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.

Multi-source Data Prediction of College Students' Psychological Stress and Model Interpretability Testing

Qiang Ma
Fuyang Normal University
Dong Li
lidongfy@fynu.edu.cn

Fuyang Normal University

Research Article

Keywords: psychological stress prediction, multi-source data fusion, tree-based models, ensemble learning, SHAP

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11304486/v1

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

Additional Declarations: No competing interests reported.

Abstract

Aiming at the drawbacks that the assessment of college students' psychological stress relies on single scales and suffers poor timeliness, this study constructs a multi-source feature set integrating questionnaire scales, wearable physiological signals and campus behaviour records. Six models covering three categories (ridge regression, logistic regression, decision tree, random forest, gradient boosting decision tree and Stacking ensemble) are compared for prediction performance. Stratified five-fold cross-validation and nested hyper-parameter search are adopted, with macro-average F1-score, AUC and Brier score as evaluation metrics. The results show that multi-source fusion improves the macro-average F1-score by approximately 8-12 percentage points compared with single-source data. Gradient boosting decision tree and Stacking ensemble achieve optimal performance. SHAP and permutation importance analyses indicate that sleep-duration fluctuation, resting heart-rate variability and proportion of nocturnal activity rank highest in feature contribution. Ablation experiments verify the complementary gains of physiological and behavioural features for prediction. All codes, random seeds and data-partition schemes are reproducible.

Highlights

Multi-source fusion raised macro-F1 by 8–12 points over single-source data.

GBDT and Stacking achieved the best performance among six compared models.

Sleep-duration variability, resting HRV, and nocturnal activity were top predictors.

Ablation verified complementary gains from physiological and behavioural features.

SHAP and permutation importance gave consistent feature rankings.

Introduction

Psychological stress has become a critical factor affecting college students' academic performance and physical-mental well-being[1][2]. Timely identification of highly-stressed individuals delivers practical value for determining intervention timing. Conventional scale assessment suffers long evaluation cycles and strong subjectivity, making it hard to reflect dynamic stress changes. Wearable devices and campus information systems accumulate massive records of sleep, heart rate, activity and daily routines, offering new data sources for objective stress characterisation. Existing studies mostly focus on a single data type or a single model; the complementary gains among data sources lack quantitative measurement, and systematic comparisons of performance differences across model types as well as the stability of feature contributions are scarce. To address the above gaps, this paper builds a multi-source feature set, conducts controlled experiments on linear, tree-based and ensemble models, and quantifies the contribution of each feature group via interpretability methods. It can provide evidence for feature selection and model selection in stress prediction.

1 Multi-source Data Collection and Construction of Feature System

1.1 Data Sources, Sample Composition and Collection-Cycle Design

Participants are undergraduate students recruited from three universities, totalling 1 286 individuals. The 16-week data collection cycle covers a complete academic semester. Valid samples with complete records for all three data types amount to 1 142, with 412, 398 and 332 valid samples from the three universities respectively. The sample sizes are close, so the schedule system of any single university cannot dominate overall results. Participants cover Grade One to Grade Four, accounting for 31%, 28%, 24% and 17% respectively. Majors include science-engineering, economics-management and arts-humanities, which can reflect stress-level disparities across different academic stages. Valid samples require participants to wear bracelets for no less than 70% of the full cycle and complete questionnaires at least three times, guaranteeing temporal continuity. The questionnaire scale consists of three sub-sections: sleep quality, academic burden and social support[3–5], administered every four weeks. Stress labels derive from graded results of the Perceived Stress Scale. Features and labels come from distinct scales to avoid information overlap. Physiological signals are continuously recorded by smart bracelets and aggregated on a daily basis. Campus behaviour records are extracted from access-control, library and canteen systems, and linked to other datasets via anonymised unified IDs. Given differences in collection frequency and population coverage among the three data types, Table 1 illustrates their composition to support subsequent missing-value handling and time alignment.

