Preplanned Studies

Establishment and Validation of a Risk Prediction Model for Non-Suicidal Self-Injury Among Adolescents Based on Machine Learning Methods — Jiangsu Province, China, 2023

Xin Wang[1]; Yan Wang[1]; Jiawen Tang[2]; Yang Wang[3]; Ran Zhang[3]; Xiyan Zhang[1]; Wenyi Yang[1]; Wei Du[4]; Fei Wang[3,#]; Jie Yang[1,#]

Summary

What is already known about this topic?

Non-suicidal self-injury (NSSI) has become increasingly common among adolescents, posing a significant public health concern that impacts both physical and mental well-being.

What is added by this report?

A total of 12.72% of adolescents aged 10–18 had engaged in NSSI in Jiangsu Province, China. A well-calibrated risk prediction model [AUC=0.800, 95% confidence interval (*CI*): 0.776, 0.823] identified 8 key predictors of NSSI: insomnia, emotional symptoms, cohesion of family environment, history of drinking alcohol, gender, conflict of family environment, conduct problems, and academic level.

What are the implications for public health practice?

This study underscores the importance of personalized prevention strategies for NSSI and highlights the necessity of implementing comprehensive behavioral interventions, such as providing mental health support, enhancing sleep quality, and cultivating supportive family environments.

ABSTRACT

Introduction: Non-suicidal self-injury (NSSI) has become increasingly prevalent among adolescents, representing a significant public health concern with profound impacts on both physical and mental well-being. This study aims to determine the prevalence of NSSI among adolescents in Jiangsu Province and develop a prediction model to facilitate early identification and intervention.

Methods: This study is based on the "School-based Evaluation and Response to Child Health (SEARCH)" project. A cross-sectional survey was conducted among students from 11 schools in Jiangsu Province, China in 2023. Following data cleaning, 11,427 students were included in the analysis. Machine learning methods were employed to establish a risk prediction model for NSSI among adolescents.

Results: The prevalence of NSSI among adolescents aged 10–18 was 12.72%. Eight key predictors of NSSI were identified: insomnia, emotional symptoms, cohesion of family environment, history of drinking alcohol, gender, conflict of family environment, conduct problems, and academic level. The XGBoost model demonstrated an area under the curve (AUC) of 0.800 [95% confidence interval (*CI*): 0.776, 0.823] and an accuracy (ACC) of 0.886 in the testing set.

Conclusions: This study underscores the importance of personalized prevention strategies for NSSI and highlights the necessity of implementing comprehensive behavioral interventions, including mental health support, sleep quality enhancement, and cultivation of supportive family environments.

Non-suicidal self-injury (NSSI) refers to deliberate and repeated self-harm without suicidal intent that is not socially sanctioned (excluding practices such as tattoos and body piercings). Common methods include cutting, burning, hitting, and scratching. In recent years, NSSI has become increasingly prevalent among adolescents, raising significant public health concerns due to its adverse impacts on both physical and mental well-being. NSSI is a key risk factor for suicidal behavior (*1*) and may increase the risk of subsequent psychopathological symptoms, repeated self-harm, and substance abuse issues (*2*). A systematic review reported that the prevalence rates of non-suicidal self-harm among youth in low- and middle-income countries ranged from 15.5% to 33.3% (*3*), with prevalence in China varying from 5.4% to 33.8% due to differences in research designs and evaluation

metrics (4). The onset of NSSI typically occurs in early adolescence (ages 11 to 14), peaks during mid-adolescence (ages 15 to 16), and then declines in late adolescence or early adulthood. Therefore, conducting epidemiological studies within community populations, particularly among student groups, is essential for investigating NSSI behavioral patterns and facilitating early detection and intervention for high-risk individuals. This study employs machine learning methods to develop and validate a risk prediction model for NSSI in adolescents, providing robust scientific evidence for early identification and prevention initiatives.

This study was based on the baseline cross-sectional data from the longitudinal cohort study "School-based Evaluation and Response to Child Health (SEARCH)," which was conducted in Jiangsu Province, China. A digital platform was used to assess the mental health status of students. Participants were recruited using a stratified cluster randomized sampling method from Hailing District (Taizhou City), Yixing (Wuxi City), and Sheyang Counties (Yancheng City) in Jiangsu Province at baseline from September 2022 to February 2023. A total of 11,427 adolescents participated in this study, with a response rate of 98.2%. Data collection involved three cities in Jiangsu Province, including three primary schools, five junior high schools, and three senior high schools. The process of data collection, model establishment and validation is illustrated in the flowchart presented in Supplementary Figure S1 (available at https://weekly.chinacdc.cn/).

