Journal of Engineering, Project, and Production Management
2026, 16(3), 2026-198

Psychological Abnormal Student Identification Based on Multi-Source Heterogeneous Educational Data

Hansong Dong[1] and Jiao Huang[2]
[1] Associate Professor, International School of Technical Education, Sichuan University of Architectural Technology, Deyang, 618000, China.
[2] Associate Professor, International School of Technical Education, Sichuan University of Architectural Technology, Deyang, 618000, China, E-mail: huangjiaohj1@163.com (corresponding author).

Project Management
Received February 7, 2026; revised March 16, 2026; accepted March 19, 2026
Available online April 8, 2026

Abstract: With the increasingly prominent mental health issues among college students, identifying students with psychological abnormalities through behavioral data has become an urgent research problem to be solved. A Hybrid Model for Psychological Abnormal Student Behavior Identification (HMPABI) based on Multi-source Heterogeneous Educational Data is proposed. By combining multidimensional behavioral characteristics and psychological assessment data from students during their school years, an efficient and accurate identification model for students with psychological abnormalities is constructed through clustering, oversampling techniques, and a mixed classification strategy combining logistic regression and support vector machines. Psychological abnormalities are defined operationally as significant behavioral deviations in social, academic, and daily routines that correlate with established clinical indicators of mental health risks. The study takes the Urumqi University Student Campus Behavior Dataset and the Adolescent Mental Health and Behavior Dataset for experimental verification. Performance was evaluated using the Mean Absolute Error framework to quantify the deviation between predicted risk scores and actual assessment labels across different behavioral observation time windows, which represent the data aggregation intervals for anomaly detection. On the Urban Underground Space-Central Business District UUS-CBD dataset, the HMPABI model consistently had lower error than the comparison models across all testing windows, achieving a maximum error of 0.28. In contrast, the errors of the other two models reached 0.45 and 0.42 at the maximum time window (45 minutes), respectively. The HMPBAI model can fully explore potential information of students behavioral characteristics. By integrating different types of data, it can more accurately predict students with psychological abnormalities. This study provides a new technological path for mental health monitoring, wherein the dynamically generated risk scores can be integrated into campus support systems to actively alert counseling centers, thereby enabling targeted, proactive early interventions for at-risk college students.

Keywords: Psychological abnormalities, behavioral characteristics, K-means clustering, smote oversampling, machine learning, decision support systems, educational management, risk monitoring, data-driven intervention.

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).
DOI 10.32738/JEPPM-2026-198

1. Introduction

In universities, students mental health issues are increasingly becoming an important topic for the academic community, educators, and various sectors of society (Chen et al., 2023). According to statistical data, an increasing number of college students are exhibiting varying degrees of psychological abnormalities when facing academic pressure, social anxiety, emotional distress, and other issues. This not only affects students learning outcomes and quality of life but may even have a profound impact on their future development (Kurcer et al., 2022; Yang and Ge, 2022). Therefore, a critical challenge lies in leveraging educational big data mining and behavioral feature analysis to construct an intelligent and proactive identification model. Developing such a model is crucial for achieving early, effective interventions, marking a significant step forward from traditional, often passive approaches to student mental health. Traditional methods for identifying students with psychological abnormalities mainly rely on psychological assessment tools, questionnaire surveys, and face-to-face psychological counseling. Although these methods can provide individual psychological state assessment, due to the long assessment cycle and poor intervention timeliness, they are often difficult to achieve early warning, and the assessment results are easily influenced by students self-report or external environment, resulting in low recognition accuracy (Guo, 2022; Firkey et al., 2022). Especially in large-scale universities, this method often fails to meet the

increasingly severe psychological health management needs due to high time and manpower costs and low efficiency. In recent years, with the rapid development of educational big data and machine learning technology, psychological anomaly recognition methods based on student behavioral data have gradually become a new research direction. Despite recent advancements, several critical research gaps remain unaddressed in the current literature, limiting the real-world effectiveness of these approaches (Zhong et al., 2022). Methodologically, there is a significant gap in effectively managing sample imbalance. Standard models frequently develop a bias toward the normal majority, resulting in high false-negative rates for at-risk students. Furthermore, a substantial data fusion gap exists concerning feature heterogeneity, as the naive combination of disparate financial, academic, and social data often obscures critical predictive patterns. Lastly, there is a distinct gap in model architecture and generalizability. Many prior studies rely on isolated data sources and simplistic frameworks that fail to capture the complex, non-linear interplay of student behaviors across different campus cultures (Quadir et al., 2022; Nti et al., 2022). Addressing these specific research gaps is the primary motivation for developing a more robust, multidimensional identification framework.

In response to these identified gaps, the Hybrid Model for Psychological Abnormal Student Behavior Identification (HMPABI) is designed with specific objectives to overcome such shortcomings. Its novelty is rooted in a comprehensive architecture that simultaneously addresses feature heterogeneity and sample imbalance. This is achieved by fusing multidimensional behavioral records with psychological assessment data to create a more holistic feature space, while incorporating the Synthetic Minority Over-sampling Technique (SMOTE) to enhance sensitivity toward the underrepresented "abnormal" category. Furthermore, architecture advances beyond simplistic designs by introducing an innovative hybrid classification framework that synergistically combines K-means clustering with a cascade of Logistic Regression (LR) and Support Vector Machine (SVM) classifiers, thereby improving predictive accuracy and reliability. A HMPABI is proposed to address the aforementioned issues. This model effectively solves the data imbalance and feature heterogeneity by integrating multidimensional behavioral features (such as student's consumption behavior, course grades, social activity, etc.) with psychological assessment data, and combining K-means clustering, SMOTE, LR, and SVM. The innovation of the HMPABI model lies in integrating systematic feature engineering with a hybrid machine learning architecture, enabling accurate identification of students with psychological abnormalities through multidimensional data fusion. It is expected to provide a new approach for universities that can improve the timeliness of psychological health monitoring and the accuracy of identification, and to provide theoretical and technical support for psychological health management and intervention in universities.

To explicitly position the proposed framework as an operational decision-support mechanism within educational service management rather than merely a predictive algorithm, the multi-source data inputs are directly aligned with institutional governance actions. Specifically, academic administration data, such as failing rates, serve to trigger early tutoring interventions and academic advisory workflows for underperforming individuals. Concurrently, one-card system metrics, including average daily expenses and breakfast frequencies, function as operational indicators to optimize targeted financial aid distribution and guide nutritional health monitoring initiatives. Furthermore, campus routine monitoring, driven by access control and regular library visits, enables the strategic allocation of psychological counseling resources. Finally, social isolation scores derived from co-occurrence matrices initiate campus peer-support programs and proactive professional check-ins, thereby systematically integrating behavioral analytics into comprehensive institutional risk management and targeted intervention planning.