Table 1
Sources and sample composition of multi-source data

Data Type	Primary Source	Collection Frequency	Valid Sample Size	Number of Raw Features
Questionnaire Scale	Sleep-quality, academic-burden and social-support scales	Every 4 weeks	1 271	32
Wearable Physiological Signals	Smart bracelets	Continuous collection, aggregated daily	1 183	46
Campus Behaviour Records	Access-control, library and canteen systems	Aggregated daily	1 254	38
Complete for all three sources	—	—	1 142	116

1.2 Missing-value Processing, Outlier Detection and Time-window Alignment

Physiological signals contain the largest proportion of missing values caused by interrupted bracelet wear. Nine features with a missing rate exceeding 30% are eliminated. For remaining missing entries, adjacent-date interpolation combined with individual-median imputation is applied. Outlier detection mainly adopts the inter-quartile-range method aided by physiological-range constraints. For instance, resting heart rate is constrained between 40 and 120 beats per minute; out-of-range records are winsorised. The overall winsorisation proportion is below 1.5%, exerting limited influence on feature distributions. In terms of time alignment, all datasets adopt the 14-day observation window prior to questionnaire completion and perform weekly aggregation. Days with bracelet wear duration shorter than 16 hours are excluded from calculation, unifying temporal scales for differently-sampled data. The 14-day window covers two full daily-routine cycles, smoothing single-day fluctuations while preserving phase-specific variations such as examination weeks. After preprocessing, 107 out of the original 116 features are retained, including 32 questionnaire features, 39 physiological features and 36 behavioural features. Eliminated features are predominantly physiological ones, consistent with bracelet-wear interruptions[6][7]. The cleaned feature matrix maintains identical observation units across data sources, laying a unified foundation for feature extraction and screening in Section 1.3.

1.3 Extraction and Screening of Questionnaire, Physiological and Behavioural Features

Feature extraction is performed separately per data source to preserve physical interpretations and consistent calculation standards. Questionnaire features adopt total scores and dimension-level sub-scores of each scale[8][9]. Physiological features centre on sleep and heart-rate metrics: mean sleep duration and its fluctuation, deep-sleep ratio, resting heart-rate variability. Behavioural features characterise daily routines and learning engagement: proportion of nocturnal activity, library visiting duration and regularity of three daily meals[10][11]. During feature screening, features with extremely low variance and redundant features whose correlation coefficient exceeds 0.9 are first removed. Recursive feature elimination is then executed according to feature-importance ranking within training folds. The recursion terminates when the macro-average F1-score on validation folds shows no substantial improvement across three consecutive rounds. Thus information from test folds never leaks into feature selection. The 107 features are finally condensed into 58 (14 questionnaire, 24 physiological, 20 behavioural), with roughly balanced counts across categories to avert bias stemming from disproportionate feature volume of any single source. Approximately 85% overlap exists among selected features across different folds, demonstrating that screening outcomes do not depend on one specific data partition. All features are computed within the 14-day observation window described in Section 1.2, enabling cross-source comparison on the identical time span (see Fig. 1).

2 Predictive-model Design and Training Scheme

2.1 Construction of Linear-model Baselines: Ridge Regression and Logistic Regression

Linear models serve as baselines to quantify extra performance gains brought by non-linear approaches. Ridge regression first predicts continuous stress scores, which are then categorised into low-, medium- and high-stress classes according to grading thresholds. Multinomial logistic regression directly outputs three-class probabilities[12–14]. Both utilise penalty terms to mitigate coefficient instability induced by collinear features. The penalty coefficient for ridge regression is searched within [0.01, 100]; regularisation strength for logistic regression ranges from 0.001 to 10, with maximum iterations set to 500 for convergence. Given that high-stress samples only account for roughly 22%, class-weighting proportional to class frequency is implemented for both models to prevent prediction bias towards majority classes. All features are standardised on training-fold data prior to model input for comparable coefficient magnitudes. Linear models share identical fold partitions and feature sets with subsequent tree-based models. Hence performance gaps among models originate purely from model architectures rather than divergent preprocessing. Thanks to simple structures and stable training, linear models act as benchmarks and provide references for interpreting non-linear effects in Section 4.2.

2.2 Tree-based and Ensemble Models: Decision Tree, Random Forest, Gradient Boosting Decision Tree and Stacking Construction

A single decision tree represents standalone tree-based modelling; its maximum depth is constrained to 3-8 layers to mitigate over-fitting on training samples[15]. Its intuitive rule-based structure offers a reference for subsequent interpretation. Random forest suppresses prediction variance via averaging outputs of multiple trees, while gradient boosting decision tree iteratively corrects residuals to approximate complex relationships; the two differ in bias-variance trade-offs. The random forest comprises 300 base trees with random feature sampling at each split to reduce variance of individual trees. Gradient boosting decision tree uses a learning rate of 0.05, base learners of depth 4, maximum iterations of 400 and early-stopping guided by validation-fold performance. For Stacking ensemble, logistic regression, random forest and gradient boosting decision tree are selected as base learners, and logistic regression functions as the meta-learner. Inputs for meta-learner are out-of-fold predicted probabilities generated by base learners, avoiding repeated utilisation of training-set labels. Only three base learners are adopted to control model complexity and training overhead. Tree-based models are insensitive to feature scaling and therefore require no standardisation; missing values have been imputed in Section 1.2, so no further input transformation is needed. The above four models plus the two linear baselines constitute the six-model comparison framework covering linear, single-tree and ensemble paradigms.