The questionnaire comprised two sections. The first section collected sociodemographic information, including gender, academic level, regional economic level, family structure, parental marriage status, the frequency of parental quarrels, and students' academic performance ranking. The second section assessed NSSI using the Chinese version of Ottawa Self-Injury Inventory (OSI). Additional assessments included the Strengths and Difficulties Questionnaire (SDQ) for emotional and behavioral issues, the Chinese version of Family Environment Scale (FES-CV) for family environment, and Insomnia Severity Index (ISI) for insomnia status. The Cronbach's alpha for OSI, SDQ, FES-CV and ISI were 0.72, 0.78, 0.78, and 0.89, respectively, indicating substantial internal validity. Health-related variables, such as alcohol and cigarette use, were recorded as binary outcomes (yes/no). A priori evaluations of potential factors associated with NSSI were conducted based on established scientific knowledge, public health relevance, and predictors highlighted in previous research findings. The presence or absence of NSSI was used as the outcome variable, with a total of 21 variables employed as predictors. Categorical variables were compared between students with and without NSSI using chi-square tests. Shapley Additive exPlanation (SHAP) and extreme gradient boosting (XGBoost) were used to filter predictors and establish a risk prediction model for NSSI among adolescents (5). To refine the prediction model, a 5-fold cross-validation and manual fine-tuning process was utilized to identify the optimal parameters. Participants were randomly divided into a training set (1,034 students with NSSI) and a testing set (419 with NSSI) in a 7∶3 ratio. Variable importance and SHAP beeswarm plots were generated to visualize the results. Receiver operating characteristic (ROC) curves, area under the curve (AUC), accuracy (ACC), and calibration plots were used to evaluate the prediction accuracy of the model. All statistical analyses were conducted using R Statistical Software (version 4.3.3, R Development Core Team, Vienna, Austria). A P value below 0.05 was considered statistically significant.

A total of 11,427 students aged 10 to 18 participated in the survey, comprising 6,083 (53.2%) boys and 5,344 (46.8%) girls, with a mean age of 14.0±2.4 years. Among these participants, 12.72% reported engaging in NSSI (Table 1). Using the XGBoost algorithm, this study evaluated the importance of 21 predictive factors and identified optimal parameters through 5-fold cross-validation. The variable importance plot (Figure 1) reveals that the eight most influential variables in the optimal model are insomnia, emotional symptoms, cohesion of family environment, history of drinking alcohol, gender, conflict of family environment, conduct problems, and academic level. The SHAP beeswarm plot (Supplementary Figure S2, available at https://weekly.chinacdc.cn/) illustrates the predictive contribution of each factor to NSSI risk. The AUC and ACC values of the XGBoost model were 0.817 [95% confidence interval (*CI*): 0.803–0.831] and 0.882 in the training set, and 0.800 (95% *CI*: 0.776–0.823) and 0.886 in the testing set, respectively. The ROC curves are presented in Figure 2. The calibration curve for the XGBoost model shows dots closely aligned with the 45° diagonal line (Supplementary Figure S3, available at https://weekly.chinacdc.cn/), indicating strong concordance between predicted and observed values.

TABLE 1. Comparisons of characteristics between Non-NSSI students and NSSI students [*N* (%)].