While this study does not perform clinical diagnoses, the term "psychological abnormalities" is operationally understood in alignment with behavioral indicators frequently associated with common mental health disorders defined by established clinical standards such as the DSM-5 and ICD-11. These indicators include, but are not limited to, persistent changes in social interaction, sleep and eating patterns, and academic engagement that may signify underlying conditions such as depression and anxiety disorders.

2. Development of Psychological Anomaly Student Recognition Model based on Multi-Source Data Fusion

2.1. Feature Engineering

Before modeling the identification of students with psychological abnormalities, it is necessary to first prepare the data and extract behavioral features (Mcbride and Philippou, 2022). Based on student behavior data collected from multiple business systems at a certain university, this study constructs behavior feature vectors that reflect individual differences in student's psychological states through systematic data preprocessing.

The data used come from five systems: the one-card system, the access control system, the academic administration system, the library system, and the psychological assessment data. The one-card system records in detail the consumption information of student's during their school period, including consumption time, amount, and consumption location. The access control system collects time data for students entering specific locations such as libraries and dormitory buildings. The academic administration system provides students with basic information, course grades, and academic performance data, such as failing grades. The library system includes student borrowing book records and records of entering and exiting the library. The psychological assessment data section includes labels for students with confirmed psychological abnormalities identified by professional teachers, which can be used to train supervision models (Jeong et al., 2023). Before conducting data analysis, all data undergoes strict anonymization, retaining only necessary behavior fields and using student ID hash values as unique identifiers to ensure data security and privacy (Lindquist et al., 2023). Building upon this data privacy protocol, a comprehensive data governance framework was established to systematically address the ethical implications of predictive mental health monitoring. Ethical approval for this retrospective analysis of pre-existing datasets was secured in strict alignment with institutional administrative guidelines, confirming that no direct student interventions occurred during the model development phase. To proactively mitigate algorithmic bias, targeted resampling techniques

are structurally embedded within the analytical design to ensure equitable representation of minority behavioral profiles. Furthermore, the operational risk mitigation strategy dictates that the algorithm be specifically engineered to optimize both precision and recall, with its outputs strictly positioned as advisory decision-support metrics. Consequently, the designed governance workflow mandates that any subsequent institutional management actions be independently evaluated by certified counseling staff, thereby embedding systemic accountability directly into the implementation framework. The raw data contains a large number of missing values, duplicate items, and inconsistent formats. Therefore, the data preprocessing steps are shown in Fig. 1.

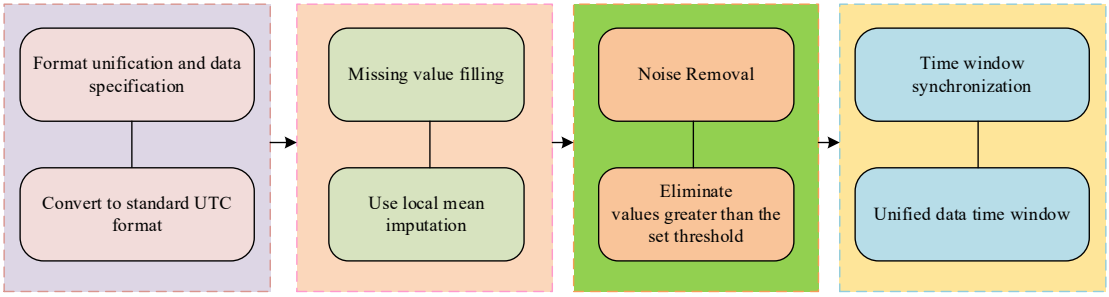

Fig. 1. Data preprocessing steps

As shown in Fig. 1, during the data preprocessing stage, the study addresses timestamp format inconsistencies by converting records from different systems to the standard Universal Time Coordinated (UTC) format and uniformly using Chinese Yuan Fen as the unit of measure to ensure data consistency and standardization. The local mean imputation method is used to make reasonable estimates for missing values in course grades, borrowing records, and other data. To eliminate noisy data from consumption records, the data distribution was analyzed to identify outliers. The threshold for a single transaction amount is 200 Renminbi (RMB), as values exceeding this point are very rare and disconnected from the central trend of student spending. This threshold closely aligns with the upper boundary for outliers identified using the standard Interquartile Range (IQR) method (Q3+1.5*IQR), providing a statistical basis for its selection. Consequently, any transaction amount exceeding this value is flagged as an outlier and removed to enhance overall data quality. In addition, to ensure comparability of data across systems in the time dimension, the time window is unified to semester units, allowing data from each system to be analyzed within the same time frame. To measure student's daily routines, the study takes the standard deviation of water fetching time as the evaluation criterion, as shown in Eq. (1).

$$\sigma_u^{water} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - \overline{t})^2} \tag{1}$$

In Eq. (1), σ_u^{water} represents the standard deviation of the water fetching time. t_i is the timestamp (in seconds) of the water fetching time on day i. $\overline{t}$ is the average of all water fetching times. n is the number of observation days. The regularity of daily life is shown in Eq. (2).

$$R_u = 1 - \frac{\sigma_u^{water}}{\max(\sigma^{water})} \tag{2}$$

In Eq. (2), R_u is the score for regularity in daily life, and the higher the score, the stronger the regularity in daily life (Lui, Sagar-Ouriaghli, and Brown, 2024). The diligence level of students is measured by the frequency of library visits and breakfast dining, and the frequency of library visits is shown in Eq. (3).

$$F_u^{lib} = \frac{N_u^{lib}}{D} \tag{3}$$

In Eq. (3), F_u^{lib} is the number of times a student u entered the library during the analysis period. N_u^{lib} represents the total number of times a student u entered the library during the analysis period. D represents the number of days. The breakfast check-in rate B_u is shown in Eq. (4).

$$B_u = \frac{N_u^{breakfast}}{D} \tag{4}$$

In Eq. (4), $N_u^{breakfast}$ represents the number of times a student u eats breakfast during the analysis period [13]. The

joint definition of the diligence comprehensive indicator C_u is shown in Eq. (5).

