ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Advanced Machine Learning Model for Early Mental Health Risk Detection in Rural Chinese Adolescents

Fang Cheng[1,*]

[1] Hubei Engineering University, XiaoGan, HuBei Province, China

SUMMARY: *The mental health risks of rural adolescents have the characteristics of strong concealment, fast evolution speed and obvious scene differences. Traditional questionnaire screening and manual interview are difficult to meet the needs of early identification. This paper proposes an advanced machine learning detection model for rural adolescents in China, which integrates learning behavior, sleep rhythm, social support, psychological assessment and weekly diary text into a unified computing framework, and realizes low risk, attention risk and high risk identification through multi-source feature construction, emotional risk embedding and two-stage hierarchical classification. Experimental results show that the model achieves 91.40% Accuracy, 0.891 Macro-F1 and 0.931 AUROC under 5-fold cross validation, and the high risk recall reaches 0.860. The research shows that the combination of multimodal fusion and interpretable analysis can more effectively capture the early psychological risk signals of rural adolescents, and provide a computable basis for school grading intervention.*

KEYWORDS: *Rural youth; Mental health risk detection; Multi-source feature fusion; Explainable Machine Learning*

1 Introduction

With the development of digital mental health monitoring, adolescent psychological risk identification is shifting from post-intervention to front-end warning. For rural areas, this shift is particularly important. Due to the combined influence of family mobility, insufficient coverage of school psychological services, long-term left-behind, sleep imbalance and weak peer support, some adolescents' emotional fluctuations do not appear as obvious symptoms in the early stage, but leave measurable traces of changes in learning behavior, sleep rhythm, classroom participation and text expression. Traditional screening methods mostly rely on questionnaires, interviews or periodic manual observation. Although they have the advantages of intuitive interpretation, they have obvious limitations in high-frequency monitoring, continuous tracking and large sample coverage, which are especially difficult to adapt to the real scene of limited resources and rapid changes in students' status in rural schools. Machine learning methods provide new technical entrances to this problem. By jointly modeling structured behavior data, psychological scale results, text feedback and life rule information, the model can identify potential risks from subtle and scattered signals, and give hierarchical judgments when risks are not yet explicit. However, there are still three shortcomings in the existing research. Firstly, the models are mostly oriented to general adolescent samples, and the data heterogeneity in rural scenes in China is not considered enough. Second,

*13469792940@sina.cn
https://doi.org/10.65102/is2026327
Published: 30 April, 2026

single-modal modeling is common, and the linkage between emotion, behavior and environmental variables is still weak. Third, although some methods have high classification accuracy, it is difficult to explain "why they are judged as high risk", which weakens their practical usability in school early warning and graded intervention. Based on this, this paper constructs an early mental health risk detection model for rural adolescents in China, which integrates multi-source feature representation, emotional signal embedding and hierarchical classification mechanism into the same computational framework to improve the sensitivity, robustness and interpretability of the model.

1.1 Definition of research questions and application scenarios

The core issue of this paper is how to use multi-source data to establish a scalable, interpretable, and early responsive mental health risk detection model in the rural education and growth environment, so that the system can not only identify the risk level, but also provide a basis for subsequent intervention. The application scenarios of the model mainly include psychological screening in township middle schools, student status tracking in boarding schools, risk warning for left-behind adolescents, and group trend identification in regional education governance. Around this problem, the research objectives of this paper are as follows:

(1) A psychological risk identification framework for rural adolescents was constructed, and the basic individual attributes, learning behaviors, sleeping and living habits, emotional scales and text expression information were uniformly coded and fused.

(2) An early hierarchical detection mechanism is designed to divide the samples into three categories: low risk, concern risk and high risk, so as to improve the recognition ability of the model for a few high-risk individuals.

(3) The interpretative analysis method was introduced to identify the key factors affecting the classification results, so as to provide traceable evidence support for the collaborative intervention of school psychology teachers, class teachers and family school.

1.2 Model idea and technical route

In response to the above problems, this paper adopts the technical route of "data acquisition - feature construction - emotion representation - risk detection - interpretation output". The structured input included family company, sleep duration, physical activity frequency, learning engagement, peer relationship score and psychological scale sub-item results. The unstructured input mainly comes from student weekly diaries, open-ended questions and answers, and short self-explanatory texts. Aiming at the characteristics of many missing values, unbalanced class distribution and fuzzy risk boundaries in rural samples, this paper introduces missing compensation, anomaly correction and class reweighting strategies in the preprocessing stage. In the representation learning stage, the multi-branch coding structure is used to extract behavioral features, emotion features and text semantic features, and the attention fusion mechanism is used to form a unified risk representation. Then, the model used the hierarchical classification strategy to complete the risk level discrimination, and combined with SHAP analysis to output the class-level influencing factors. In the experimental stage, cross-validation is used to evaluate the performance of the model on indicators such as accuracy, Macro-F1, AUROC and high-risk recall, and its stability in different gender, grade and left-behind subgroups is investigated to verify the promotion value of the model in real education scenarios.

2 Related Research

2.1 Research on machine learning methods in mental health risk identification

In recent years, computational research on mental health risk identification has gradually shifted from traditional scale statistical analysis to machine learning driven multi-source data modeling. Liu et al. [2] systematically reviewed the research on depression detection based on social media, and pointed out that text features, emotional polarity and behavior frequency can provide persistent signals for mental state recognition, but platform context noise and sample bias still affect the model extrapolation ability. Mendes et al. [3] and Bufano et al. [4] further summarized the application of mobile perception, wearable devices and public datasets in psychological monitoring from the perspective of digital phenotypes, indicating that sleep, activity intensity, position change and device interaction behavior have become important computational inputs. Andrew et al. [5] discussed the application of artificial intelligence from the three levels of diagnosis, prognosis assessment and intervention support of adolescent psychological disorders, and believed that in addition to model performance improvement, fairness, privacy protection and scene adaptation also determine whether the system can really be implemented. In general, these studies have demonstrated that machine learning can extract risk patterns from heterogeneous data, but more detailed scenario modeling is still needed when targeting specific groups.

2.2 Research on early detection model of adolescent psychological risk

Within the adolescent group, the early detection model has been extended from single depression prediction to multiple directions such as suicidal ideation, non-suicidal NSSI, and comprehensive psychological risk identification. Huang et al. [6] compared the performance of three machine learning models in the prediction of depression and suicidal ideation in Chinese adolescents, and showed that the multivariate machine learning method was superior to simple linear discrimination. Zhou et al. [7] combined artificial neural network with random forest to identify depression risk in large samples of adolescents, which reflects the adaptability of integrated modeling to complex nonlinear relationships. Kim et al. [8] constructed a prediction model of suicidal ideation in a cross-regional adolescent cohort, indicating that external validation is crucial to improve the robustness of the model. Li et al. [9] and Zhong et al. [10] respectively modeled depression, suicidal ideation and non-suicidal NSSI, which further showed that although there were common risk factors for different psychological problems, their characteristic contribution structures were different. Yang et al. [11] used longitudinal data to predict the suicidal ideation of preadolescent children, and Yoo et al. [12] tried to introduce earlier growth information into the prediction of adolescent depression, showing the research trend of "early risk signal moving forward". Luo et al. [13] established a risk prediction tool for Chinese rural adolescents, which directly revealed the distinctive scene characteristics of rural samples in terms of family companishment, social support and growth pressure. Although these studies have promoted the development of psychological early warning models for adolescents, many works still remain in the cross-sectional data framework, and do not pay enough attention to the problems of unbalanced samples, sparse text information and insufficient explanation feedback in rural areas.

2.3 Hierarchical intervention and explanation support strategies in AI-driven psychological early warning

The value of a psychological early warning system is not only in "identifying who is at risk", but also in "explaining where the risk comes from and how to respond hierarchously". Jiang et al. [14] and She et al. [15] pointed out that the depression, anxiety and emotional distress of children and adolescents in rural areas had a more complex social environmental background, and the urban-rural differences should not be simply regarded as a single regional difference, but should be understood as the comprehensive projection of differences in family resources, educational support and development opportunities at the psychological level. Zhang et al. [16] analyzed the influencing factors of Chinese adolescents' mental health from the three levels of individual, school and province, suggesting that psychological risks have obvious hierarchical structure, which is suitable for combining with hierarchical early warning mechanism. Jing et al. [17] analyzed and compared the association patterns of depressive symptoms among urban and rural adolescents through network analysis, indicating that the symptoms association patterns within different groups were not consistent, which provided a basis for category-level interpretation and precise intervention. Zhao et al. [18], Liu et al. [19], Ruan et al. [20], Mei et al. [21] and Bao et al. [22] respectively revealed the important trigger factors of rural adolescents' psychological risks from the perspectives of living habits, sleep quality, left-behind experience, family disorder and social support. While not all of these studies directly employ complex deep learning architectures, they provide a solid foundation for feature selection, risk stratification, and interpretation of output in AI systems. In other words, if the warning model only outputs a single probability value, but cannot map the risk to the intervalable dimensions such as sleep imbalance, insufficient family support, impaired peer relationship or elevated learning pressure, its educational application value is still limited.