Variables	Non-NSSI students (*N*=9,974)	NSSI students (*N*=1,453)	Total (*N*=11,427)	χ^2
Gender				
Male	5,417 (54.3)	666 (45.8)	6,083 (53.2)	36.59*
Female	4,557 (45.7)	787 (54.2)	5,344 (46.8)	
Academic level				
Primary school	2,939 (29.5)	270 (18.6)	3,209 (28.1)	78.19*
Middle school	3,697 (37.1)	656 (45.1)	4,353 (38.1)	
High school	3,338 (33.5)	527 (36.3)	3,865 (33.8)	
Regional economic level				
Low	3,257 (32.7)	543 (37.4)	3,800 (33.3)	151.73*
Middle	2,865 (28.7)	582 (40.1)	3,447 (30.2)	
High	3,852 (38.6)	328 (22.6)	4,180 (36.6)	
Family structure				
Core family	4,599 (46.1)	586 (40.3)	5,185 (45.4)	17.09*
Non-core family	5,375 (53.9)	867 (59.7)	6,242 (54.6)	
Parental marriage status				
Married	7,890 (79.1)	1,058 (72.8)	8,948 (78.3)	29.57*
Others	636 (6.4)	122 (8.4)	758 (6.6)	
Unknown	1,448 (14.5)	273 (18.8)	1,721 (15.1)	
Frequency of parental quarrels				
Never	4,230 (42.4)	335 (23.1)	4,565 (39.9)	371.87*
Sometimes	5,438 (54.5)	948 (65.2)	6,386 (55.9)	
Often	306 (3.1)	170 (11.7)	476 (4.2)	
Academic performance ranking				
Top 25%	3,406 (34.1)	463 (31.9)	3,869 (33.9)	97.70*
26%–50%	1,809 (18.1)	294 (20.2)	2,103 (18.4)	
51%–75%	1,118 (11.2)	205 (14.1)	1,323 (11.6)	
Bottom 25%	885 (8.9)	216 (14.9)	1,101 (9.6)	
Not disclosed	2,756 (27.6)	275 (18.9)	3,031 (26.5)	
SDQ_emotional symptoms				
Normal	9,369 (93.9)	965 (66.4)	10,334 (90.4)	1,183.43*
Marginal	284 (2.8)	146 (10.0)	430 (3.8)	
Abnormal	321 (3.2)	342 (23.5)	663 (5.8)	
SDQ_conduct problems				
Normal	8,358 (83.8)	853 (58.7)	9,211 (80.6)	598.39*
Marginal	974 (9.8)	267 (18.4)	1,241 (10.9)	
Abnormal	642 (6.4)	333 (22.9)	975 (8.5)	
SDQ_hyperactivity				
Normal	8,859 (88.8)	877 (60.4)	9,736 (85.2)	896.61*
Marginal	588 (5.9)	210 (14.5)	798 (7.0)	
Abnormal	527 (5.3)	366 (25.2)	893 (7.8)	
SDQ_peer problems				
Normal	6,848 (68.7)	733 (50.4)	7,581 (66.3)	280.98*
Marginal	2,653 (26.6)	518 (35.7)	3,171 (27.8)	

Continued

Variables	Non-NSSI students (*N*=9,974)	NSSI students (*N*=1,453)	Total (*N*=11,427)	χ^2
Abnormal	473 (4.7)	202 (13.9)	675 (5.9)	
SDQ_prosocial behavior				
Normal	7,910 (79.3)	1,049 (72.2)	8,959 (78.4)	38.28*
Marginal	1,049 (10.5)	199 (13.7)	1,248 (10.9)	
Abnormal	1,015 (10.2)	205 (14.1)	1,220 (10.7)	
FES_cohesion				
Low	1,030 (10.3)	530 (36.5)	1,560 (13.7)	897.37*
Medium	3,809 (38.2)	624 (42.9)	4,433 (38.8)	
High	5,135 (51.5)	299 (20.6)	5,434 (47.6)	
FES_conflict				
Low	4,891 (49.0)	310 (21.3)	5,201 (45.5)	859.38*
Medium	4,369 (43.8)	713 (49.1)	5,082 (44.5)	
High	714 (7.2)	430 (29.6)	1,144 (10.0)	
FES_achievement				
Low	4,116 (41.3)	671 (46.2)	4,787 (41.9)	12.63*
Medium	5,533 (55.5)	737 (50.7)	6,270 (54.9)	
High	325 (3.3)	45 (3.1)	370 (3.2)	
FES_intellectual-cultural				
Low	3,103 (31.1)	672 (46.2)	3,775 (33.0)	159.05*
Medium	5,572 (55.9)	699 (48.1)	6,271 (54.9)	
High	1,299 (13.0)	82 (5.6)	1,381 (12.1)	
FES_active recreational				
Low	2,012 (20.2)	578 (39.8)	2,590 (22.7)	354.57*
Medium	3,571 (35.8)	544 (37.4)	4,115 (36.0)	
High	4,391 (44.0)	331 (22.8)	4,722 (41.3)	
FES_organization				
Low	2,645 (26.5)	754 (51.9)	3,399 (29.7)	408.23*
Medium	5,935 (59.5)	619 (42.6)	6,554 (57.4)	
High	1,394 (14.0)	80 (5.5)	1,474 (12.9)	
FES_control				
Low	4,115 (41.3)	599 (41.2)	4,714 (41.3)	5.90
Medium	4,509 (45.2)	625 (43.0)	5,134 (44.9)	
High	1,350 (13.5)	229 (15.8)	1,579 (13.8)	
Insomnia				
No	8,019 (80.4)	609 (41.9)	8,628 (75.5)	1,015.67*
Yes	1,955 (19.6)	844 (58.1)	2,799 (24.5)	
Smoking history				
No	9,417 (94.4)	1,209 (83.2)	10,626 (93.0)	244.43*
Yes	557 (5.6)	244 (16.8)	801 (7.0)	
Drinking history				
No	8,405 (84.3)	905 (62.3)	9,310 (81.5)	406.09*
Yes	1,569 (15.7)	548 (37.7)	2,117 (18.5)	

Abbreviation: *N*=number; NSSI=non-suicidal self-injury; SDQ=Strengths and Difficulties Questionnaire; FES=family environment scale.
* $P<0.01$.