$$C_u = \omega_1 F_u^{lib} + \omega_2 B_u \tag{5}$$

In Eq. (5), ω_1 and ω_2 are normalized weight parameters, each set to 0.5 by default. Based on the time and location of card swiping, a social graph is constructed using a co-occurrence matrix. If two students swipe their cards at the same location within time interval $\Delta t \le 60s$, it is considered a social co-occurrence. The social intensity $S_{i,j}$ is shown in Eq. (6).

$$S_{i,j} = \frac{N_{i,j}^{co}}{N_i} \tag{6}$$

In Eq. (6), $N_{i,j}^{co}$ is the co-occurrence frequency of the student i and j. N_i is the total number of card swipes by student i. τ represents the social connection threshold. If $S_{i,j} > \tau$, there is a social connection between i and j. The degree D_u^{social} of a student's social network is the number of their friends, as shown in Eq. (7).

$$D_u^{social} = \sum_{j=1}^{N} \mathrm{I}(S_{u,j} > \tau) \tag{7}$$

The social isolation degree I_u is further defined, as shown in Eq. (8).

$$I_u = 1 - \frac{D_u^{social}}{\max(D^{social})} \tag{8}$$

In Eq. (8), I is the indicator function, with a value of 1 when the condition in parentheses is true. Otherwise, it is 0. The average grade $\overline{S}_u$ of all required courses for students is extracted, as shown in Eq. (9).

$$\overline{S}_u = \frac{1}{m} \sum_{k=1}^{m} s_k \tag{9}$$

In Eq. (9), s_k represents the grade of the student u in course k. m is the number of courses. The failure rate is shown in Eq. (10).

$$F_u^{fail} = \frac{N_u^{fail}}{m} \tag{10}$$

In Eq. (10), F_u^{fail} represents the failure rate. N_u^{fail} represents the number of failed courses. The daily average consumption E_u based on the student's consumption amount is shown in Eq. (11).

$$E_u = \frac{\sum_{i=1}^{n} e_i}{D} \tag{11}$$

In Eq. (11), e_i represents the i-th consumption amount, which is further annotated with the aid records and labels of impoverished students. Due to the inconsistent range of original feature values, which may affect the convergence effect of the model, the Min-Max normalization method is adopted, as shown in Eq. (12) (Chinchanikar and Shaikh, 2022).

$$x_u^{norm} = \frac{x_u - \min(x)}{\max(x) - \min(x)} \tag{12}$$

In addition, to reduce redundancy and dimension disasters, Principal Component Analysis (PCA) is used to extract the main behavioral factors, as shown in Fig. 2.

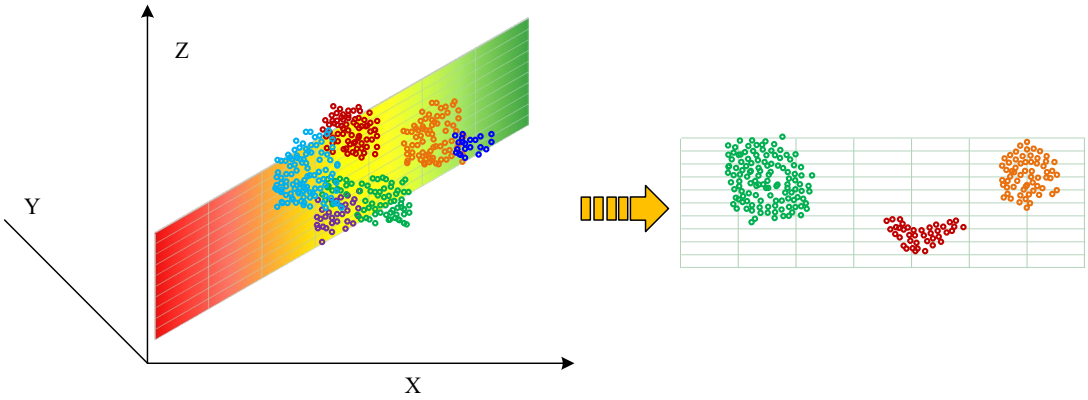

Fig. 2. Schematic diagram of PCA data dimensionality reduction mechanism

As shown in Fig. 2, the left side is the three-dimensional raw data space, where the data is mapped to a two-dimensional plane, and the structural features of the original data are preserved. The first k principal components with a cumulative explanatory variance of over 85% are retained, as shown in Eq. (13).

$$Z = XW_k \tag{13}$$

In Eq. (13), X is the original feature matrix. W_k is the first k principal component vectors. Behavioral feature extraction has recently been applied to the detection of psychological anomalies using multimodal data sources. Studies have shown that behavioral cues, including social interactions and digital activity patterns, can serve as effective indicators for detecting mental health risks. For example, machine learning techniques have been employed to analyze user-generated content and behavioral data from online platforms to proactively identify tendencies toward depression (Hemtanon et al., 2022). In parallel, system architectures designed for general behavioral recognition have demonstrated their applicability in mental health modeling by integrating structured and unstructured data features (Zhu et al., 2022). Additionally, forensic psychiatric research has highlighted the advantages of machine learning in uncovering complex behavioral dynamics related to psychological disorders (Hofmann et al., 2022). These findings provide empirical support for using behavioral features as valid inputs in models such as HMPABI (Vieira et al., 2022).

2.2. Identification of Psychological Abnormal Students based on Hybrid Model

After completing data preprocessing and feature extraction, a feature space with clear structure and significant differences is obtained through feature normalization and dimensionality reduction. However, due to the fact that students with psychological abnormalities only account for a small proportion of the real college student population, their labeled samples are far less than those of the psychologically normal group, resulting in a highly imbalanced distribution of the dataset. Therefore, to fully explore the potential structure of student's behavior characteristics in school and identify psychological abnormalities, the total number of samples in the entire dataset is N, with N_+ being the number of psychologically normal students and N_- being the number of psychologically abnormal students, and N_- □ N_+. The imbalance rate is defined in Eq. (14).