Table 1: Overview of the main directions and shortcomings of related research

Research Direction	Representative References	Data / Objects	Computational Method Characteristics	Existing Limitations and Implications for This Study
General mental health identification	[2]–[5]	Social media, mobile sensing, wearable devices, and review-based data	Text mining, sentiment analysis, digital phenotyping, and multimodal sensing	Strong generalizability across scenarios, but insufficient adaptation to rural Chinese adolescents
Prediction of adolescent depression and suicide risk	[6]–[12]	Chinese and international adolescent samples	Random forest, neural networks, ensemble learning, and longitudinal prediction	Greater focus on high-risk category detection, but interpretability and integration of localized contextual variables remain limited
Specialized research on rural adolescents	[13]–[15], [18]–[22]	Rural middle school students, left-behind children, and urban–rural comparative samples	Risk factor analysis, predictive tool construction, and support-variable modeling	Provides key risk clues, but has not yet formed a unified intelligent detection framework based on multi-source fusion
Hierarchical intervention and interpretability support	[16]–[17], [21]–[22]	Multi-level samples spanning individuals, schools, and regions	Multi-level factor analysis, network relationship analysis, and support-path identification	Helpful for intervention design, but still insufficiently integrated with interpretable machine learning

It can be seen from Table 1 that the existing research has provided a relatively rich method basis for psychological risk identification of adolescents, and also revealed the action paths of several key variables in rural scenes. However, the real early detection system for rural adolescents in China still needs to incorporate behavioral data, emotional text, life rules and social support factors into the unified representation space, and output interpretable risk basis while ensuring classification performance. In this paper, we address this gap and strive to build advanced machine learning models that balance accuracy, interpretability and scene adaptability.

3 Model Framework

The key to early detection of mental health risk for rural adolescents in China is not the static interpretation of single scale scores, but the transformation of weak signals scattered in the learning process, life rules, family environment and emotional expression into a unified representation that can be calculated, comparable and traceable. Based on this understanding, this paper constructs a model framework consisting of "multi-source data acquisition - quality control and preprocessing - feature representation learning - risk signal embedding - hierarchical detection output". The framework not only retains the realistic oriented variables such as family support, left-behind state, sleep quality and school participation in rural scenes, but also introduces text emotion information and machine learning representation mechanism to enhance the sensitivity of the model to early abnormal fluctuations. Figure 1 shows the overall process of the model.

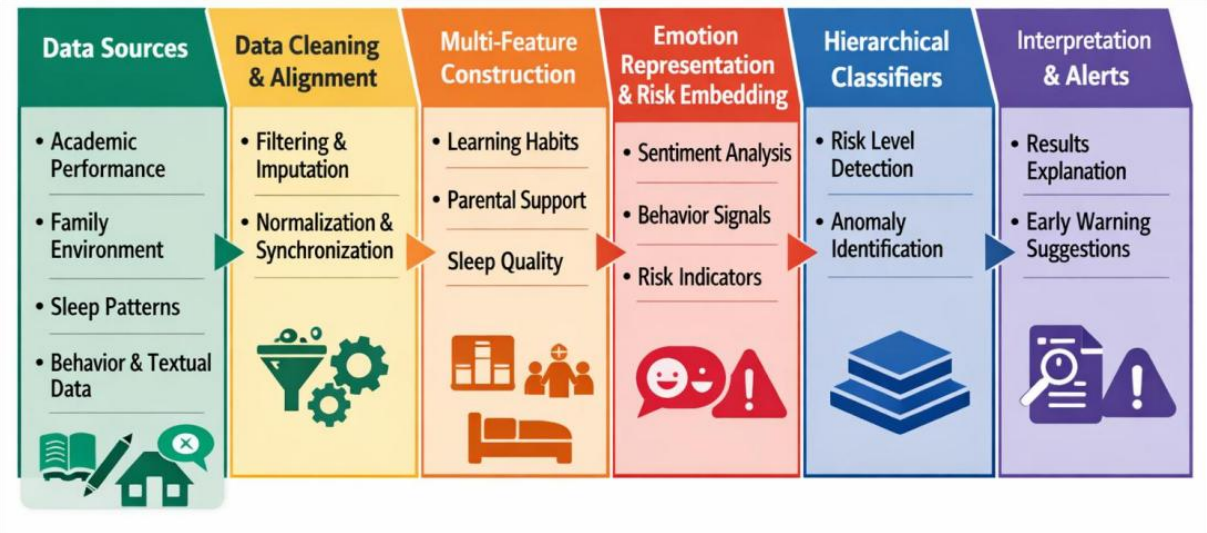

Figure 1: Overall flow chart of the early detection model of mental health risk for rural adolescents

3.1 Data collection and preprocessing of rural adolescent mental health

In order to ensure sufficient stability and scene adaptability of subsequent model training, the design of data layer is no longer limited to a single psychological questionnaire, but integrates individual basic attributes, learning behaviors, living habits, emotional evaluation and open text into a unified collection system. The structured variables included gender, grade, whether left behind, frequency of family companionship, sleep duration in the past month, frequency of physical activity, classroom participation, homework completion rate, peer relationship score, and scale items of depression, anxiety, and stress. The unstructured variables were

mainly from weekly notes, self-reporting essays and open-ended question and answer texts. For the ith sample, its original observation is denoted as in this paper:

$$x_i = \left[x_i^{(b)}, x_i^{(l)}, x_i^{(p)}, x_i^{(t)}\right] \tag{1}$$

Among them, $x_i^{(b)}$ represents basic demographic and family background characteristics, $x_i^{(l)}$ represents learning and life behavior characteristics, $x_i^{(p)}$ represents psychological evaluation characteristics, and xi(t) represents text characteristics. Such a representation is able to preserve the coupling relationship between behavior, emotion and environment in the same input space.

In the original collection stage, rural samples are often accompanied by missing, abnormal and different recording caliber. Some students have missing weekly records, missing sleep data, or partial vacancies in questionnaires. If they directly enter the training process, it is easy to cause the deviation of category boundaries. To this end, this paper adopts a joint processing strategy of "missing indication + classification filling". For continuous variables, the mean values of the same age group and the same school group were used to compensate. For discrete variables, the inpainting is completed by the intra-group mode or the nearest neighbor estimation based on adjacent observations. The form can be written as:

$$\tilde{x}_{ij} = m_{ij}x_{ij} + (1 - m_{ij})\bar{x}_j^{(g)} \tag{2}$$

Here, $m_{ij} \in \{0,1\}$ is the missing indicator variable, and $\bar{x}_j^{(g)}$ is the statistical compensation value of the JTH feature under the corresponding group g. This process can preserve group differences as much as possible and reduce the structural deviation caused by large-scale mean imputation.

After the repair of continuous variables is completed, the problem of dimension inconsistency still needs to be solved. The numerical ranges of sleep duration, exercise frequency, scale scores and classroom behavior indicators are quite different. If unconstrained, the variables with large numerical spans will occupy unreasonable weights in the optimization process. Therefore, min-max normalization is applied to continuous features in this paper:

$$x_{ij}^{*} = \frac{\tilde{x}_{ij} - \min(\tilde{x}_j)}{\max(\tilde{x}_j) - \min(\tilde{x}_j) + \epsilon} \tag{3}$$

where ϵ is a tiny constant that prevents the denominator from being zero. After standardization, the features of each dimension are mapped to a unified range, which is conducive to subsequent multimodal fusion and classification boundary learning. The extreme values are constrained by a truncation strategy based on the interquartile range to weaken the disturbance caused by abnormal filling or equipment errors.