2.3 Cross-validation, Hyper-parameter Search and Reproducible Experimental Configuration

Group-stratified five-fold cross-validation grouped by student ID is adopted. Records belonging to the same student never simultaneously appear in training and test folds, and class-label proportions are maintained across folds. Hyper-parameter search is nested inside each outer fold: inner three-fold

validation combined with 50-times random search optimises macro-average F1-score. Outer test folds are never involved in parameter tuning. Parameters for missing-value imputation and standardisation are estimated exclusively on training folds and then applied to corresponding test folds to eliminate information leakage in preprocessing[16–18]. Evaluation metrics include macro-average F1-score, AUC and Brier score, measuring class-balanced classification performance, ranking capacity and probability-calibration quality respectively. All metrics are calculated on prediction outputs from outer test folds. The whole pipeline is repeated under five random seeds (primary seed fixed as 42). Data-partition schemes and full configurations are archived with source code for complete reproducibility. Results of five-fold cross-validation under each random seed are averaged before aggregation across seeds. Pairwise tests are applied to judge statistical significance of inter-model differences at the significance level $\alpha = 0.05$. This protocol supplies unbiased evaluation without information leakage for model comparison in Section 3.1.

3 Experimental Results and Comparative Analysis

3.1 Overall Prediction-performance Comparison across Three Model Categories

Under unified evaluation settings, performance presents clear gradients. Linear models yield macro-average F1-scores between 0.681 and 0.702; the single decision tree reaches 0.688, all at relatively low levels. Random forest achieves 0.751; gradient boosting decision tree and Stacking ensemble obtain 0.779 and 0.786 respectively, with corresponding AUC values rising to 0.896 and 0.901. Random forest outperforms the single decision tree by 6.3 percentage points, confirming the effectiveness of multi-tree averaging for mitigating single-model fluctuations. Stacking surpasses logistic regression by 8.4 percentage points. Meanwhile the decision tree yields an AUC of merely 0.794, inferior to both linear models, indicating insufficient stability of standalone trees given substantial feature collinearity. The Brier score drops from 0.196 (logistic regression) to 0.138 (Stacking), demonstrating superior probability-calibration quality of ensemble models. Stacking outperforms gradient boosting decision tree by only 0.7 percentage points, with standard deviation around 0.006 across five random seeds and p-value = 0.21 from pairwise testing. The performance gap lies within random-variation range; the two are treated as jointly optimal. All above comparisons utilise the full feature set. Contributions of individual data sources are examined in Section 3.2 (see Table 2).

Table 2
Overall prediction performance of six models (mean values from stratified five-fold cross-validation)

Model Category	Model	Macro-average F1-score	AUC	Brier Score
Linear Model	Ridge Regression	0.681	0.812	—
Linear Model	Logistic Regression	0.702	0.831	0.196
Tree-based Model	Decision Tree	0.688	0.794	0.213
Ensemble Model	Random Forest	0.751	0.874	0.158
Ensemble Model	Gradient Boosting Decision Tree	0.779	0.896	0.141
Ensemble Model	Stacking	0.786	0.901	0.138

3.2 Ablation Test and Fusion-gain Validation for Data Sources

Ablation experiments are performed on the Stacking model by retraining under different data-source combinations. Macro-average F1-scores range from 0.672 to 0.694 with single-source inputs, and reach 0.786 under three-source fusion, yielding improvements of 9.2-11.4 percentage points (consistent with the 8-12-percentage-point interval reported in the abstract). Among single-source settings, questionnaire features deliver the best performance while physiological features produce the weakest results, partly attributable to substantial missing entries and high noise in individual physiological indicators[19–21]. Behavioural features occupy an intermediate position, suggesting campus logs contain stable stress-related signals. Among two-source combinations, the physiological-plus-behavioural setting achieves the highest macro-average F1-score (0.752), implying low information overlap between these two objective data modalities. Adding questionnaire features further lifts performance by 3.4 percentage points, proving subjective assessments carry independent predictive information. Removing any single data source from the full-feature set reduces macro-average F1-score by 3.4-5.2 percentage points. The largest performance drop occurs when physiological features are excluded. Although physiological signals perform poorly in isolation, they deliver irreplaceable complementary information in multi-source fusion. Hence three feature groups exhibit complementary gains instead of simple additive effects[22–24]. This finding is further corroborated via feature-contribution analysis in Section 4.1 and demonstrates that prediction schemes relying on a single data type suffer notable information deficits (see Table 3).