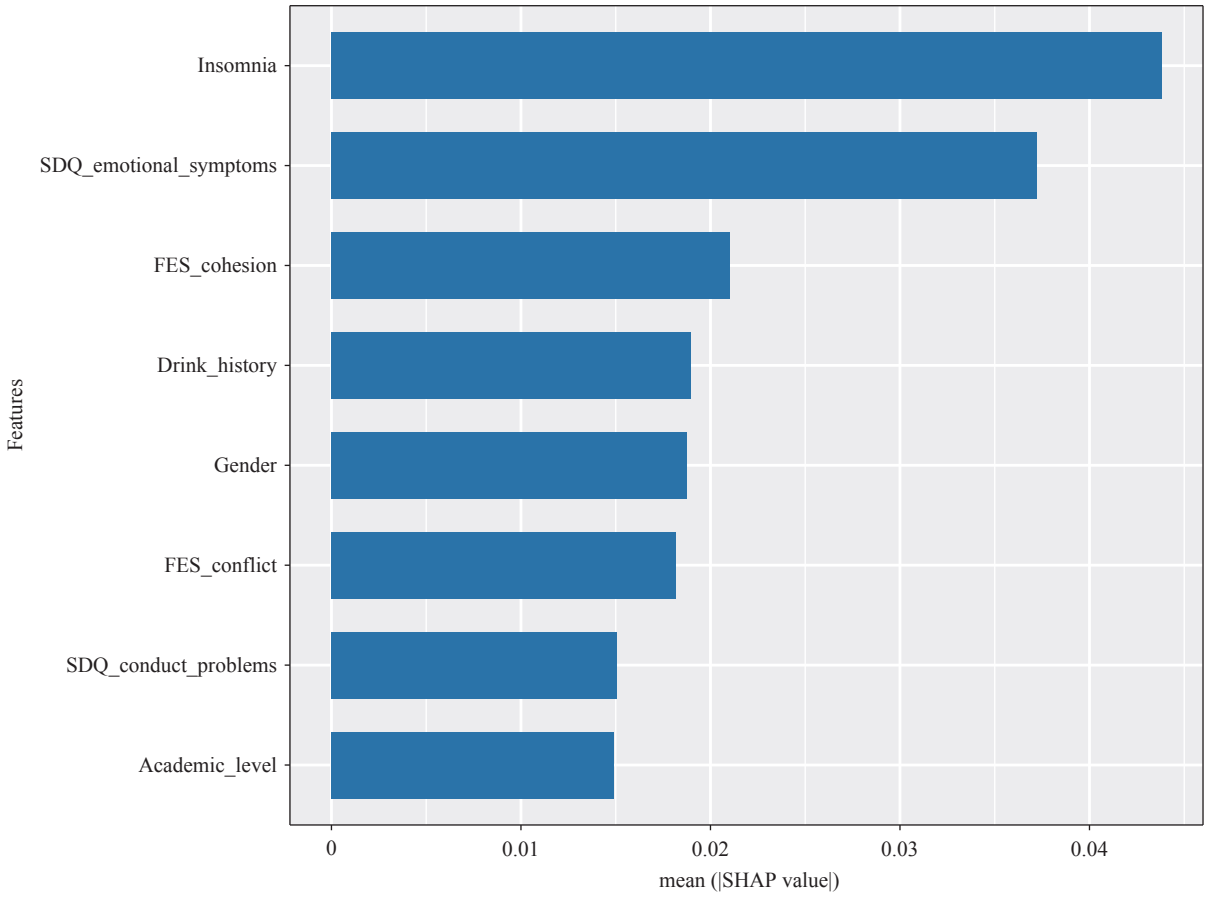


FIGURE 1. The SHAP value importance of features.
Abbreviation: SDQ=strengths and difficulties questionnaire; FES=family environment scale; SHAP=Shapley Additive exPlanations.