The preprocessing of text data takes a different technical path than that of structured data. Considering that rural teenagers' texts are usually short in length, colloquialized, and emotional expression is implicit and not direct, this paper performs word segmentation, noise symbol cleaning, stop word screening and emotional dictionary mapping in turn on the premise of keeping negative words, adverbs of degree and emotional trigger words. Then, the term is weighted by TF-IDF and input into the text encoding module together with the sentiment polarity score. The weight of a term t in a text d is defined as:

$$w_{t,d} = tf_{t,d} \cdot \log\left(\frac{N+1}{df_t+1}\right) \tag{4}$$

where $tf_{t,d}$ is the term frequency, df_t is the number of documents containing term t, and N is the total number of texts. This representation can highlight low-frequency words with strong risk-pointing meaning, such as "can't sleep", "don't want to talk", "always want to escape", etc.

After completing numerical inpainting, scale unification and text encoding, the model further fuses structural features and text vectors into a unified input representation:

$$h_i = W_s x_i^{(s)} \oplus W_t e_i^{(t)} \tag{5}$$

Here, $x_i^{(s)}$ is the set of structured features, $e_i^{(t)}$ is the text encoding result, $\oplus$ represents the concatenation operation, W_s and W_t are linear mapping parameters. The unified representation thus formed not only contains behavioral and environmental information, but also retains the implicit risk cues in emotional expression.

The whole data acquisition and preprocessing process is shown in Figure 2. This process focuses on "multi-source collection, quality control, unified coding, and fusion representation", and integrates heterogeneous information such as individual background, learning behavior, life rules, psychological evaluation and text expression into a unified input for subsequent model training. After missing repair, anomaly correction, normalization processing and text coding, the discrete risk cues in the original samples are transformed into feature representations with clear structure and consistent scale, which provides a stable data basis for subsequent analysis and hierarchical detection of emotion representation.

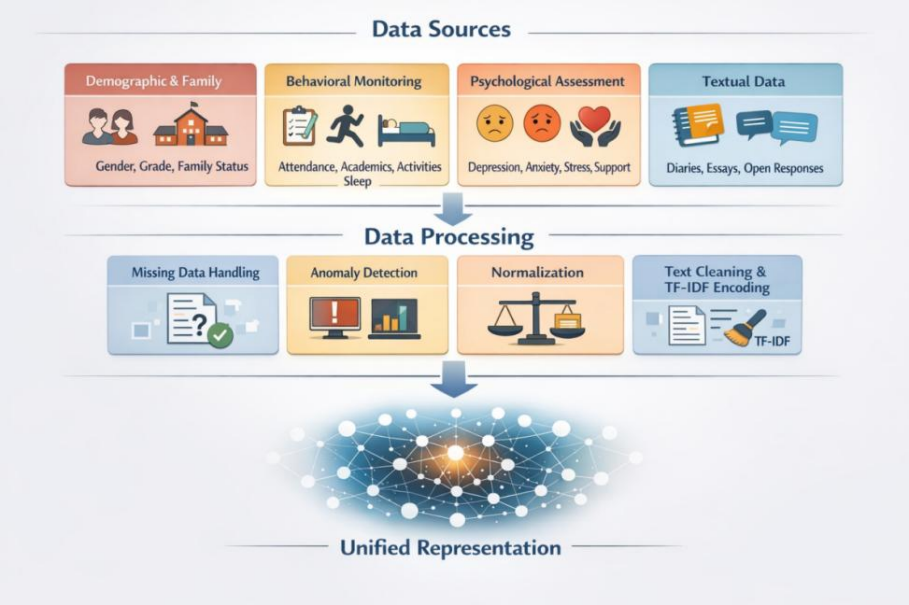

Figure 2: Structure of data collection and preprocessing for rural adolescents' mental health

3.2 Multi-source feature construction and representation learning methods

After the completion of data collection, missing repair and basic cleaning, the model inputs still show the characteristics of different sources, different dimensions, and uneven correlation

strength. Concatenating these variables directly not only amplifies the effect of high-variance features, but may also mask the linkage between sleep imbalance, low mood, and learning withdrawal. Based on this, the multi-source feature construction is divided into four steps: continuous behavior feature standardization, psychological evaluation representation compression, text emotion coding and cross-modal fusion representation, and the input vector for early risk detection is formed in a unified representation space. Its core structure is shown in Figure 3.

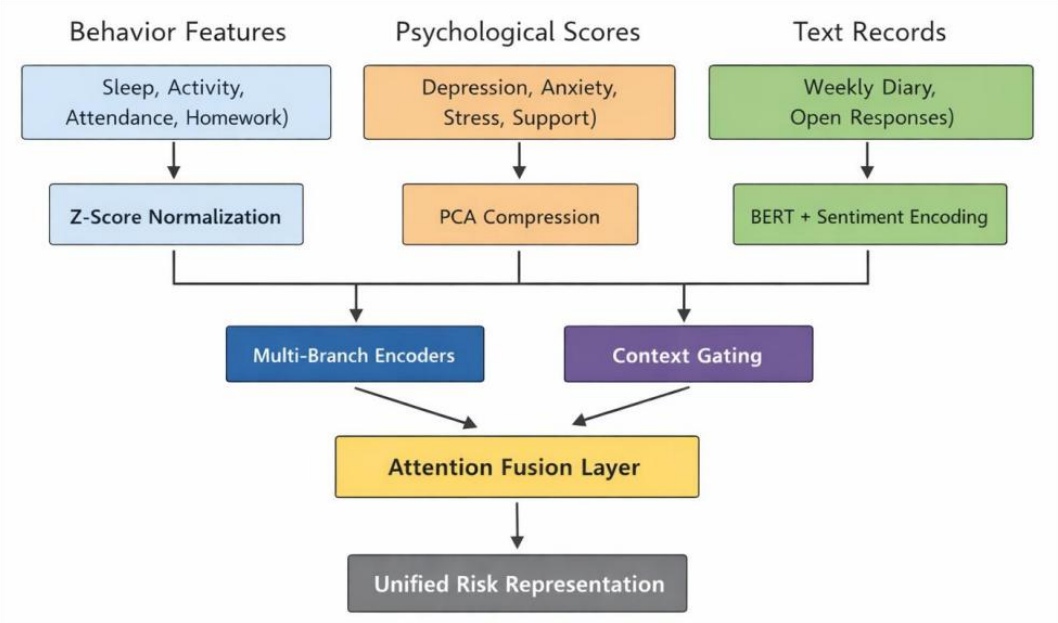

Figure 3: Schematic diagram of multi-source feature construction and representation learning method

a. Standardization of continuous behavioral characteristics

Although sleep duration, physical activity frequency, classroom participation and assignment completion rate were all related to mental state, their numerical ranges were quite different. In order to avoid a certain class of high-amplitude indicators dominating the training process, Z-score normalization is used to deal with continuous behavioral variables in this paper. For the JTH feature, the normalized result is denoted as:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j + \epsilon} \tag{6}$$

Here, x_{ij} is the observed value of the ith sample on the JTH feature, μ_j and σ_j represent the mean and standard deviation of the feature, and ϵ is the minimum constant. After this processing, behavioral data from different sources are mapped to comparable statistical scales, which is conducive to the stable identification of subsequent abnormal patterns.

b. Compressed representation of psychological assessment features

Scale variables such as depression, anxiety, stress and social support often have strong statistical correlation, and direct input into the model is easy to cause redundant accumulation and local collinear. In this paper, the psychological assessment vector is denoted as p_i, and then it is compressed by principal components to obtain a low-dimensional representation:

$$u_i = W_{pca}^{\top} p_i \tag{7}$$

And choose to satisfy:

$$\frac{\sum_{k=1}^{K}\lambda_k}{\sum_{k=1}^{m}\lambda_k} \geq 0.95 \tag{8}$$

The first K principal components of serve as compact representations of mental states. Here, λ_k is the KTH characteristic root. This process not only retains the main psychological fluctuation information, but also reduces the interference of high correlation dimensions on classification boundaries.

c. Text emotion feature representation

Early risk for rural adolescents is not always directly apparent through scale scores, and some of the more subtle changes are often hidden in weekly notes, self-statements, and open-ended texts. To this end, we use pre-trained language models to extract contextual semantic vectors and combine them with sentiment polarity scores to form text representations:

$$t_i = \left[\mathrm{BERT}(d_i)\,; s_i^{pos}, s_i^{neg}, s_i^{neu}\right] \tag{9}$$

Here, di represents the text input of the ith sample, and $s_i^{pos}, s_i^{neg}, s_i^{neu}$ represent positive, negative, and neutral sentiment intensities, respectively. Such a representation can capture low-explicit emotional cues such as "fatigue", "depression" and "reluctance to communicate", and make up for the lack of emotional details in structured data.

d. Cross-modal attention fusion representation

The contribution of behavioral branch, psychological branch and textual branch in risk discrimination is not constant. For some samples, the plunge in sleep was more distinct from the decrease in activity; For other samples, negative text and high stress scale are more revealing of potential risks. To this end, this paper uses the attention mechanism to adaptively weight the multi-branch representation. Let the MTH modal branch output be $h_i^{(m)}$, then its weight is defined as:

$$\alpha_i^{(m)} = \frac{\exp\left(q^{\top}\tanh(W_m h_i^{(m)})\right)}{\sum_r \exp\left(q^{\top}\tanh(W_r h_i^{(r)})\right)}, h_i = \sum_m \alpha_i^{(m)} h_i^{(m)} \tag{10}$$

The unified representation h_i can integrate behavioral, psychological and textual information in the same space, which provides a more discriminative input basis for subsequent emotion representation analysis and hierarchical risk detection. On the whole, this representation learning process is not simply to expand the number of features, but to improve the ability of the model to capture the weak signal of psychological risk in the early stage of rural adolescents under the premise of controlling redundancy and noise.