Table 3
Data-source ablation outcomes (Stacking, macro-average F1-score)

Feature Combination	Number of Features	Macro-average F1-score	Difference vs. full-feature setting (percentage points)
Questionnaire only	14	0.694	−9.2
Physiological only	24	0.672	−11.4
Behavioural only	20	0.683	−10.3
Questionnaire + Physiological	38	0.741	−4.5
Questionnaire + Behavioural	34	0.734	−5.2
Physiological + Behavioural	44	0.752	−3.4
All three sources	58	0.786	—

4 Interpretability Analysis and Identification of Influencing Factors

4.1 Global Feature Importance: Permutation Importance and SHAP Summary

Global interpretability analysis targets the gradient boosting decision tree, which is compatible with tree-based SHAP algorithms and delivers performance comparable to Stacking. Permutation importance and aggregated SHAP values produce consistent feature-importance rankings. The mean absolute SHAP value for sleep-duration variability equals approximately 0.31; resting heart-rate variability is about 0.24; proportion of nocturnal activity reaches roughly 0.19. Together these three features account for nearly 40% of total predictive contribution. Permutation importance quantifies performance loss via macro-average F1-score decay after feature shuffling. Random permutation of sleep-duration variability reduces F1-score by about 4.1 percentage points, consistent with SHAP-based ordering[25–27]. Two of the top-three features (sleep-duration variability, resting heart-rate variability) originate from wearable bracelets[28][29], whereas proportion of nocturnal activity comes from campus-behaviour logs. Therefore top-ranked features are distributed across two objective-data modalities rather than concentrating on one single source. Aggregated by data source, physiological features contribute about 46%, behavioural features 34% and questionnaire features 20% of total importance, corroborating the finding in Section 3.2 that physiological-plus-behavioural combination outperforms other two-source configurations. Eight out of the top-ten features overlap between permutation-importance and SHAP rankings, demonstrating stable importance conclusions (see Fig. 2).

4.2 Non-linear Effects and Threshold Intervals of Key Features

Non-linear effects are identified using SHAP-dependence plots and segment-wise analysis. Features are partitioned into ten quantile-based bins. Average SHAP contributions for high-stress samples are computed for each bin; threshold candidates are located where contribution signs invert between adjacent bins[30–32]. When the standard deviation of 14-day sleep duration is below approximately 0.8 hours, its contribution to high-stress prediction approximates zero; contribution rises sharply once the metric exceeds 1.2 hours. This suggests sleep-schedule stability delivers more diagnostic information than mean sleep duration. When resting heart-rate variability falls below roughly 35 ms, its positive contribution to high-stress prediction rises prominently and gradually weakens as values increase. Proportion of nocturnal activity saturates beyond 25% with no further contribution growth[33][34]. Threshold estimates shift by less than 10% across five random seeds, proving turning points are not artefacts of random fold sampling. Monitoring systems may adopt these values to set graded alerts, e.g. triggering concern if the 14-day sleep-duration standard deviation exceeds 1.2 hours for two consecutive weeks. Nevertheless, single-threshold discrimination has limited power and must be combined with multi-feature evidence[35][36]. These quantified turning points capture non-linear patterns inaccessible to linear models and serve as reference intervals for instance-level attribution in Section 4.3.

4.3 Instance-level Prediction Attribution and Feature Patterns of High-stress Samples

SHAP local attribution is applied at individual level. For roughly 250 students predicted as high-stress within test folds, prediction probabilities are decomposed feature-by-feature starting from the baseline value. Every student's predicted probability can be traced back to concrete feature deviations. Clustering identifies three major feature-pattern subgroups among high-stress instances[37–39] Around 58% of high-stress samples exhibit concurrently elevated sleep-duration variability and nocturnal-activity ratio, mostly observed around examination weeks. Approximately 27% are dominated by low resting heart-rate variability alongside relatively regular daily routines. The remaining 15% stem from superposition of minor deviations across multiple features with no single dominant predictor, representing the hardest-to-detect subgroup requiring comprehensive judgement[40]. The first two patterns correspond to top-ranked global features in Section 4.1. Individual-level attributions align with global importance rankings and additionally reveal subgroup heterogeneity invisible in global summaries. Subgroup-proportional fluctuations stay within 5 percentage points across test folds, supporting their utility for stratified intervention. For samples driven primarily by low heart-rate variability, behavioural logs show few anomalies while physiological signals diverge early. Continuous physiological monitoring therefore adds value for early warning. Identical prediction outputs may arise from divergent underlying drivers; instance-level attribution yields more reliable insights than global-importance rankings alone.