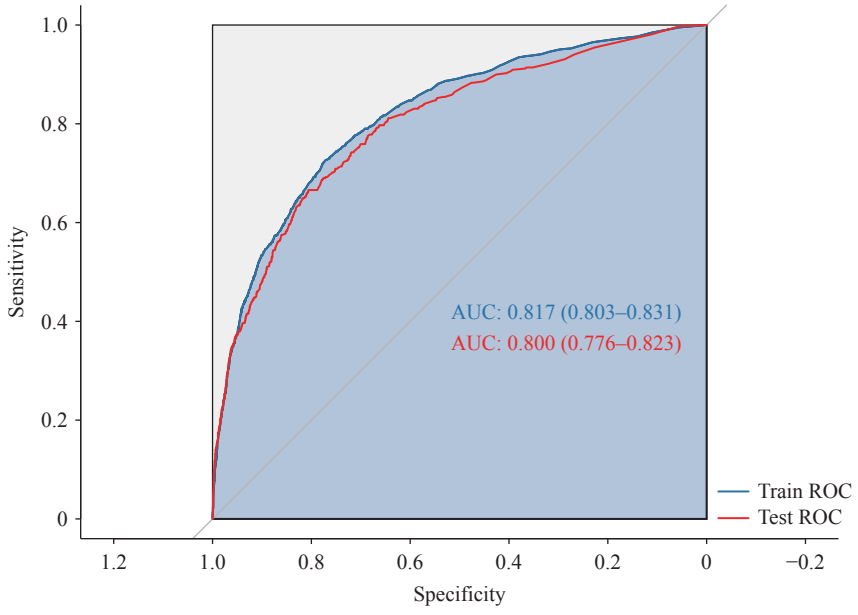


FIGURE 2. The ROC curves of training set and testing set.
Abbreviation: ROC=receiver operating characteristic; AUC=area under curve.

DISCUSSION

The prevalence of NSSI among adolescents aged 10 to 18 in Jiangsu Province is 12.72%, which is notably higher than the 4.64% reported among the same age group in Sichuan Province (*5*), yet lower than the 30.2% observed in Anhui Province (*6*). This study is the first to apply the XGBoost algorithm to develop a risk prediction model for adolescent NSSI in Jiangsu, China. Using data from 11,427 school-aged adolescents, this study's authors built a risk prediction model by filtering variables based on SHAP value importance. The model demonstrated good predictive performance, identifying insomnia, emotional symptoms, family environment cohesion, history of drinking alcohol, gender, family conflict, conduct problems, and academic level as key predictors for NSSI. Current research on NSSI prediction across various populations faces several limitations. Notably, there is a lack of studies focusing on predictive models specifically for adolescents, and existing models often include a limited range of independent variables. Additionally, few studies have explored the potential of XGBoost in NSSI prediction. This approach offers several advantages, including flexibility in data type handling, efficient training, and strong predictive performance. This study's final results confirm that the model has high predictive accuracy and a well-calibrated curve fit.

Zhou et al. (*7*) constructed a random forest model to predict NSSI among junior and senior high school students, identifying key predictors including adolescent depression, insomnia, family conflict, and gender. Similarly, Jiang et al. (*8*) reported that family environment cohesion and conflict were significant contributors to NSSI, which aligns with this study's findings. Marti-Puig et al. (*9*), using machine learning models with leave-one-subject-out (LOSO) cross-validation, demonstrated that recent adverse emotional experiences can trigger NSSI in adolescents. Consistent with these studies, this study research identified emotional symptoms and insomnia as significant predictors of NSSI (*10*). Additionally, within China's cultural and educational context, adolescence represents a transition from junior to senior high school characterized by increasing academic pressure. Consequently, academic level emerged as a predictor in this study's model. Furthermore, this study's model included drinking history and conduct problems as predictive factors, supplementing previous literature.

This study has several limitations. First, its cross-sectional design precludes causal inferences; therefore, results should be interpreted cautiously. Future longitudinal studies are needed to better quantify the influence of independent variables on NSSI. This study's "SEARCH" project is currently conducting subsequent cohort studies to address this limitation. Second, as this research was conducted in Jiangsu Province, the findings may not be generalizable to other regions or populations. External validation through multi-center or nationwide cohort studies will be essential in future research. Third, this study employed self-assessment questionnaires, particularly concerning sensitive topics such as NSSI, which may be susceptible to recall bias or underreporting. These limitations should be considered when interpreting the findings.