3.3 Emotion representation analysis and risk signal embedding

In the early detection of rural adolescent mental health, text is not ancillary information. Compared with the scale scores and behavior records, the self-report text often exposes the implicit changes such as emotional withdrawal, social withdrawal and cognitive depression earlier. Some students may not show significant abnormalities in the questionnaire, but in the weekly notes, messages or open questions and answers, they have appeared "always feel tired", "don't want to talk to others", "always wake up at night" and other expressions with

persistent risks. Therefore, this paper separates the emotion representation from ordinary text features, and designs a processing chain of "semantic encoder-polarity estimation -gated fusion -risk embedding" to enhance the model's ability to perceive early weak signals. The process is shown in Figure 4.

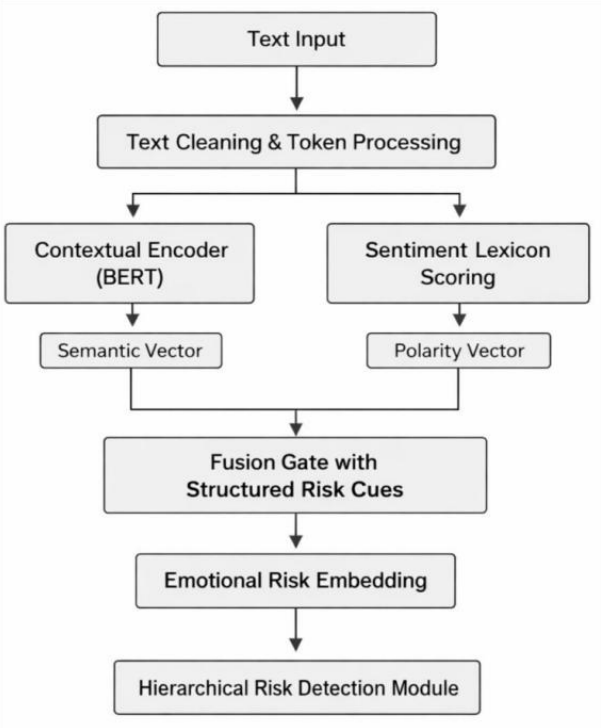

Figure 4: Flowchart of emotion representation analysis and risk signal embedding

Let the text of the ith sample be d_i, whose contextual semantic representation is extracted by the pre-trained encoder as:

$$c_i = \mathrm{BERT}(d_i) \tag{11}$$

c_i preserves the emotional orientation and semantic dependency of words in the specific context. In order to avoid the masking of explicit emotion polarity by deep semantic representation, we also introduce lexicon-level emotion score and construct polarity vector:

$$l_i = \left[p_i^{pos}, p_i^{neg}, p_i^{neu}, p_i^{comp}\right] \tag{12}$$

Among them, $p_i^{pos}, p_i^{neg}, p_i^{neu}$ and p_i^{comp} represent positive, negative, neutral and composite polarity intensities, respectively. After this process, the explicit emotions such as "fatigue", "depression" and "irritability" in the text can be preserved at the same time as the more implicit negative semantic tendency.

Text alone is still not enough to judge risk intensity, because the same negative expression has different meanings in different life contexts. To this end, this paper introduces the structured risk cue s_i as a gating condition to adaptively regulate the emotion semantics. The fused representation is defined as:

$$e_i = \tanh(W_c c_i + W_l l_i + b) \tag{13}$$

The gating weight is further written as:

$$g_i = \sigma\left(W_g[e_i; s_i] + b_g\right), r_i = g_i \odot e_i, \tag{14}$$

Here, $[\cdot;\cdot]$ represents vector concatenation, $\odot$ represents element-wise multiplication, and r_i is the final emotional risk embedding. The significance of this process is that when there is a mild negative expression in the text, and the sample is accompanied by decreased sleep, weakened classroom participation or low social support, the gated unit will automatically amplify the corresponding emotional dimension. On the contrary, if the negative words of the text are only short-term emotional fluctuations, the risk embedding will not be over-strengthened.

Through the above modeling, the originally scattered language cues are transformed into discriminative continuous vectors, and enter the subsequent hierarchical detection module together with the aforementioned behavioral features and psychological evaluation features. In this way, the model is not just making a judgment based on whether the high score is depressed or not, but is able to catch earlier signs of risk before the semantic expression, life order, and behavior withdrawal have completely deteriorated in sync.

3.4 Design of early psychological risk stratification detection process

After completing the multi-source feature construction and emotional risk embedding, the model needs to further answer a question closer to real applications: how to identify rural adolescents that need to be focused on as early as possible under the condition of unbalanced class distribution. If the flat triple classifier is used directly, the number of low-risk samples is dominant, and the discrimination boundary between the risk and high-risk categories is easy to be compressed, which reduces the early warning ability. Based on this, this paper designs a two-stage hierarchical detection process, which separates "whether there is a significant risk" and "how to subdivide the risk degree", so as to improve the sensitivity of the model to a few high-risk samples and take into account the overall classification stability. The process structure is shown in Figure 5.

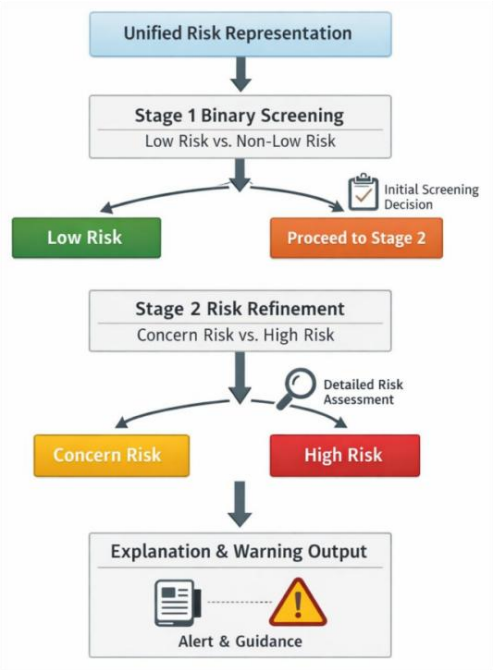

Figure 5: Schematic diagram of the early psychological risk stratification detection process

a. Hierarchical input representation

The unified risk representation output by the aforementioned module is denoted as r_i. In order to preserve the interactive relationship between behavioral features, text emotion and psychological evaluation, this paper does not perform manual rule segmentation, but directly input them into the hierarchical detector. The nonlinear mapping of stage one is defined as:

$$p_i^{(1)} = \sigma(W_1 r_i + b_1) \tag{15}$$

where $p_i^{(1)}$ is the probability that the sample is "not low risk" and $\sigma(\cdot)$ is the Sigmoid function. The goal of this stage is not to complete all the risk determination, but to first separate the most stable samples from the subsequent fine classification.

b. Stage 1 binary screening

In view of the high proportion of low risk in rural school samples, a binary screen was used to complete the "low risk-non-low risk" discrimination in the first stage. The optimization objective is written as:

$$\mathcal{L}_{bin} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log p_i^{(1)} + (1-y_i)\log\left(1-p_i^{(1)}\right)\right] \tag{16}$$

Here, y_i=1 indicates that the sample belongs to the non-low risk group. After this stage is completed, the model only sends the samples that pass the screening to the next layer, thus weakening the squeezing effect of the majority class on the classification boundary.

c. Stage II Risk refinement

For the samples that are judged as non-low risk, the model further distinguishes between "concerned risk" and "high risk". Let the second-stage output probability vector be $p_i^{(2)}$, then we have:

$$p_i^{(2)} = \mathrm{softmax}(W_2 r_i + b_2) \tag{17}$$

And it is trained by category weighted cross entropy:

$$\mathcal{L}_{ref} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{c=1}^{2}\omega_c\, y_{ic} \log p_{ic}^{(2)} \tag{18}$$

Here, ω_c is the category weight, which is used to improve the influence strength of high-risk samples in training. In this way, the model is able to learn a clearer local boundary between the risk of interest and the high risk, without being disturbed by the large presence of low risk samples.

d. Joint decision making and early warning output

The total objective function of the whole process is written as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{bin} + \lambda_2 \mathcal{L}_{ref} \tag{19}$$

Here, λ_1 and λ_2 are used to balance the two-stage learning task. In the prediction, the model completed the hierarchical decision according to the output probability of stage 1 and stage 2. If the stage 1 was judged as low risk, the low risk label would be output directly. If it enters the second stage, the attention risk or high risk will be output according to the fine classification results, and the corresponding explanation information and warning level will

be generated simultaneously. Such a process design makes the system not only technically complete the three classifications, but also form a continuous chain of "initial screening, subdivision, explanation and warning" in a way close to the actual disposal logic of the school, so as to enhance the availability and response efficiency of early psychological risk detection.