Conclusion

Experiments confirm that integrating questionnaire, physiological and behavioural data substantially improves psychological-stress prediction performance relative to single-source inputs. Ablation results verify complementary relationships among the three feature groups. Among linear, tree-based and ensemble models, gradient boosting decision tree and Stacking ensemble achieve optimal overall performance. Linear models possess advantages in interpretability and stability and can serve as deployment baselines. Interpretability analyses reveal sleep-schedule stability, heart-rate variability and proportion of nocturnal activity as dominant predictive features, with evident threshold effects for several metrics, implying psychological-stress dynamics are better characterised by non-linear relationships. Cross-semester validation indicates moderate generalisation capability with limited performance decay. Future work should validate conclusion stability on larger datasets over longer time spans and explore real-time feature-update mechanisms.

Declarations

Authorship Contribution

Qiang Ma: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft.

Dong Li: Validation, Resources, Supervision, Project administration, Funding acquisition, Writing – review & editing.

All authors read and approved the final manuscript.

Data Availability Statement

The data and code that support the findings of this study are available from the corresponding author upon reasonable request. The raw wearable and campus behaviour data are not publicly available due to privacy and ethical restrictions.

Declaration of Competing Interest

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

Funding

This work was supported by the Anhui Provincial Higher Education Humanities and Social Sciences Research Youth Project [grant number 2025AHGXSK40171].

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors (Qiang Ma and Dong Li) used ChatGPT (OpenAI) for language translation from Chinese to English. After using this tool/service, the authors reviewed and

edited the content as needed and take full responsibility for the content of the publication. ChatGPT was not used for data collection, data analysis, interpretation of results, or generation of scientific conclusions.

Clinical Trial Number

Clinical trial number: not applicable.

Ethics Approval and Consent to Participate

This study was conducted in accordance with the Declaration of Helsinki. As Altay Vocational and Technical College does not have an Institutional Review Board (IRB) or a formal Ethics Committee, the study was reviewed and approved by the Academic Committee of Altay Vocational and Technical College, which confirmed that it meets ethical requirements. All participants were fully informed of the purpose and procedures of the study and provided written informed consent prior to participation. Participation was voluntary, and participants had the right to withdraw at any time without consequence.

Consent for Publication

All participants provided informed consent for the publication of their anonymized data. The data presented in this study are anonymized and do not compromise participant privacy. No individually identifiable information is included.

References

1. LI M,. MULTIMODAL FUSION OF PHYSIOLOGICAL SIGNALS AND BEHAVIORAL DATA FOR MENTAL HEALTH SHANXCHENX, ASSESSMENT: A DEEP LEARNING APPROACH BASED ON HIERARCHICAL CNN [J]. J Mech Med Biology. 2026;prepublish10.1142/S0219519426400981.
2. Yang S, Wang S, Wang Y, et al. A generalized supervised contrastive learning framework for integrative multi-omics prediction models [J]. Front Microbiomes. 2026;5:1825540–1825540. 10.3389/FRMBI.2026.1825540.
3. Girón ACD, Rodriguez MJW, Rojas ZGM, et al. AI. 2026;7(7):268–268. 10.3390/AI7070268. Predicting Student Stress Using Machine Learning Ensemble Models: A Multi-Criteria Comparison with Explainable Artificial Intelligence Analysis [J].
4. Omarbekova A, Nazyrova A, Bekmanova G, et al. An explainable machine learning framework for academic stress classification among university students: a comparative multi-dataset study with ablation analysis and statistical significance testing [J]. Front Comput Sci. 2026;8:1886274–1886274. 10.3389/FCOMP.2026.1886274.
5. Hamd YZ, Mohmed AZ, Alroqaiba N, et al. Correction: Utilizing artificial intelligence to assess academic exam anxiety, perceived stress, and achievement motivation among college students [J]. Front Psychiatry. 2026;17:1830610–1830610. 10.3389/FPSYT.2026.1830610.