Currently, only a few countries have established comprehensive self-injury monitoring systems. Given that most individuals who self-injure do not seek medical attention or assistance, and considering the significant social stigma associated with self-injury behavior, medical systems can only capture a small fraction of the broader self-injury population. Most existing studies are limited to hospitalized patients or clinical cohorts. Therefore, epidemiological studies within community populations, particularly among students, are essential for understanding the behavioral characteristics of adolescent NSSI and enabling early identification and intervention for high-risk groups. Establishing a risk prediction model for NSSI among adolescents allows for more targeted interventions, ensuring that those most vulnerable receive personalized support. This study's findings highlight the necessity for personalized anti-NSSI strategies that effectively address specific behaviors and practical needs. Implementing comprehensive behavioral interventions, such as providing mental health support, improving sleep quality, and creating a harmonious family atmosphere, may help reduce NSSI among adolescents.

Conflicts of interest: No conflicts of interest.

Acknowledgements: All study participants for their contributions.

Ethical statement: Ethical approval was obtained from the Ethics Committee of the Affiliated Brain Hospital of Nanjing Medical University (Approval No. 2022-KY095-02).

Funding: Supported by the Jiangsu Provincial Key Research and Development Program (BE2021617).

doi: 10.46234/ccdcw2025.160

Corresponding authors: Jie Yang, july-summer@jscdc.cn; Fei Wang, fei.wang@yale.edu.

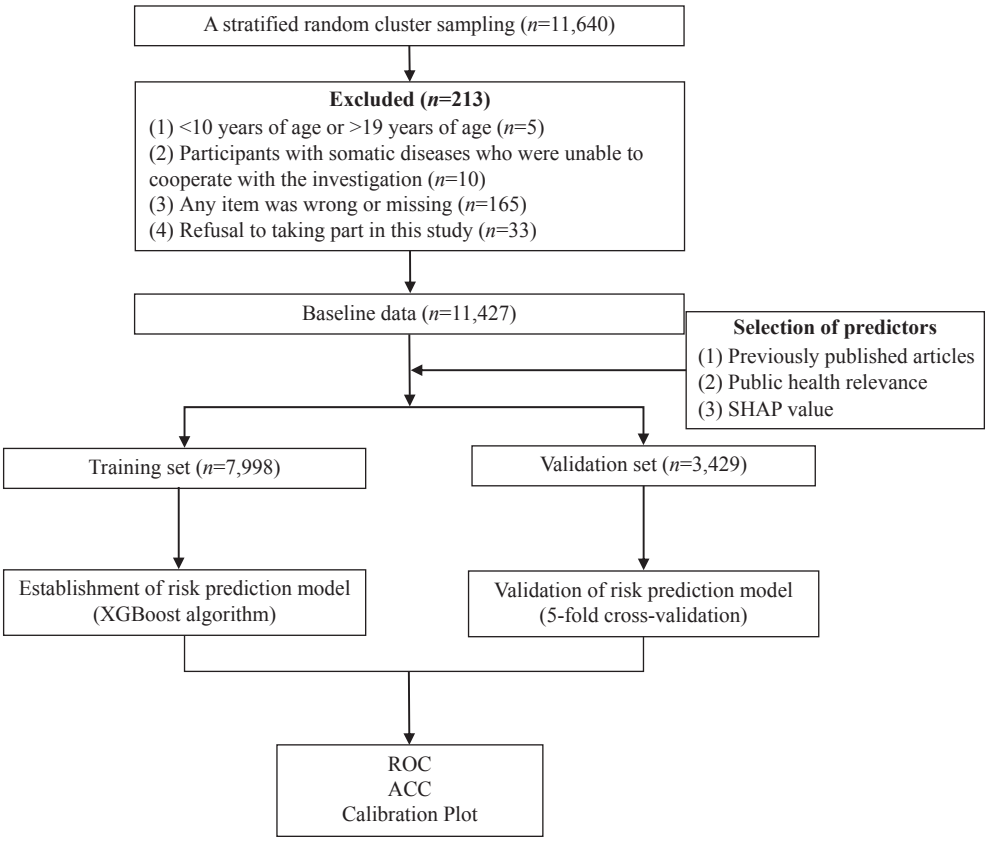
[1] Department of Child and Adolescent Health Promotion, Jiangsu Provincial Center for Disease Control and Prevention, Nanjing City, Jiangsu Province, China; [2] School of Public Health, Nanjing Medical University, Nanjing City, Jiangsu Province, China; [3] Early Intervention Unit, Department of Psychiatry, The Affiliated Brain Hospital of Nanjing Medical University, Nanjing City, Jiangsu Province, China; [4] School of Public Health, Southeast University, Nanjing City, Jiangsu Province, China.