3.5 Model evaluation index system

In order to test the effectiveness of the constructed model in the early detection of rural adolescents' mental health, this paper established an evaluation index system from four levels: overall discrimination accuracy, class equilibrium recognition ability, threshold discrimination ability and high-risk sample capture ability. Considering that the samples are unevenly distributed among low risk, concern risk and high risk, if the overall accuracy is only used as the basis for judgment, it is easy to cover up the insufficient recognition of a few high-risk individuals by the model. Therefore, in this study, Accuracy, Macro-F1, AUROC, and High-Risk Recall are jointly used as core indicators, and the model output is characterized by a confusion matrix in a unified way. The evaluation framework is shown in Figure 6.

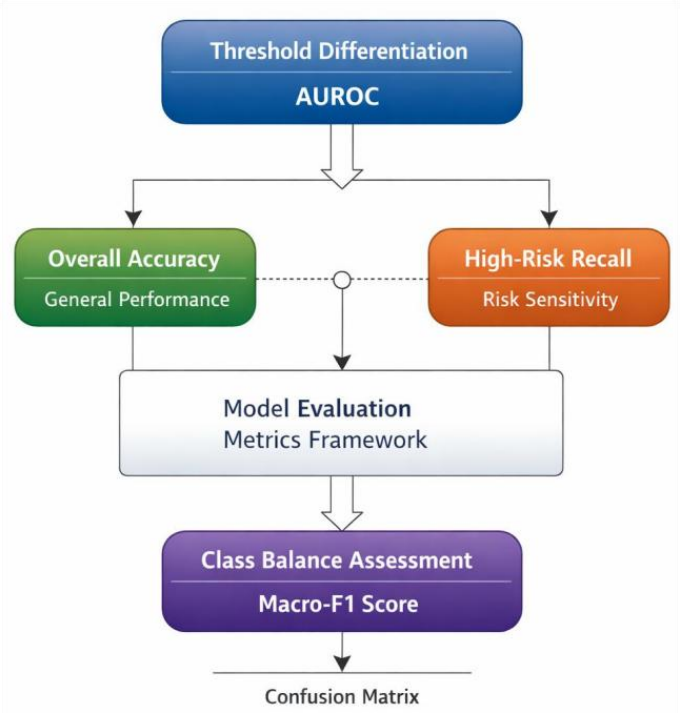

Figure 6: Schematic diagram of the model evaluation index system

Let the multiclass confusion matrix be $C = [c_{ij}]$, where c_{ij} denotes the number of samples with true class i and predicted class j, then the overall accuracy is defined as:

$$Acc = \frac{\sum_{k=1}^{K} c_{kk}}{\sum_{i=1}^{K} \sum_{j=1}^{K} c_{ij}} \tag{20}$$

where K is equal to 3. This index reflects the overall discrimination level of the model for all samples, which is suitable to measure the basic stability of the system in large-scale screening scenarios, but it cannot represent the early warning ability alone.

In order to avoid excessive dominance of the majority class samples, the macro average F1 value is used to evaluate the class balance recognition effect. For the KTH class, the

precision and recall are denoted as P_k and R_k, respectively, then:

$$\text{Macro} - \text{F1} = \frac{1}{K}\sum_{k=1}^{K}\frac{2P_kR_k}{P_k + R_k} \tag{21}$$

This index can simultaneously examine the classification quality of the model on three types of risk levels, especially for the education monitoring data with a large proportion of low-risk samples. If the model is only good at identifying low risks and ignores risks and high risks, Macro-F1 will decrease significantly.

The threshold discrimination ability was evaluated by AUROC. For any threshold t, the true positive rate is denoted as TPR(t) and the false positive rate is denoted as FPR(t), then:

$$\text{AUROC} = \int_0^1 \text{T PR}(\text{FPR}^{-1}(u))\,du \tag{22}$$

AUROC does not rely on a fixed classification threshold, and is more suitable for evaluating the robustness of the model under different warning intensity Settings. For mental health screening, this is particularly important, because different schools differ in teachers and intervention resources, and the setting of warning thresholds is not completely consistent.

Considering that high-risk individuals are prioritized for early intervention, this paper further reports high-risk recall separately. Let the high-risk category be denoted as h, then:

$$\text{Recall}_h = \frac{c_{hh}}{\sum_{j=1}^{K} c_{hj}} \tag{23}$$

This metric directly reflects the ability of the model to capture real high-risk samples. For psychological monitoring of rural adolescents, the cost of missing high-risk individuals is usually higher than misjudging general samples, so Recallh has a stronger application direction than overall accuracy in this study. Taken together, the above indicators constitute an evaluation system for early risk detection, which makes the model performance judgment no longer stay in a single numerical comparison, and can more truly reflect its usability in actual education warning scenarios.

4 Analysis of results

4.1 Explanation of data source and sample information

The experimental data in this paper are from the anonymous student sample database of rural middle schools in several counties. The collection objects are adolescents aged 12-18 years old in school, and a total of 1286 valid records are collected. The data sources covered five dimensions of students' basic information, learning behavior, living habits, psychological assessment and text self-report, which not only retained the family and school context factors in the formation of rural adolescents' psychological risk, but also provided computational input for subsequent machine learning modeling. Each record corresponded to a comprehensive portrait of a student in a continuous observation period. The structured fields included age, gender, whether to stay behind, sleep duration, physical activity frequency, classroom participation, homework completion rate, stress score, anxiety score, depression score and social support score, and the unstructured fields mainly included weekly self-report text and brief emotional description.

In the sample composition, male students accounted for 51.2% and female students accounted for 48.8%. The left-behind samples accounted for 37.6%. The risk labels were jointly labeled according to the scale scores, school observation records and text emotion results, and were divided into three categories: low risk, concerned risk and high risk, with the sample proportions of 55.4%, 28.7% and 15.9%, respectively. This hierarchical labeling makes the data suitable for both regular supervised classification and validation of the proposed two-stage hierarchical detection process. On the whole, the data set contains behavioral signals, emotional signals and environmental signals, which can completely present the dynamic differences of rural adolescents 'psychological states. The main fields are described in Table 2.

Table 2: Description of main attributes and meanings of data

Attribute Name	Description
Student_ID	Anonymous student identifier
Age	Age
Gender	Gender
Left_Behind_Status	Whether the student is left behind
Sleep_Hours	Average daily sleep duration
Activity_Frequency	Weekly frequency of physical activity
Class_Participation	Classroom participation level
Homework_Completion	Homework completion rate
Stress_Score	Stress assessment score
Anxiety_Score	Anxiety assessment score
Depression_Score	Depression assessment score
Social_Support	Social support score
Weekly_Reflection	Weekly reflection / open-ended self-report text
Mood_Description	Brief mood description
Mental_Health_Risk	Target label (Low-risk / Concern-risk / High-risk)

4.2 Model implementation and experimental environment setup

The model is implemented, trained and evaluated in a unified computing environment. The experimental process includes structured data preprocessing, text encoding, multi-source feature fusion, hierarchical classification training and interpretation analysis. Structural variables were repaired for missing values, cut off outliers and standardized before input. After word segmentation, noise cleaning and stop word filtering, the context semantic vector and sentiment polarity score are extracted from the text data, and the model input is composed of behavior, scale and social support features. In order to ensure a fair comparison, the baseline model and the proposed model adopt the same data partition and preprocessing aperture.