$$r = \frac{N_+}{N_-} \tag{14}$$

In actual data collection, r usually exceeds 30. Directly training the classification model will make the model tend to predict as a normal class, resulting in extremely high false negatives (Pallavicini et al., 2022). The cluster-based sampling strategy adopted in this work is supported by recent advances in behavior-driven machine learning for psychological anomaly detection. Reviewing predictive analysis methods underscores the need to address data imbalance and feature heterogeneity in mental health classification tasks (Islam et al., 2024). Clustering and passive sensing approaches have also been shown to be effective in multimodal behavioral monitoring, particularly within unsupervised or semi-supervised learning frameworks (Khoo et al., 2024). In addition, bio signal-based clustering methods have been recognized for their ability to enhance machine learning pipelines in mental health applications by capturing latent behavioral variations and uncertainty structures (Sajno et al., 2023). To solve this problem, K-Means clustering is used to select K cluster center samples from normal classes, and representative k sample points within each cluster are retained to ensure uniform distribution of the retained normal samples (Ikegwu et al., 2022; Scotta et al., 2022; Buizza et al., 2022), as shown in Eq. (15).

$$x_{\text{new}} = x_i + \delta \cdot (x_n - x_i) \tag{15}$$

In Eq. (15), x_i represents a minority class sample. x_n is a neighbor of x_i. $\delta \sim U(0,1)$ is a random number between 0 and 1. x_{new} is the generated new sample (Li et al., 2022). The flowchart for identifying psychologically abnormal students driven by K-means clustering is shown in Fig. 3.

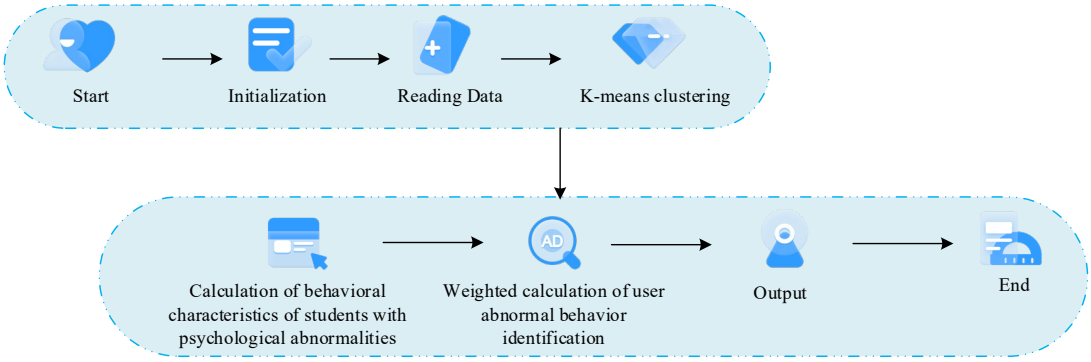

Fig. 3. Flow chart of identifying psychologically abnormal students driven by K-means clustering (Icons in the picture are sourced from: https://iconpark.oceanengine.com/illustrations/6)

As shown in Fig. 3, after reading student behavior data, the K-means algorithm is first used to cluster the samples and obtain preliminary clustering based on behavioral features. Subsequently, by combining clustering labels with extracted features of psychologically abnormal students, a weighted strategy is used to calculate the probability score of abnormalities, identifying user behavior abnormalities. The final output prediction results are used to label high-risk students and achieve a closed-loop identification. SMOTE is used to generate synthetic samples from abnormal-class samples and to fill the sample space. The SMOTE algorithm is shown in Fig. 4.

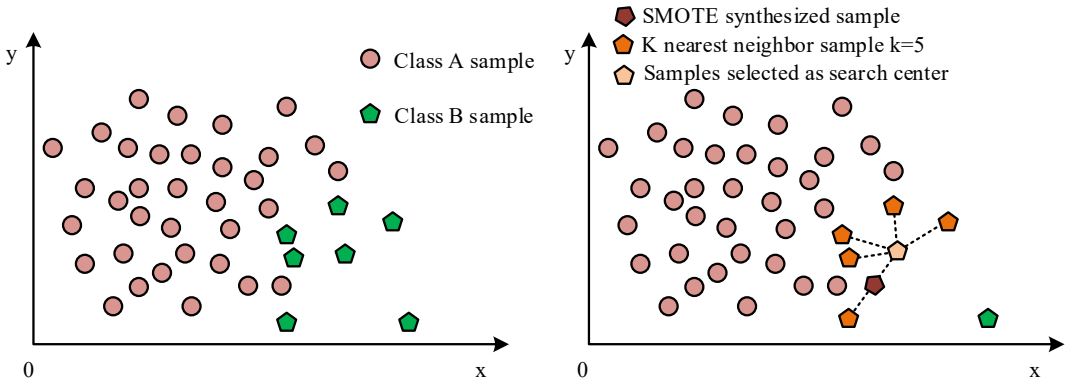

Fig. 4. SMOTE algorithm

To further improve recognition ability, a hybrid classification model structure combining LR and SVM is designed. The predictive function of the LR model is shown in Eq. (16).

$$P(y=1 \mid x)=\frac{1}{1+e^{-(\mathbf{w}^T x+b)}} \tag{16}$$

In Eq. (16), $x \in \mathbb{R}^d$ is the input feature vector. $\boldsymbol{w} \in \mathbb{R}^d$ is the model weight vector. $b \in \mathbb{R}^d$ is the bias term. The output probability can be interpreted as the risk score for students with psychological abnormalities. The training objective is to minimize the logarithmic loss function with a regularization term, as shown in Eq. (17).

$$L_{\text{LR}}(\mathbf{w},b)=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i)+(1-y_i)\log(1-p_i)\right]+\lambda \ \|\mathbf{w}\|_2^2 \tag{17}$$

In Eq. (17), $y_i \in \{0,1\}$ is the true label of the sample. λ is the regularization strength parameter. Due to the scarcity of abnormal class samples, SVM is used as a supplementary component to improve the model's ability to recognize "abnormal boundaries" (Rotanov et al., 2023). The SVM model training process is shown in Fig. 5.