Submitted: April 21, 2025
Accepted: June 24, 2025
Issued: July 11, 2025

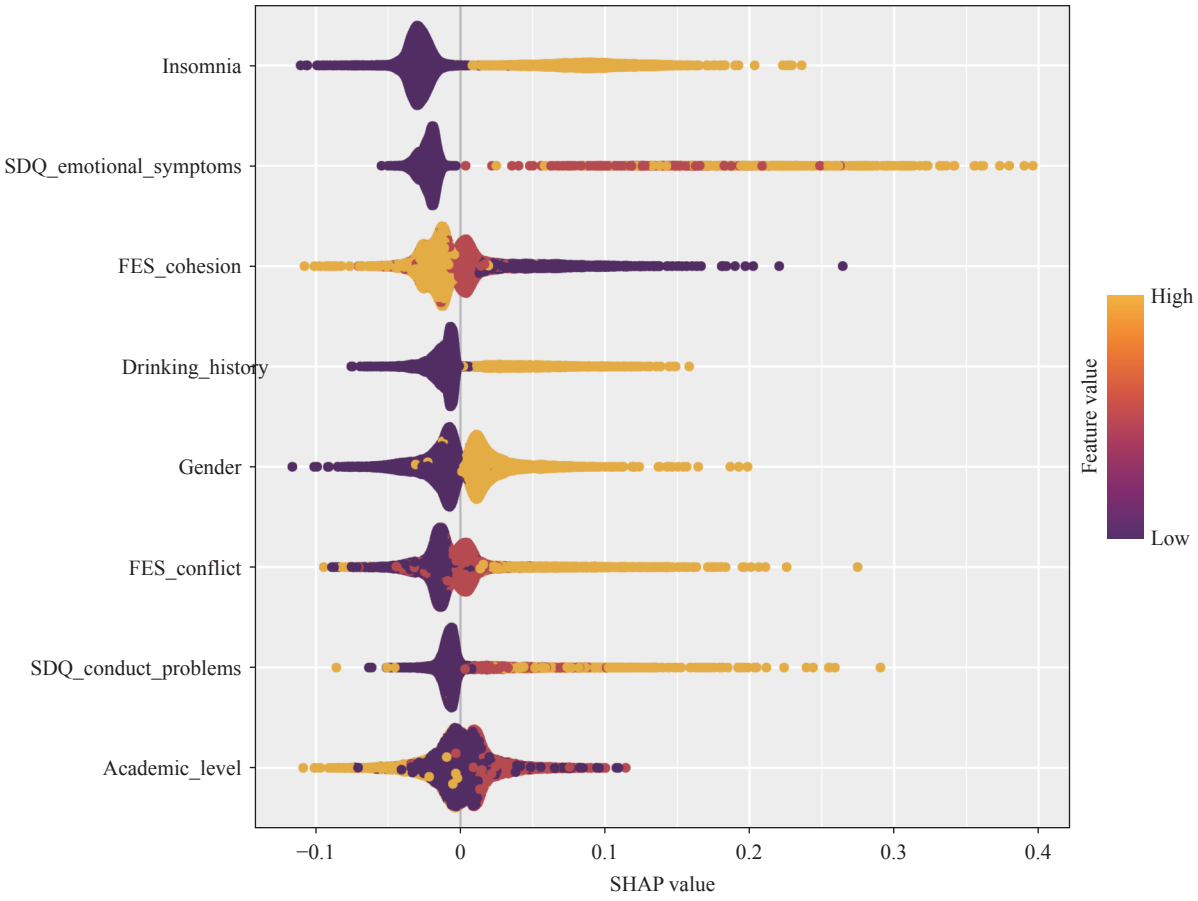
REFERENCES

1. Victor SE, Klonsky ED. Correlates of suicide attempts among self-injurers: a meta-analysis. Clin Psychol Rev 2014;34(4):282 – 97. https://doi.org/10.1016/j.cpr.2014.03.005.
2. Mars B, Heron J, Crane C, Hawton K, Lewis G, Macleod J, et al. Clinical and social outcomes of adolescent self harm: population based birth cohort study. BMJ 2014;349:g5954. https://doi.org/10.1136/bmj.g5954.
3. Aggarwal S, Patton G, Reavley N, Sreenivasan SA, Berk M. Youth self-harm in low- and middle-income countries: systematic review of the risk and protective factors. Int J Soc Psychiatry 2017;63(4):359 – 75. https://doi.org/10.1177/0020764017700175.
4. Yu G, Zhang YY, Tang LH, Xu FZ. Progress in health management of adolescent non suicidal self injury behavior. Chin J Health Manage 2022;16(1):43 – 6. https://doi.org/10.3760/cma.j.cn115624-20210902-00511.
5. Zhong YL, He JL, Luo J, Zhao JY, Cen Y, Song YQ, 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 Affect Disord 2024;345:369 – 77. https://doi.org/10.1016/j.jad.2023.10.110.
6. Lang JJ, Yao YS. Prevalence of nonsuicidal self-injury in Chinese middle school and high school students: a meta-analysis. Medicine 2018;97(42):e12916. https://doi.org/10.1097/MD.0000000000012916.