At the implementation level, the traditional baselines included Random Forest, XGBoost and single-stage DNN. In the first stage, low risk and non-low risk samples were screened. In the second stage, non-low risk samples were further divided into concern risk and high risk. In the training process, five-fold cross validation with stratified sampling is used, and class weight is introduced to alleviate the impact of uneven distribution of samples on high-risk identification. The machine learning module mainly relies on scikit-learn, PyTorch and Transformers. The text processing calls jieba, NLTK and regular cleaning tools. The statistics and plots are supported by NumPy, Pandas, and Matplotlib. The overall experimental platform can handle structured variables and text vector input at the same time, and meet the needs of

repeated training under multi-round cross validation. See Table 3 for the specific implementation environment.

Table 3: Model implementation environment and main configurations

Category	Configuration
Programming Language	Python 3.11
Operating System	Ubuntu 22.04
Development Environment	Jupyter Notebook
Data Processing	Pandas, NumPy
Machine Learning Libraries	scikit-learn, PyTorch
NLP Libraries	Transformers, NLTK, jieba
Visualization Tools	Matplotlib
CPU	Intel Core i7-12700
RAM	32 GB
GPU	NVIDIA RTX 4080 16 GB
Validation Strategy	Stratified 5-fold Cross-Validation
Baseline Models	Random Forest, XGBoost, Single-stage DNN
Proposed Model	Hierarchical Multimodal Risk Detection Model

5 Performance analysis

5.1 Analysis of model classification performance and generalization ability

In order to test the classification ability and cross-fold stability of the proposed model in the early detection task of rural adolescents' mental health, this paper compared it with three representative baseline methods such as Random Forest, XGBoost and single-stage DNN. All models were trained and tested under the same data partition, the same preprocessing process and the same evaluation index system, and the mean and standard deviation were reported by stratified five-fold cross validation. The comparison results show that the model in this paper achieves the best performance in four core indicators of Accuracy, Macro-F1, AUROC and High-Risk Recall, indicating that the multi-source feature fusion and two-stage hierarchical detection strategy can effectively improve the category confusion problem in the psychological risk identification of rural adolescents.

From the overall classification results, the Accuracy of the proposed model reaches 91.40%±0.47, which is higher than that of Random Forest (82.10%±0.77), XGBoost (85.74%±0.59) and single-stage DNN (87.94%±0.55). Macro-F1 reaches 0.891±0.008, which is 0.077 higher than XGBoost and 0.044 higher than single-stage DNN. This means that the model does not only have an advantage on the low-risk samples, but maintains a more balanced discrimination ability across the three risk classes. Table 4 summarizes the comprehensive performance of each model. It can be seen that the traditional tree model has certain stability when dealing with structured variables, but it is insufficient to describe the implicit association between text emotion and behavior withdrawal. Although single-stage DNN makes good use of fusion representation, it is still affected by the uneven distribution of classes, and there is a certain miss judgment on high-risk samples.

Table 4: Classification performance comparison of different models under 5-fold cross validation

Model	Accuracy (%)	Macro-F1	AUROC	High-Risk Recall
Random Forest	82.10 ± 0.77	0.763 ± 0.011	0.845 ± 0.007	0.709 ± 0.016
XGBoost	85.74 ± 0.59	0.814 ± 0.010	0.883 ± 0.007	0.761 ± 0.012
Single-stage DNN	87.94 ± 0.55	0.847 ± 0.009	0.904 ± 0.006	0.798 ± 0.011
Proposed Model	91.40 ± 0.47	0.891 ± 0.008	0.931 ± 0.006	0.860 ± 0.012

The boost in high-stakes recall is particularly noteworthy. The model in this paper reaches 0.860±0.012 on High-Risk Recall, which is 0.151 higher than 0.709±0.016 of Random Forest, 0.099 and 0.062 higher than XGBoost and single-stage DNN, respectively. For the early detection of mental health, this result has direct application, because the high-risk category itself is the priority object for early warning and follow-up intervention in schools. If the model can only maintain a high overall accuracy, but cannot stably identify the samples that really need to be focused on, its practical value is still limited. Figure 7 further shows the comparison of each model on the four core indicators. It can be seen that the proposed model keeps ahead in the four dimensions, and the advantage is not concentrated on a single indicator, but reflected in the synchronous improvement of the overall performance.

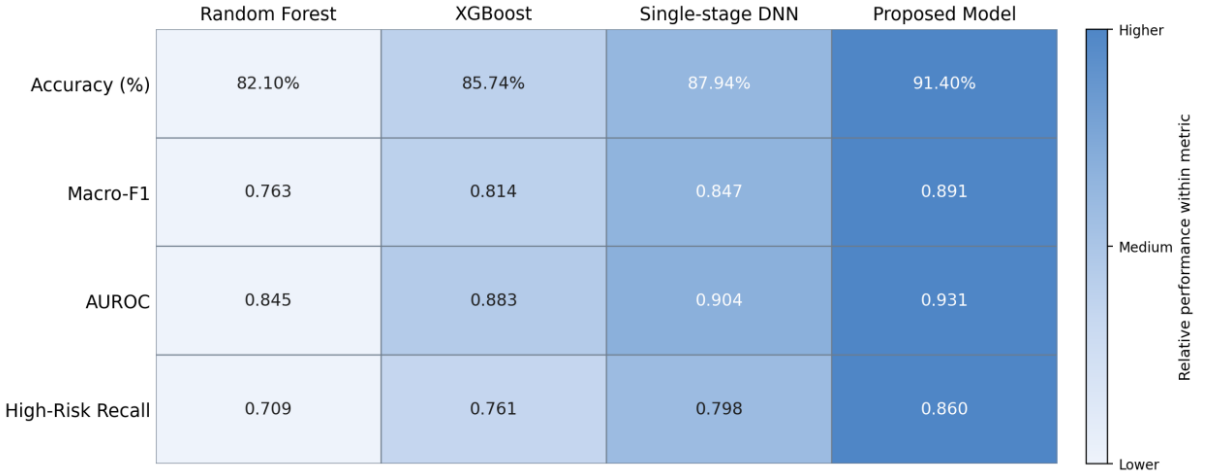

Figure 7: Comparison of core indicators of different models

From the generalization ability, Figure 8 shows the variation trend of Macro-F1 of each model under 5-fold cross validation. The Macro-F1 of the five folds of the model in this paper are 0.881, 0.892, 0.901, 0.887 and 0.894, respectively, with a small fluctuation range, indicating that the model maintains high consistency under different training-testing divisions. In contrast, Random Forest and XGBoost have more obvious fluctuations in partial folds, reflecting that they are more sensitive to the change of sample distribution. The overall stability of the single-stage DNN is better than the traditional baseline, but still worse than the proposed method. This shows that the hierarchical detection process adopted in this paper can make the second-stage classifier more focused on learning the local boundary between "risk-high risk" after filtering the low-risk samples, so as to maintain a more stable risk discrimination ability in different data folds.