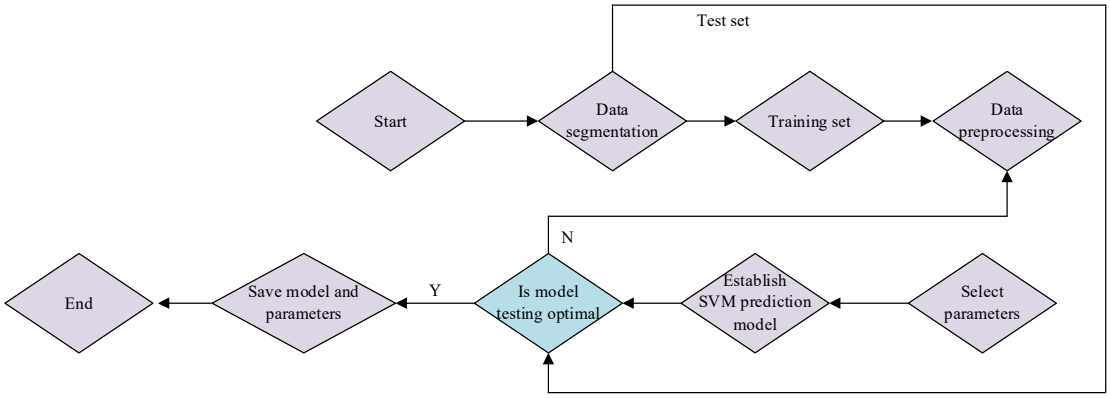

Fig. 5. SVM model training process

As shown in Fig. 5, starting from the starting node, data segmentation is performed first, dividing the data into a training set and a testing set. After data preprocessing, the training set selects appropriate parameters to establish an SVM prediction model. The test set is used to test the model and determine whether the test results are optimal. If the optimal conditions are not met, the parameters are re-selected, and the model is adjusted. If the test results reach the optimal level, the model and related parameters are saved. Whether the new sample falls outside the boundary is determined by Eq. (18).

$$\begin{cases} \min_{\mathbf{w},\xi_i,\rho} \dfrac{1}{2}\|\mathbf{w}\|^2 + \dfrac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho \\ \text{s.t.}\mathbf{w}^T\phi(x_i) \geq \rho - \xi_i, \quad \xi_i \geq 0 \end{cases} \tag{18}$$

In Eq. (18), $\phi(x_i)$ is the kernel function mapping. ν can control the proportion of abnormal samples. ρ is the decision boundary threshold. ξ_i is a relaxation variable that allows for error. To bridge the gap between technical prediction and operational management, the predictive outputs of the HMPABI models are systematically integrated into a multi-tiered institutional intervention pathway. Specifically, when a student's calculated risk score exceeds a dynamically calibrated threshold, the system automatically generates an alert directed to designated academic advisors. This triggers a preliminary, non-invasive assessment to contextualize the flagged behavioral deviations within the student's daily academic life. Subsequently, validated high-risk cases are formally escalated to the institutional counseling center for professional evaluation and intervention planning. This structured workflow seamlessly translates raw predictive analytics into actionable management protocols.

3. Model Comparison And Behavioral Feature Sensitivity Verification Experiment

To ensure the transparency and replicability of the experiments, all models are evaluated under the same hardware environment and data partitioning strategy. The experiment employs the extracted multi-source behavior feature vectors and a normalized feature space. The datasets, specifically the Urumqi University Student Campus Behavior Dataset (UUS-CBD) and the Adolescent Mental Health and Behavior Dataset (AMHBD), contain highly sensitive personal behavioral and psychological information. To strictly comply with institutional data privacy regulations and ethical governance protocols, the raw data are not publicly hosted on open repositories. However, de-identified data subsets supporting the conclusions of this article can be made available by the corresponding author upon reasonable academic request, subject to mandatory non-disclosure agreements. The detailed experimental setup and key hyperparameter settings for the HMPABI model are summarized in Table 1. These settings are kept consistent across all relevant training and testing procedures to ensure a fair comparison.

To comprehensively evaluate the proposed HMPABI model, it is compared against several baseline models representing different state-of-the-art approaches adapted from established architectures in recent references. The Temporal Deep Anomaly Detection Model (T-DADM) is representative of methods that specialize in identifying deviations within time-series behavioral data, drawing upon foundational deep learning frameworks for sequential anomaly detection (Choi et al., 2021). Graph-based models, including the Graph Attention Network for Student Anomaly Detection (GAT-SAD) based on multivariate graph deviation structures (Deng and Hooi, 2021), and the Multi-View Behavioral Sequence Graph Neural Network (MBS-GNN) adapted from multi-behavior relational models (Xia et al., 2023), focus on modeling students within a social network structure to find relational or community-based anomalies. Furthermore, the Deep Ensemble Model for Psychological Stress Detection (DEM-PSD) stands for ensemble methods that aggregate predictions from multiple algorithms to improve overall robustness in mental health screening tasks (Bobade and Vani, 2020).

Fig. 6(a) shows the comparison results of MAE. As the time window T increases from 5 minutes to 45 minutes, the MAE of each model shows an overall upward trend. However, HMPABI maintains the lowest error value at all time points, indicating that the model had higher prediction accuracy and stability in abnormal behavior recognition tasks. Fig. 6(b) shows the MAPE indicators. The results are consistent with MAE. HMPABI consistently outperforms other models throughout the entire testing range, with the smallest increase in error, demonstrating good robustness and generalization

ability. The study continues to test the convergence speed and performance stability of different models at different training stages, as shown in Fig. 7.

Table 1. Summary of Experimental Setup and Hyperparameter Settings

Category	Parameter	Value/Setting
Environment and Data	Hardware Environment	NVIDIA RTX 3090 (24GB)
	Datasets	UUS-CBD, AMHBD
	Data Split (Train/Val/Test)	70%/15%/15%
Preprocessing and Sampling	PCA Variance Threshold	85%
	K-Means Clusters (k) for undersampling	Determined empirically via elbow method
	SMOTE k_neighbors for Oversampling	5
HMPABI Model and Training	Optimizer	Adam
	Learning Rate	0.001
	Training Epochs	10
	Batch Size	32
	Loss Function	Cross-Entropy with L2 Regularization
	SVM Kernel	Radial Basis Function (RBF)

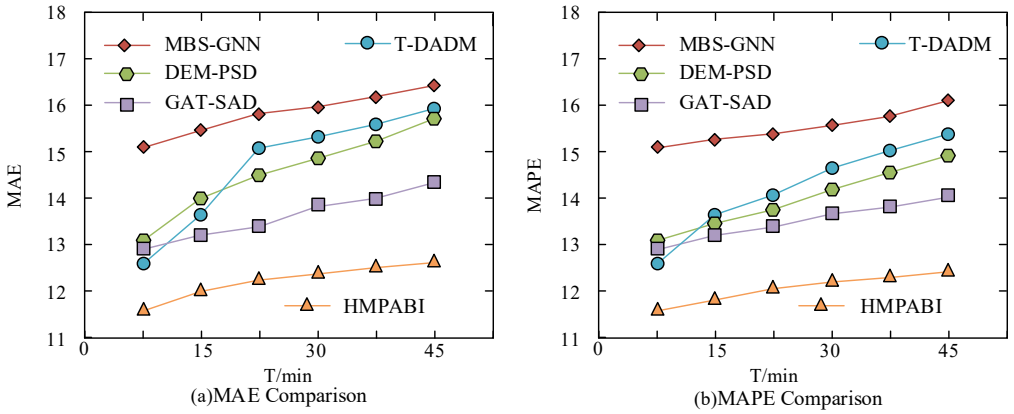

Fig. 6. Comparison of predictive performance of different models under different time windows

