental health and quality of life in students: structural equation modelling,” BMC Psychiatry, vol. 22, no. 1, p. 451, 2022, doi: 10.1186/s12888-022-04097-6.

[19] R. Garriga et al., “Machine learning model to predict mental health crises from electronic health records,” Nat. Med., vol. 28, no. 6, pp. 1240–1248, 2022, doi: 10.1038/s41591-022-01811-5.

[20] I. Chien et al., “A Machine Learning Approach to Understanding Patterns of Engagement With Internet-Delivered Mental Health Interventions,” JAMA Netw. Open, vol. 3, no. 7, pp. e2010791–e2010791, Jul. 2020, doi: 10.1001/jamanetworkopen.2020.10791.

[21] K. Dadi, G. Varoquaux, J. Houenou, D. Bzdok, B. Thirion, and D. Engemann, “Population modeling with machine learning can enhance measures of mental health,” Gigascience, vol. 10, no. 10, p. giab071, Oct. 2021, doi: 10.1093/gigascience/giab071.

[22] H. W. Yeung et al., “Predicting sex, age, general cognition and mental health with machine learning on brain structural connectomes,” Hum. Brain Mapp., vol. 44, no. 5, pp. 1913–1933, Apr. 2023, doi: https://doi.org/10.1002/hbm.26182.

[23] M. Wan and S. Zou, “Adolescent mental health state assessment framework by combining YOLO with random forest,” Appl. Soft Comput., vol. 168, p. 112497, 2025.

[24] R. D. Riley et al., “A general sample size framework for developing or updating a predictive algorithm: with application to clinical prediction models,” BMC Med. Res. Methodol., vol. 26, no. 1, p. 121, 2026.

[25] L. Wang, E. Shi, B. Meyers, P. Vlachos, J. Tcheng, and S. Denardo, “Machine learning performance for a small dataset: random oversampling improves data imbalances and fairness,” BMC Med. Res. Methodol., 2026.

[26] B. Sun, J. Zhang, Y. Ma, and H. He, “A proof-of-concept machine learning model for short-term suicide risk stratification in depressed youth,” Transl. Psychiatry, vol. 16, no. 1, p. 187, 2026.

[27] K. Zantvoort, B. Nacke, D. Görlich, S. Hornstein, C. Jacobi, and B. Funk, “Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions,” NPJ Digit. Med., vol. 7, no. 1, p. 361, 2024, doi: 10.1038/s41746-024-01360-w.

[28] S. Ö. Arik and T. Pfister, “Tabnet: Attentive interpretable tabular learning,” in Proceedings of the AAAI conference on artificial intelligence, 2021, pp. 6679–6687.

[29] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, “Tabpfn: A transformer that solves small tabular classification problems in a second,” arXiv preprint arXiv:2207.01848, 2022.

[30] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, “Revisiting deep learning models for tabular data,” Adv. Neural Inf. Process. Syst., vol. 34, pp. 18932–18943, 2021.

[31] S. J. J. Jui, R. C. Deo, R. Acharya, P. D. Barua, J. Soar, and A. Devi, “Automated stress detection with explainable-AI transformer TabNet technique and physiological signals,” Biomed. Signal Process. Control, vol. 113, p. 109208, 2026.

[32] Q. Yuan, W. Sun, F. Li, X. Dong, and Y. Yuan, “A dual attention transformer modelling for explainable mental health analysis in academic environments using TaBERT,” Sci. Rep., vol. 16, no. 1, p. 11201, 2026.

[33] Y.-H. Lee, J. H. Lee, Q.-S. Auh, S. Lee, D. Nixdorf, and A. Chaurasia, “TMD diagnosis using a masked self-supervised tabular transformer,” J. Dent. Res., vol. 105, no. 6, pp. 717–726, 2026.

[34] U. Habiba and S. A. Ludwig, “PSO-optimized TabTransformer architecture with feature engineering for enhanced cervical cancer risk prediction,” Expert Syst. Appl., vol. 300, p. 130484, 2026.

[35] R. Shwartz-Ziv and A. Armon, “Tabular data: Deep learning is not all you need,” Information Fusion, vol. 81, pp. 84–90, 2022, doi: https://doi.org/10.1016/j.inffus.2021.11.011.

[36] L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why do tree-based models still outperform deep learning on typical tabular data?,” Adv. Neural Inf. Process. Syst., vol. 35, pp. 507–520, 2022.

[37] G. Kang, L. Ding, J. Liu, B. Cao, and Y. Xu, “Web API recommendation based on self-attentional neural factorization machines with domain interactions,” IEEE Trans. Netw. Sci. Eng., vol. 10, no. 6, pp. 3953–3963, 2023.

[38] Y. Chen, Y. Wang, P. Ren, M. Wang, and M. de Rijke, “Bayesian feature interaction selection for factorization machines,” Artif. Intell., vol. 302, p. 103589, 2022.

[39] J. Liu, T. Tian, Y. Liu, S. Hu, and M. Li, “iTabNet: an improved neural network for tabular data and its application to predict socioeconomic and environmental attributes,” Neural Comput. Appl., vol. 35, no. 15, pp. 11389–11402, 2023.

[40] W. Xu and Z. Xie, “Deep categorical clustering via symbolization and masking mechanisms,” Pattern Recognit., p. 113002, 2025.

[41] Y. Guo, J. Chen, Q. Du, A. Van Den Hengel, Q. Shi, and M. Tan, “Multi-way backpropagation for training compact deep neural networks,” Neural networks, vol. 126, pp. 250–261, 2020.

[42] J. Luo and S. Xu, “NCART: Neural Classification and Regression Tree for tabular data,” Pattern Recognit., vol. 154, p. 110578, 2024.

[43] Z. Li, H. Lei, E. Ma, J. Lai, and J. Qiu, “Ensemble technique to predict post-earthquake damage of buildings integrating tree-based models and tabular neural networks,” Comput. Struct., vol. 287, p. 107114, 2023.

[44] X. Zhu, J. Hu, T. Xiao, S. Huang, Y. Wen, and D. Shang, “An interpretable stacking ensemble learning framework based on multi-dimensional data for real-time prediction of drug concentration: The example of olanzapine,” Front. Pharmacol., vol. 13, p. 975855, 2022.