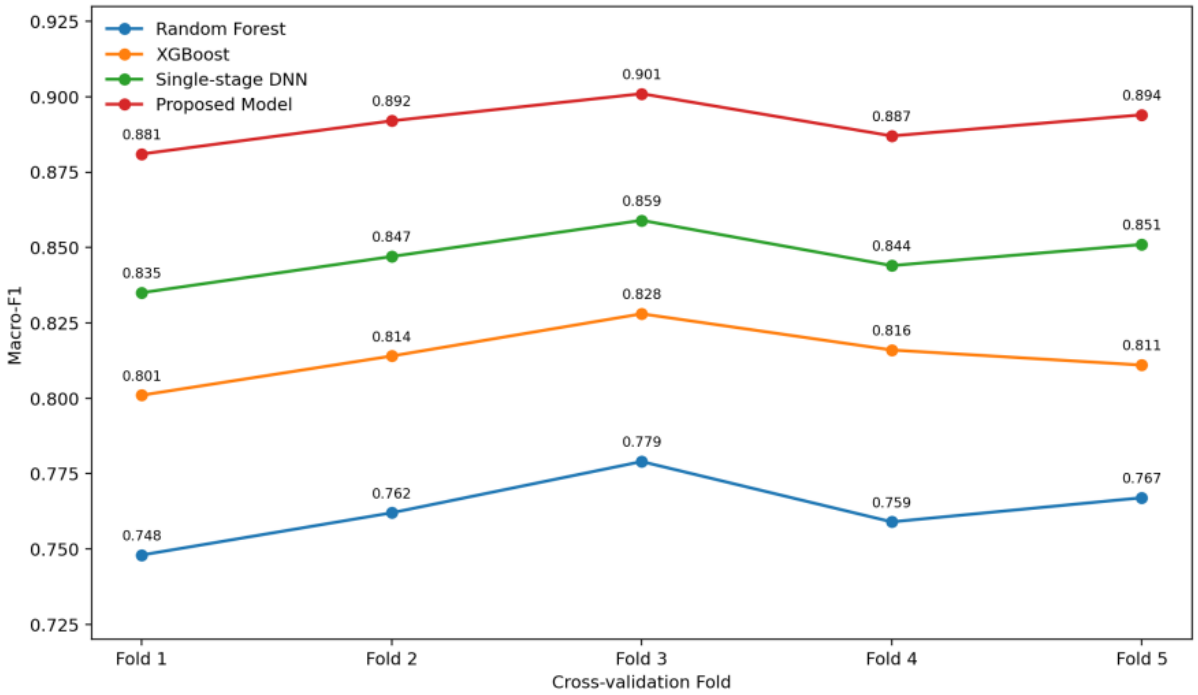

Figure 8: Macro-F1 changes of each model under 5-fold cross validation

It can be seen from Table 4, Figure 7 and Figure 8 that the performance advantage of the proposed model does not come from the accidental single division result, but from the joint modeling of multi-source behavior features, psychological evaluation features and text emotion representation, as well as the targeted treatment of class imbalance problem by the hierarchical classification mechanism. In other words, our model not only outperforms baseline methods in overall accuracy, but also shows stronger robustness in high-risk identification and cross-fold generalization, which lays a foundation for subsequent category-level interpretation analysis.

5.2 SHAP based analysis of class-level feature interpretation

In the task of early detection of rural adolescent mental health, if the model output only stays at the risk label level, its application value is still limited. What the school scenario really needs is not only "who is judged as high risk", but also "what factors are driving the risk" and "what signals have been superimposed". Based on this consideration, after the performance verification, this paper analyzes the SHAP interpretation of the proposed model, and discusses the global feature contribution and individual local decision respectively. The explanatory objects cover key variables such as depression, anxiety, sleep, classroom participation, social support, and negative emotion of the text, so that the discriminant logic of the model can be directly understood by educational managers and psychology teachers.

Figure 9 illustrates the distribution of the mean absolute SHAP values of the main features under the three classes of risk labels. It can be seen that the explanatory structure of the low-risk category is dominated by the "stability" variable. The mean absolute SHAP values for depression score, anxiety score, sleep duration and negative emotion expression are 0.041, 0.036, 0.033 and 0.027, respectively, which means that when the above variables are in a relatively stable interval, The model is more likely to give low-risk judgments. After entering the concern risk category, the importance of depression score, anxiety score and sleep duration increased significantly, and the corresponding mean absolute SHAP values increased

to 0.066, 0.059 and 0.051, respectively, indicating that the model has begun to capture the linkage trend between emotional distress and life rhythm fluctuations. The mean absolute SHAP values of depression score, anxiety score, sleep duration and negative text emotion reached 0.104, 0.093, 0.081 and 0.072, respectively. The decreased classroom participation and insufficient social support also reached 0.058 and 0.052, respectively. Figure 9 indicates that the identification of high-risk samples by the model does not rely on a single scale threshold, but forms stronger risk inference when multidimensional anomalies are co-accumulated.

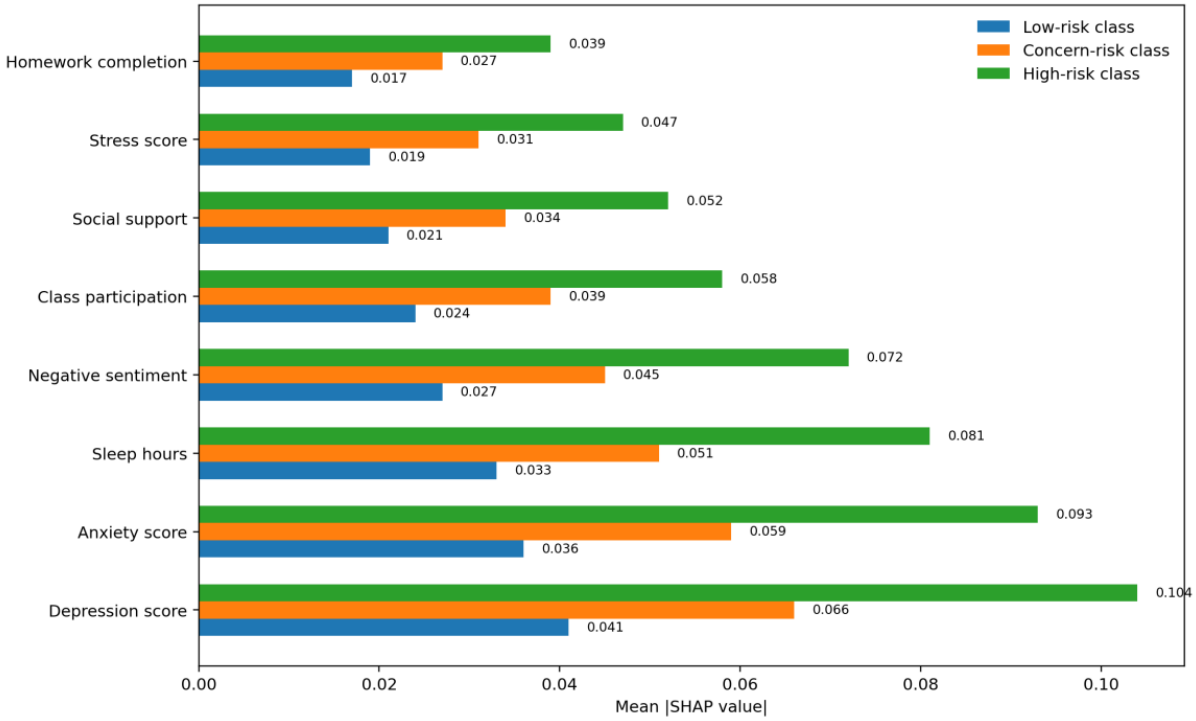

Figure 9: Category level global SHAP feature importance map

For comparison purposes, Table 5 summarizes the core variables with the most explanatory power among the three categories of risk labels. The low-risk category was mainly supported by lower depression, anxiety and more stable sleep; Attention to the risk category showed the characteristics of the transition from mild emotional fluctuations to persistent risks. In the high-risk category, depression, anxiety, negative emotional expression, and lack of sleep together constitute the main driving force. Such class-level differences suggest that the model does not simply compress all samples into the same "anomaly pattern", but instead learns different driving structures at different risk stages.

Table 5: Results of SHAP interpretation of category-level core features

Risk Class	Key Feature	Mean |SHAP|	Interpretation
Low-risk	Depression score	0.041	A low depression level helps maintain a stable psychological state.
Low-risk	Anxiety score	0.036	Lower anxiety supports a low-risk classification.
Low-risk	Sleep hours	0.033	Regular sleep is closely associated with psychological stability.
Concern-risk	Depression score	0.066	Emotional distress begins to play a clear role in increasing risk.
Concern-risk	Anxiety score	0.059	Rising anxiety is an important signal of concern-level risk.
Concern-risk	Sleep hours	0.051	Reduced sleep strengthens the judgment of concern-level risk.
High-risk	Depression score	0.104	It is the primary variable driving high-risk predictions.
High-risk	Anxiety score	0.093	It shows a stable and strong association with the high-risk label.
High-risk	Negative sentiment	0.072	Negative textual expression significantly strengthens high-risk predictions.

Figure 10 further presents the local SHAP interpretation results for a high-risk sample. The baseline prediction for this sample was 0.312, which was pushed up to 1.044 after a combination of adverse factors. Specifically, a high depression score contributed +0.184, a high anxiety score contributed +0.146, a low sleep duration contributed +0.121, and a negative weekly text contributed +0.108. Insufficient classroom participation, low social support and high stress scores brought +0.081, +0.067 and +0.053 increments respectively. Only "relatively stable activity frequency" produced a slight cancellation of -0.028. Figure 10 clearly shows that the model's judgment of high-risk individuals is not triggered by a single indicator, but is advanced by a combination of deteriorating mood, disrupted life rhythms, decreased school participation, and weak support systems.