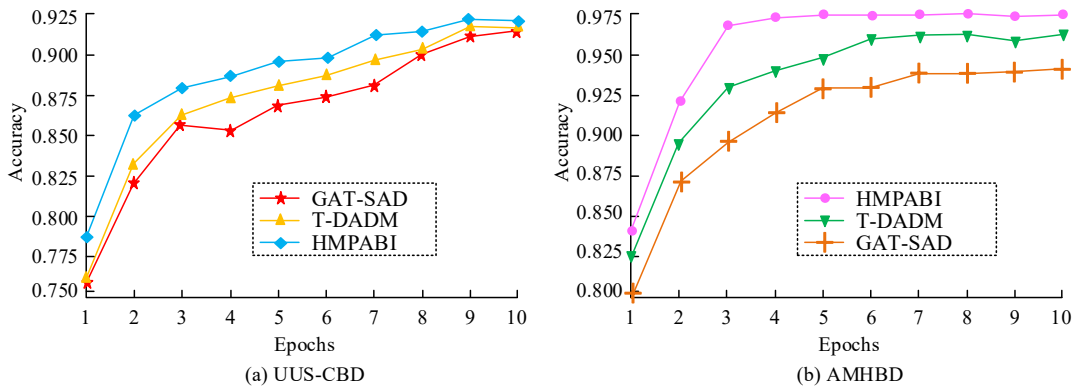

Fig. 7. Convergence speed and performance stability of different models at different training stages

On the UUS-CBD dataset shown in Fig. 7(a), as the training epoch increases, the accuracy of the three models gradually improves. HMPABI shows high initial accuracy from the initial stage, quickly approaching and stabilizing in the high accuracy range after the 6th round, and finally reaching the highest level in the 10th round, outperforming other comparison models. It has a faster learning ability and stronger generalization performance in campus behavior data modeling tasks. The test results on the AMHBD dataset in Fig. 7(b) further validate the advantages of HMPABI. The model quickly approaches optimal accuracy after the second round of training and maintains a leading position in subsequent rounds, achieving higher recognition accuracy and faster convergence than T-DADM and GAT-SAD. This indicates that HMPABI

has greater stability and accuracy in the recognition task and higher practical value in applications. The comparative analysis results of identifying psychological abnormalities under different student behavioral characteristics are shown in Table 2.

Table 2. Recognition of psychological anomalies across student behaviors

Student ID	Avg. Daily Expense (CNY)	Breakfast Frequency (%)	Library Visits (per week)	Social Isolation Score	Failing Rate (%)	Regularity Score	HMPABI Prediction	Manual Annotation
S1004	13.28	7.69	0.4	0.91	42.86	0.238	Abnormal	Abnormal
S1021	25.91	53.85	3.1	0.32	7.14	0.612	Normal	Normal
S1042	8.67	3.85	0.3	0.97	57.14	0.127	Abnormal	Abnormal
S1085	19.74	38.46	2.7	0.41	14.29	0.488	Normal	Normal
S1103	10.02	11.54	0.5	0.86	35.71	0.196	Abnormal	Abnormal
S1129	22.89	73.08	4.4	0.15	0.00	0.758	Normal	Normal
S1150	12.45	15.38	1.0	0.72	21.43	0.361	Abnormal	Normal
S1187	27.56	57.69	2.8	0.34	0.00	0.625	Normal	Normal
S1222	9.84	19.23	0.7	0.79	28.57	0.284	Abnormal	Abnormal

The sensitivity results presented in Table 2 reveal the model's reliance on a holistic interpretation of behavioral shifts. The model's prediction is not sensitive to minor fluctuations but reacts to significant perturbations in key indicators, such as library visit frequency. For student S1301, the sharp decline in library visits from 3.6 to 0.8 per week does more than signal a drop in diligence. It also corresponds to a dramatic increase in the social isolation score from 0.33 to 0.89. The HMPBAI model, as a multivariate system, interprets the fusion of negative signals, namely decreased academic engagement and increased social withdrawal, as a significant deviation from the normal pattern, thereby transforming its prediction from "normal" to "abnormal". Conversely, when a student's behavior improves dramatically, as with S1330 whose library visits increased from 0.9 to 3.0, the model readjusts its prediction to "normal", showcasing its ability to recognize positive changes. Specifically, the term "Post-Perturbation Prediction" denotes the updated classification output generated by the model after specific behavioral input variables, such as library visit frequency, are artificially altered. By comparing this newly generated outcome with the original prediction, the analysis directly illustrates the model's dynamic responsiveness and sensitivity to critical shifts in individual student behavior. The sensitivity analysis of student's psychological recognition under different behavioral variable disturbances is shown in Table 3.

Table 3. Sensitivity analysis of students psychological recognition under different behavioral variable disturbances

Student ID	Original Library Visits (per week)	After Perturbation	Original Prediction	Post-Perturbation Prediction	Change in Frequency	Original Isolation Score	Post-Isolation Score
S1301	3.6	0.8	Normal	Abnormal	Significant drop	0.33	0.89
S1314	2.2	2.1	Normal	Normal	Slight drop	0.44	0.47
S1330	0.9	3.0	Abnormal	Normal	Significant rise	0.81	0.30
S1342	4.1	4.3	Normal	Normal	Stable	0.26	0.24
S1355	1.0	1.0	Abnormal	Abnormal	No change	0.91	0.93
S1376	3.2	0.9	Normal	Abnormal	Clear drop	0.36	0.84
S1389	2.5	2.6	Normal	Normal	Slight rise	0.41	0.40
S1400	0.7	3.5	Abnormal	Normal	Significant rise	0.88	0.29
S1418	1.8	1.0	Normal	Abnormal	Drop	0.49	0.77
S1432	3.9	3.7	Normal	Normal	Stable	0.27	0.28

To further enhance persuasiveness, Table 3 provides specific examples of how HMPABI calculates abnormality risk scores based on a combination of key behavioral inputs. These cases serve as concrete validation of the model's decision-

making logic. For example, student S3053 exhibits multiple high-risk indicators: an extremely low breakfast frequency (7.7%), a very high failing rate (57.1%), and minimal library visits. The model aggregates these negative factors, resulting in a high abnormal risk score of 0.93. In stark contrast, student S3011 displays consistently positive behaviors across these metrics, resulting in a very low risk score of 0.17. This side-by-side comparison demonstrates that the HMPABI model does not rely on a single variable but rather predicts based on the collective weight of evidence across the entire student behavior profile, providing a more robust and detailed evaluation. The student behavior characteristics and HMPABI model abnormal risk score are shown in Table 4.