6. Zhang W, Xie Y, Ma L. Front Psychol. 2026;17:1777211–1777211. 10.3389/FPSYG.2026.1777211. Pathways for the precise prevention and improvement of mental health among university freshmen: a network analysis and simulated intervention study based on the biopsychosocial model [J].
7. Li F, Wang L, Chen W, et al. A multimodal deep learning approach for mental health classification of university students: an intelligent early warning system [J]. Front Artif Intell. 2026;9:1679030–1679030. 10.3389/FRAI.2026.1679030.
8. Yasmin NM, Rahman MM, Ahmed YN, et al. An explainable machine learning framework for analyzing and predicting mental health problems among university students in Bangladesh. [J] Sci Rep. 2026;16(1):23179–23179. 10.1038/S41598-026-52351-8.
9. Tariq R, Castillo D O G M, Zamir TM, et al. Retraction Note: Explainable artificial intelligence for predictive modeling of student stress in higher education [J]. Sci Rep. 2026;16(1):14717–14717. 10.1038/S41598-026-51945-6.
10. Yan H, Zheng Y. Discover Comput. 2026;29(1):205–205. 10.1007/S10791-026-10062-8. Early warning system for mental health risks among college students based on artificial intelligence and multimodal data [J].
11. Sathyanarayana A, Amin BO, An J, et al. Examining the Use of Consumer Wearable Devices and Digital Tools for Stress Measurement in College Students: Scoping Review of Methods. [J]. JMIR mHealth uHealth. 2026;14:e64144. 10.2196/64144.
12. Ceja GE, Uribe AJ, Ambrosio EJP, et al. Sci Data. 2026;13(1):732–732. 10.1038/S41597-026-07085-7. A Dataset of University Students' Stress and Anxiety Levels based on Questionnaires and Wearable Sensors [J].
13. Hamd YZ, Mohmed AZ, Alroqaiba N, et al. Utilizing artificial intelligence to assess academic exam anxiety, perceived stress, and achievement motivation among college students [J]. Front Psychiatry. 2026;17:1686106–1686106. 10.3389/FPSYT.2026.1686106.
14. Norbury A, Fairs G, Georgescu LA, et al. [J] Sci Rep. 2026;16(1):5397–5397. 10.1038/S41598-025-33331-W. A multimodal Bayesian network for symptom-level depression and anxiety prediction from voice and speech data.
15. Ji K, Wang X, Yuan D, et al. Stress health : J Int Soc Invest Stress. 2026;42(1):e70137. 10.1002/SMI.70137. Predicting Depression Risk Among Chinese College Students From Childhood Adversity and Recent Stress Exposure Using Interpretable Machine Learning. [J].
16. Liu C, Yu S. Front Psychol. 2026;16:1684529–1684529. 10.3389/FPSYG.2025.1684529. Students’ stress prediction and explainable analysis based on improved decision trees [J].
17. Jiang D. Predicting student mental health through entropy-based features and interpretable cross-attention transformer networks. [J] PloS one. 2026;21(4):e0347294. 10.1371/JOURNAL.PONE.0347294.
18. Grimm B, Yilmam P, Talbot B, et al. [J] NPJ Digit Med. 2025;9(1):65–65. 10.1038/S41746-025-02242-5. Multimodal machine learning for video based single question mental health assessment.