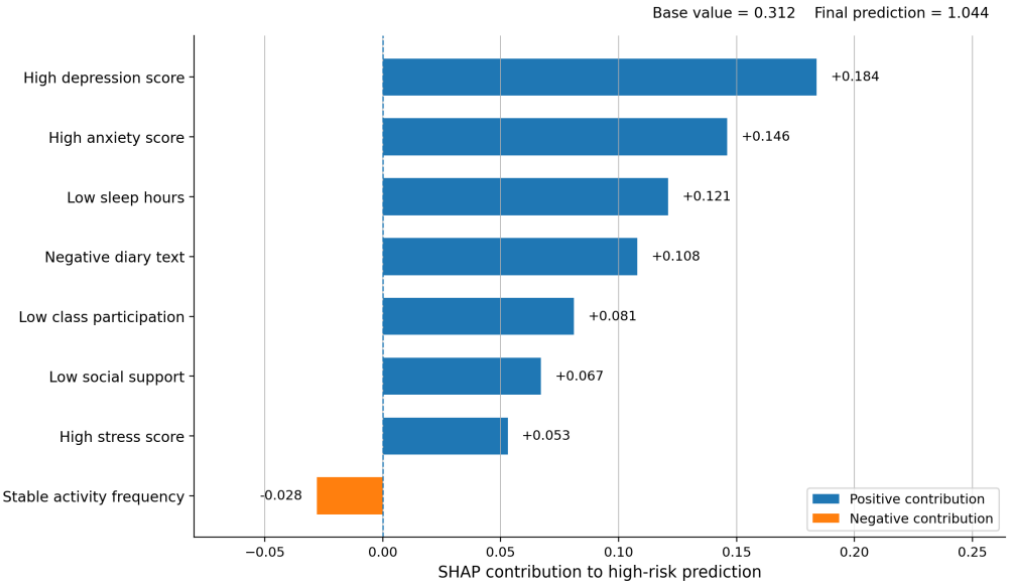

Figure 10: Local SHAP contribution diagram of high-risk samples

6 Conclusion and future improvement direction

Focusing on the task of early detection of mental health risk among rural adolescents in China, this paper constructs an advanced machine learning model that integrates behavioral characteristics, psychological assessment information and text emotional cues. In this study, multi-source feature construction, emotion representation analysis, risk signal embedding and two-stage hierarchical detection are completed under a unified data preprocessing framework, and the SHAP explanation mechanism is used to reveal the key driving factors of different risk categories. Experimental results show that the proposed model is superior to Random Forest, XGBoost and single-stage DNN in terms of Accuracy, Macro-F1, AUROC and high-risk recall rate, and shows stronger sensitivity and stability especially in the identification of high-risk samples. This indicated that integrating sleep, classroom participation, social support, depression and anxiety scores and negative text expression into the unified representation space could more effectively capture the early subtle signals in the formation of psychological risk in rural adolescents.

From the perspective of method, the value of this paper is not only to improve the classification accuracy, but also to make the risk identification process interpretable to a certain extent. The output of the model is no longer just an abstract label, but can further explain what factors are mainly driven by the increased risk, so as to provide more targeted basis for school psychology teachers, class teachers and family-school collaborative intervention. This computing framework from "detection" to "auxiliary judgment" is more in line with the practicability and landing requirements of rural education scenes.

There are several future directions worth pursuing. On the one hand, the current cross-sectional modeling can be extended to the longitudinal time series modeling, and the state change information under continuous observation periods can be introduced to improve the ability to describe the risk evolution process. On the other hand, the federated learning or parameter isolation training mechanism can be explored under the premise of ensuring privacy security, so that different schools can complete collaborative modeling without directly sharing the original data. In addition, if fine-grained behavior trajectories, speech cues or wearable device data can be further connected, and lightweight deployment and threshold adaptive calibration strategies can be combined, the generalization ability and application value of the model in the real campus environment still have room to be improved.

Funding

This research was supported by 2024 Annual General Research Project of Hubei Small Town Development Research Center"A Study on the Differences and Countermeasures of Mental Health Status between Urban and Rural College Students in Hubei Province"(Grant No.2024B001).

References

[1] Thapar A, Eyre O, Patel V, et al. Depression in young people[J]. The Lancet, 2022, 400(10352): 617-631.

[2] Liu D, Feng X L, Ahmed F, et al. Detecting and measuring depression on social media using a machine learning approach: systematic review[J]. JMIR Mental Health, 2022, 9(3): e27244.

[3] Mendes J P M, Moura I R, Van de Ven P, et al. Sensing apps and public data sets for digital phenotyping of mental health: systematic review[J]. Journal of medical Internet research, 2022, 24(2): e28735.

[4] Bufano P, Laurino M, Said S, et al. Digital phenotyping for monitoring mental disorders: systematic review[J]. Journal of medical Internet research, 2023, 25: e46778.

[5] Andrew J, Rudra M, Eunice J, et al. Artificial intelligence in adolescents mental health disorder diagnosis, prognosis, and treatment[J]. Frontiers in Public Health, 2023, 11: 1110088.

[6] Huang Y, Zhu C, Feng Y, et al. Comparison of three machine learning models to predict suicidal ideation and depression among Chinese adolescents: a cross-sectional study[J]. Journal of affective disorders, 2022, 319: 221-228.

[7] Zhou Y, Zhang X, Gong J, et al. Identifying the risk of depression in a large sample of adolescents: An artificial neural network based on random forest[J]. Journal of Adolescence, 2024, 96(7): 1485-1497.

[8] Kim H, Son Y, Lee H, et al. Machine learning–based prediction of suicidal thinking in adolescents by derivation and validation in 3 independent worldwide cohorts: Algorithm development and validation study[J]. Journal of medical internet research, 2024, 26: e55913.

[9] Li Q, Song K, Feng T, et al. Machine learning identifies different related factors associated with depression and suicidal ideation in Chinese children and adolescents[J]. Journal of affective disorders, 2024, 361: 24-35.

[10] Zhong Y, He J, Luo J, et al. A machine learning algorithm-based model for predicting the risk of non-suicidal self-injury among adolescents in western China: a multicentre cross-sectional study[J]. Journal of affective disorders, 2024, 345: 369-377.

[11] Yang C, Huebner E S, Tian L. Prediction of suicidal ideation among preadolescent children with machine learning models: a longitudinal study[J]. Journal of affective disorders, 2024, 352: 403-409.

[12] Yoo A, Li F, Youn J, et al. Prediction of adolescent depression from prenatal and childhood data from ALSPAC using machine learning[J]. Scientific Reports, 2024, 14(1): 23282.

[13] Luo Y, Wang Y, Wang Y, et al. Development and validation of a nomogram for predicting suicidal ideation among rural adolescents in China[J]. Psychology Research and Behavior Management, 2024: 4413-4429.

[14] Jiang Q, She X, Dill S E, et al. Depressive and anxiety symptoms among children and adolescents in rural China: a large-scale epidemiological study[J]. International journal of environmental research and public health, 2022, 19(9): 5026.

[15] She X, Zhao D, Li M. Adolescent mental health disparities in rural Guizhou vs. urban Beijing: A comparative analysis from China[J]. Global Pediatrics, 2022, 2: 100023.

[16] Zhang P, Yang F, Huang N, et al. Assessment of factors associated with mental well-being among Chinese youths at individual, school, and province levels[J]. JAMA network open, 2023, 6(7): e2324025.

[17] Jing Z, Ding F, Sun Y, et al. Comparing depression prevalence and associated symptoms with intolerance of uncertainty among Chinese urban and rural adolescents: a network analysis[J]. Behavioral Sciences, 2023, 13(8): 662.

[18] Zhao J C, Wang X, Xu S, et al. The influence of lifestyle habits on levels of depression among rural middle school students in Northeastern China[J]. Frontiers in public health, 2024, 12: 1293445.

[19] Liu W, Guan H, Chen X, et al. Insights into adolescent sleep and mental health in rural area of Northwestern China[J]. Scientific Reports, 2024, 14(1): 31082.

[20] Ruan Q N, Shen G H, Xu S, et al. Depressive symptoms among rural left-behind children and adolescents in China: a large-scale cross-sectional study[J]. BMC Public Health, 2024, 24(1): 3160.

[21] Mei K, Zhang F, Zhang J, et al. Perceived social support mitigates the associations among household chaos and health and well-being in rural early adolescents[J]. Journal of Adolescence, 2024, 96(1): 112-123.

[22] Bao X, Guo T, Xu L, et al. Suicidal ideation in Chinese adolescents: prevalence, risk factors, and partial mediation by family support, a cross-sectional study[J]. Frontiers in Psychiatry, 2024, 15: 1427560.