Table 4. Student behavioral characteristics and HMPABI model abnormal risk score

Student ID	Breakfast Frequency (%)	Failing Rate (%)	Library Visits (per week)	Abnormality Risk Score
S3004	11.5	42.9	0.4	0.86
S3011	69.2	0.0	3.9	0.17
S3027	23.1	28.6	1.1	0.58
S3053	7.7	57.1	0.2	0.93
S3079	46.2	14.3	2.6	0.32

Table 4 shows the relationship between some key behavioral characteristics of five students (breakfast frequency, failure rate, and library visit frequency) and the probability score of psychological abnormalities output by the HMPABI model. The more irregular the behavior, the more severe the failure, and the lower the library visit frequency, the higher the probability score predicted by the model for psychological abnormalities. For example, the failure rate of S3053 was 57.1%, the breakfast frequency was extremely low, at 7.7%, and the final abnormal probability determined by the model was 0.93, indicating a high risk. In contrast, S3011 exhibited positive behavior and was judged by the model to have a normal psychological state.

To further substantiate the quantitative superiority of the HMPABI model, a statistical significance analysis was conducted. The model was evaluated over multiple independent runs to account for variations stemming from data partitioning and stochastic model initialization. In addition to accuracy and MAPE, the key classification metrics of Precision and Recall were also assessed. An independent samples t-test was performed to compare the performance of HMPABI against each baseline model. The comprehensive results, including the mean, standard deviation (Std. Dev.), and the resultant *p*-values from the significance tests on the AMHBD dataset, are presented in Table 5.

Table 5. Statistical Comparison of Model Performance on the AMHBD Dataset

Model	Accuracy (%) (Mean ± Std. Dev.)	Precision (%) (Mean ± Std. Dev.)	Recall (%) (Mean ± Std. Dev.)	*p*-value (vs. HMPABI)
HMPABI	/	96.27 ± 0.88	95.15 ± 1.10	97.02 ± 0.95
T-DADM	93.55 ± 1.21	91.80 ± 1.35	94.13 ± 1.28	< 0.01
GAT-SAD	91.34 ± 1.53	89.52 ± 1.68	92.44 ± 1.45	< 0.01

As shown in Table 5, the analysis reveals that HMPABI not only achieves higher mean performance across all key metrics but also shows statistically significant improvements. The *p*-value is well below the standard threshold of 0.05, indicating that HMPABI's excellent performance is not a random or accidental result. This provides strong statistical evidence that the architectural design of the HMPABI model leads to a demonstrably more effective and reliable solution for identifying students with psychological abnormalities compared to the other models. In order to further validate the robustness, long-term stability, and contributions of each component of the proposed model, a series of targeted experiments was conducted. These included noise-injection tests, longitudinal data stability evaluation, and a comprehensive ablation study. For the noise injection test, Gaussian noise (mean=0, std.dev=0.1) was added to all normalized continuous features to simulate real-world data imperfections. For the longitudinal test, the model's performance was evaluated on a continuous three-month data stream to assess its stability over time. The ablation study involved systematically deactivating or replacing key components within the HMPABI architecture to quantify their impact on overall performance. The results of these experiments are summarized in Table 6.

Table 6 clearly demonstrates the synergistic contribution of each component within the HMPABI framework. Removing the SMOTE module (HMPABI w/o SMOTE) caused a drastic drop in recall from 97.02% to 86.43%, highlighting its critical role in addressing class imbalance and ensuring that the model can effectively identify a small number of "abnormal" classes. Deactivating other components, such as the K-Means-based undersampling or the LR classifier, also led to a noticeable degradation in overall performance, confirming that the integrated hybrid design is superior to its constituent parts alone.

4. Discussion

The significance of this study can be understood from two aspects: its theoretical and practical contributions. From a theoretical perspective, this study proposes the HMPABI model, a novel hybrid framework that systematically addresses

the persistent challenges of sample imbalance and feature heterogeneity in educational data mining. By integrating clustering, oversampling, and a mixed classification strategy, it provides a new methodological reference for analyzing complex, multi-source behavioral data in related academic fields. From a practical perspective, the developed model serves as a valuable tool for higher education institutions. It offers a data-driven, automated early warning system that can help counseling centers and student affairs departments to identify at-risk students proactively. This efficiently allocates limited mental health resources and facilitates a shift from reactive to preventative student support, holding significant potential to improve the timeliness and effectiveness of campus mental health services. The practical contribution of this research fundamentally lies in detecting psychological risks through continuous behavioral analytics. If effectively operationalized within educational institutions, the system is designed to seamlessly support university mental health monitoring programs, facilitate targeted counseling interventions, enhance overall institutional risk management, and optimize the strategic planning of resource allocation.

Table 6. Results of Robustness, Longitudinal, and Ablation Experiments

Model/Condition	Accuracy (%)	Precision (%)	Recall (%)	Description
HMPABI (Full Model)	96.27	95.15	97.02	The complete proposed model.
HMPABI w/o SMOTE	92.15	97.55	86.43	Oversampling module removed.
HMPABI w/o K-Means	94.08	93.12	94.88	Undersampling via random selection instead of clustering.
HMPABI (SVM only)	93.81	92.90	94.50	LR component removed.
HMPABI + 10% Noise	95.16	94.20	95.85	Performance with noise injected into features.
HMPABI (3-Month Data)	96.05	94.98	96.81	Performance on a 3-month longitudinal dataset.