19. Shi X, Pan J, Yuan D, et al. Front Psychol. 2025;16:1682083–1682083. 10.3389/FPSYG.2025.1682083. Advanced data analysis and prediction model for student mental health risk assessment [J].
20. Tariq R, Castillo D O G M, Zamir TM, et al. Explainable artificial intelligence for predictive modeling of student stress in higher education. [J] Sci Rep. 2025;15(1):38375. 10.1038/S41598-025-22171-3.
21. Ding L, Zhang H, Zhou J. Crying in the algorithm: modeling academic stress via multilayer topic construction and ERA effect [J]. Front Psychol. 2025;16:1673559–1673559. 10.3389/FPSYG.2025.1673559.
22. Shen S, Qi W, Zeng J, et al. Passive Sensing for Mental Health Monitoring Using Machine Learning With Wearables and Smartphones: Scoping Review. [J]. J Med Internet Res. 2025;27:e77066. 10.2196/77066.
23. Koushal H, Kaur R, Dhaliwal KC. Machine Learning for Mental Health: Assessing Teen Depression and Anxiety Risk Factors [J]. Cureus J Comput Sci. 2025;2(1):5123–5123. 10.7759/S44389-025-05123-W.
24. Borelli LJ, Wang Y, Li HF, et al. Detection of Depressive Symptoms in College Students Using Multimodal Passive Sensing Data and Light Gradient Boosting Machine: Longitudinal Pilot Study. [J]. Volume 9. JMIR formative research; 2025. p. e67964. 10.2196/67964.
25. Lee H, Kim J, Han B, et al. Developing an explainable Deep Neural Network for stress detection using biosignals and human-engineered features [J]. Biomed Signal Process Control. 2025;109:107960–107960. 10.1016/J.BSPC.2025.107960.
26. Filippis DR, Foysal AA. Comprehensive analysis of stress factors affecting students: a machine learning approach [J]. Discover Artif Intell. 2024;4(1):62–62. 10.1007/S44163-024-00169-6.
27. Tanwar R, Singh G, Pal KP. A hybrid transposed attention based deep learning model for wearable and explainable stress recognition [J]. Comput Electr Eng. 2024;119(PB):109551–109551. 10.1016/J.COMPELECENG.2024.109551.
28. Shvetcov A, Kupper FJ, Zheng YW, et al. Passive sensing data predicts stress in university students: a supervised machine learning method for digital phenotyping [J]. Front Psychiatry. 2024;15:1422027–1422027. 10.3389/FPSYT.2024.1422027.
29. Moser KM, Ehrhart M, Resch B. An Explainable Deep Learning Approach for Stress Detection in Wearable Sensor Measurements [J]. Sensors. 2024;24(16):5085–5085. 10.3390/S24165085.
30. Choi A, Ooi A, Lottridge D. Digital Phenotyping for Stress, Anxiety, and Mild Depression: Systematic Literature Review. [J]. JMIR mHealth and uHealth, 2024, 12 e40689-e40689. 10.2196/40689
31. Liao Z, Fan X, Ma W, et al. Mathematics. 2024;12(10). 10.3390/MATH12101501. An Examination of Mental Stress in College Students: Utilizing Intelligent Perception Data and the Mental Stress Scale [J].
32. Fontes L, Machado P, Vinkemeier D, et al. Enhancing Stress Detection: A Comprehensive Approach through rPPG Analysis and Deep Learning Techniques [J]. Sensors. 2024;24(4). 10.3390/S24041096.

33. Alrazaq AA, Alajlani M, Ahmad R, et al. The Performance of Wearable AI in Detecting Stress Among Students: Systematic Review and Meta-Analysis. [J]. J Med Internet Res. 2024;26(1):e52622–52622. 10.2196/52622.
34. Luo L. Research on practical education model of mental health development of college students in higher education institutions in the context of big data [J]. Applied Mathematics and Nonlinear Sciences, 2024, 9 (1): 10.2478/AMNS.2023.2.00244
35. Morr EC, Jammal M, Hamad BI, et al. Predictive Machine Learning Models for Assessing Lebanese University Students' Depression, Anxiety, and Stress During COVID-19. [J]. J Prim care community health. 2024;15:21501319241235588–21501319241235588. 10.1177/21501319241235588.
36. Ritu T, Chetia OP, Ghanapriya S, et al. Attention based hybrid deep learning model for wearable based stress recognition [J]. Eng Appl Artif Intell. 2024;PB127. 10.1016/J.ENGAPPAI.2023.107391.
37. Jahan IR, Muntasir MN, Fahim F, et al. Analyzing Perceived Psychological and Social Stress of University Students: A Machine Learning Approach. [J] Heliyon. 2023;9(6):e17307–17307. 10.1016/J.HELIYON.2023.E17307.
38. Vaz M, Summavielle T, Sebastião R, et al. Multimodal Classification of Anxiety Based on Physiological Signals [J]. Appl Sci. 2023;13(11). 10.3390/APP13116368.
39. Danielle C, John T. [J] J Med Internet Res. 2023;e2539258–39258. 10.2196/39258. Digital Phenotyping Data to Predict Symptom Improvement and Mental Health App Personalization in College Students: Prospective Validation of a Predictive Model.
40. Rajendran GV, Jayalalitha S, Adalarasu K, et al. A Review on Mental Stress Detection Using PSS Method and EEG Signal Method [J]. Electrochem Soc Trans. 2022;107(1). 10.1149/10701.1845ECST.