7. Zhou SC, Zhou ZH, Tang Q, Yu P, Zou HJ, Liu Q, et al. Prediction of non-suicidal self-injury in adolescents at the family level using regression methods and machine learning. J Affect Disord 2024;352:67 – 75. https://doi.org/10.1016/j.jad.2024.02.039.
8. Jiang ZL, Cui YH, Xu H, Abbey C, Xu WJ, Guo WT, et al. Prediction of non-suicidal self-injury (NSSI) among rural Chinese junior high school students: a machine learning approach. Ann Gen Psychiatry 2024;23(1):48. https://doi.org/10.1186/s12991-024-00534-w.
9. Marti-Puig P, Capra C, Vega D, Llunas L, Solé-Casals J. A machine learning approach for predicting non-suicidal self-injury in young adults. Sensors (Basel) 2022;22(13):4790. https://doi.org/10.3390/s22134790.
10. Fan YY, Liu J, Zeng YY, Conrad R, Tang YL. Factors associated with non-suicidal self-injury in Chinese adolescents: a meta-analysis. Front Psychiatry 2021;12:747031. https://doi.org/10.3389/fpsyt.2021.747031.

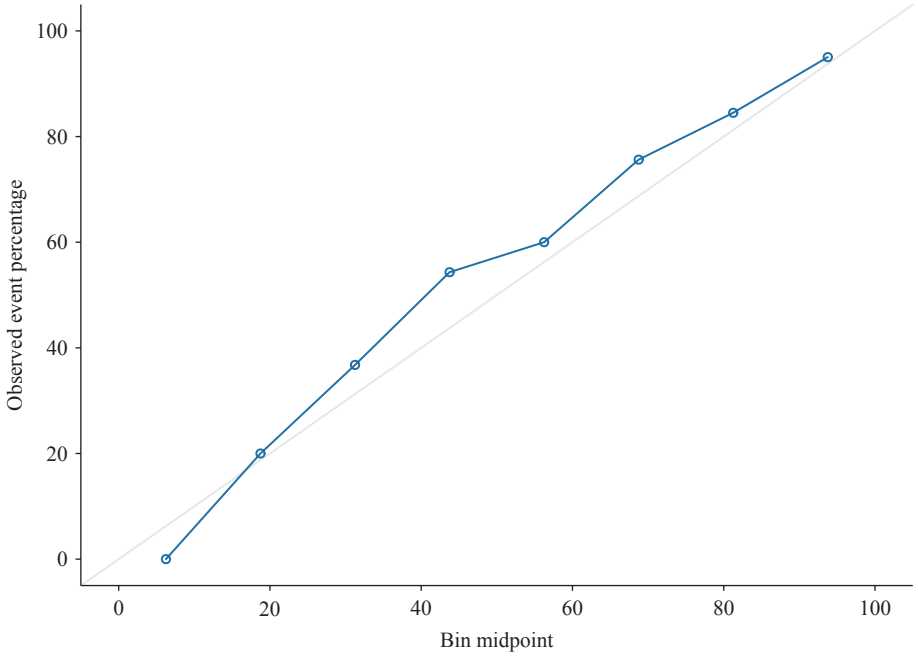
SUPPLEMENTARY MATERIAL



SUPPLEMENTARY FIGURE S1. The flowchart of data collection, model establishment and validation.
Abbreviation: SHAP=Shapley Additive exPlanations; ROC=receiver operating characteristic; ACC=accuracy.



SUPPLEMENTARY FIGURE S2. The SHAP beeswarm plot.
Abbreviation: SDQ=strengths and difficulties questionnaire; FES=family environment scale; SHAP=Shapley Additive exPlanations.



SUPPLEMENTARY FIGURE S3. The calibration curve for XGBoost model.