Figures

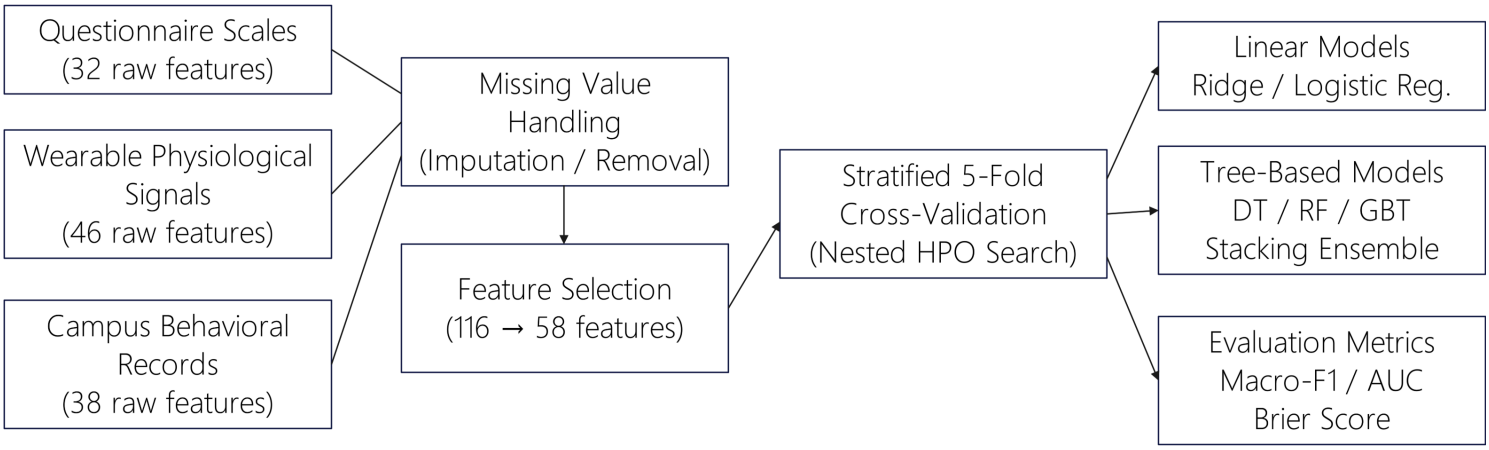


Figure 1

Multi-Source Data Fusion and Predictive Modeling Pipeline

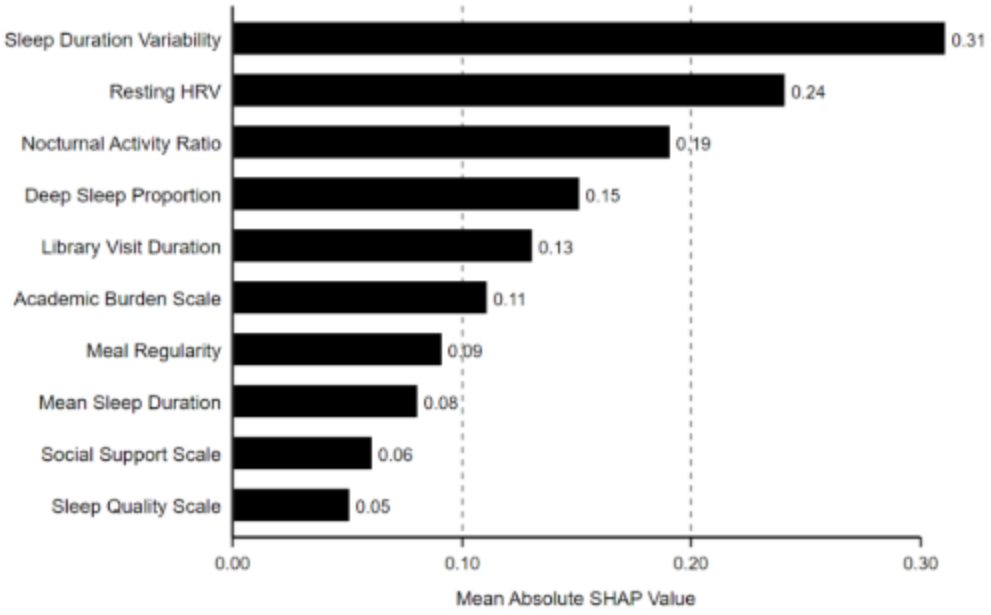


Figure 2

Global SHAP Feature Importance: Top 10 Features