In practical terms, the proposed model could be embedded into existing college psychological support systems to serve as a decision-support tool for early intervention. Once the HMPBAI framework is integrated into the university's student behavior monitoring platform, it can regularly evaluate anonymous behavior data and generate a ranking list of high-risk individuals based on dynamic abnormal risk scores. For example, students with declining behavioral regularity, reduced social engagement, or rising failure rates can be flagged and discreetly offered assistance. This application not only improves resource allocation but also aligns with ethical standards by avoiding invasive measures while enhancing early detection capabilities. This integration will provide actionable intelligence for academic advisors, bridging the gap between passive observation and proactive mental health support. To operationalize these predictions within institutional management, a structured intervention workflow is established. Initially, the system generates automated alerts for students exceeding specific risk thresholds. Subsequently, academic advisors utilize these alerts to conduct preliminary, non-invasive check-ins to contextualize behavioral data. Finally, cases requiring further attention are formally escalated to university counseling centers for comprehensive evaluation, thereby systematically integrating data-driven risk monitoring into established educational management and administrative decision-making protocols.

To ensure responsible, practical implementation and effective governance, the system requires dynamically calibrated decision thresholds to trigger graduated intervention alerts. Crucially, algorithmic outputs serve solely as advisory tools, continuous human oversight and professional validation by certified counseling staff remain universally mandatory before any management actions are taken. Given the profound sensitivity of mental health information, stringent ethical safeguards and privacy protection protocols, such as robust data anonymization and strictly audited role-based access, are structurally embedded to comply with overarching institutional policies. Furthermore, by directly leveraging pre-existing campus big data infrastructures, the proposed framework maintains a low implementation cost and demonstrates high scalability, offering a sustainable, policy-aligned solution for broad educational deployment.

Implementing predictive analytics for student welfare governance requires a rigorous ethical and risk-management framework. To ensure data privacy and comply with ethical approval procedures for retrospective research, all behavioral records are irreversibly anonymized using cryptographic hashing prior to modeling. Algorithmic bias and fairness are proactively addressed through the structural integration of undersampling and oversampling techniques, thereby mitigating dataset imbalances and ensuring equitable representation of minority behavioral patterns. Furthermore, algorithmic decision accountability is established by addressing the consequences of misclassification; a high model precision of 95.15% limits false positives and unnecessary administrative scrutiny, while a high recall of 97.02% minimizes false negatives to

prevent overlooking high-risk individuals. Ultimately, these outputs serve strictly as advisory decision-support metrics, ensuring that institutional risk responsibility and final welfare interventions remain the exclusive domain of certified counseling professionals.

Future research could further enhance the robustness and adaptability of the proposed framework by integrating additional multimodal data sources. These may include real-time sensor data, natural language student feedback, social media interactions, and physiological indicators such as heart rate and sleep patterns. Including such heterogeneous data will enable a more comprehensive simulation of students psychological states and allow for finer grained anomaly detection. Moreover, developing fusion strategies that effectively combine structured, semi-structured, and unstructured data will be critical to advancing the generalizability of the HMPABI model across diverse educational and cultural contexts.

To extend the applicability of these research findings beyond a local context to a global scale, the proposed analytical framework must be adapted to accommodate diverse cultural and institutional variations. While specific student behaviors, campus infrastructure, and data collection systems differ significantly across countries, the core methodology of fusing multi-source heterogeneous data remains universally applicable. When deploying this model in different geographical regions, educational administrators should dynamically calibrate the behavioral feature inputs and risk decision thresholds to reflect local cultural norms, such as varied commuting habits and distinct social engagement patterns. Furthermore, international implementation must strictly align with regional data privacy regulations, such as the General Data Protection Regulation (GDPR). By adopting a flexible, culturally responsive approach to feature engineering, the underlying machine learning architecture can be successfully scaled and seamlessly integrated into university mental health governance systems worldwide.

5. Conclusion

With the increasingly serious mental health problems among college students, identifying them early through effective technological means has become an urgent problem to be solved. Therefore, the study proposed to construct an efficient and accurate model for identifying students with psychological abnormalities by analyzing their behavioral characteristics based on Multi-source Heterogeneous Educational Data and combining them with psychological assessment data, namely the HMPABI. The study combined K-means clustering, the SMOTE oversampling technique, and an SVM hybrid classification strategy. The research results indicated that the HMPABI model performed excellently across different time windows on UUS-CBD. Specifically, the HMPABI model had MAE of 0.18, 0.22, 0.26, and 0.28 at 5, 15, 30, and 45-minute windows, respectively, consistently maintaining the lowest error level. Compared to other models, the error was significantly lower. In terms of MAPE, HMPABI outperformed the comparison models, especially in the 45-minute window, where its MAPE was 2.5%, compared to 4.2% and 3.9% for T-DADM and GAT-SAD, respectively. In addition, the performance on the AMHBD dataset further validated the advantages of HMPABI, with an accuracy of 96.27%, far exceeding other comparison models. The T-DADM model achieved 93.55% accuracy, and GAT-SAD achieved 91.34%. This result indicates that HMPABI not only demonstrates strong performance in the fusion of behavioral and psychological assessment data but also has high stability and generalization. However, the research still has some shortcomings. Firstly, due to the limited dataset, the model is trained and tested only on specific student groups. Furthermore, while the current study relies on quantitative data, future work could be significantly enriched by incorporating qualitative inputs. Methods such as structured interviews or contextual surveys with students could provide deeper insights into the motivations and circumstances behind observed behavioral changes. This fusion of quantitative and qualitative data would not only enhance the model's explanatory power but also help validate its findings against lived experiences, leading to a more nuanced and holistic understanding of the link between behavior and psychological well-being.

Author Contributions

Hansong Dong contributed to conceptualization, methodology, software use, and manuscript writing. Jiao Huang contributed to conceptualization, methodology, and manuscript revision.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

A generative AI tool, Grammarly, was utilized strictly for language polishing, grammatical correction, and phrasing improvements to enhance the manuscript's readability. All research design, data analysis, and intellectual content remain the entirely original work of the authors.

References

Bobade, P., and Vani, M. (2020). Stress detection with machine learning and deep learning using multimodal physiological data. *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA*), 51–57.

Buizza, C., Bazzoli, L., and Ghilardi, A. (2022). Changes in college students' mental health and lifestyle during the COVID-19 pandemic: A systematic review of longitudinal studies. *Adolescent Research Review,* 7(4), 537–550.

Chen, C., Wu, Y., Li, J., Wang, X., Zeng, Z., Xu, J., and Xia, R. (2023). TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. *Molecular Plant,* 16(11), 1733–1742.

Chinchanikar, S., and Shaikh, A. A. (2022). A review on machine learning, big data analytics, and design for additive manufacturing for aerospace applications. *Journal of Materials Engineering and Performance,* 31(8), 6112–6130.

Choi, K., Yi, J., Park, C., and Yoon, S. (2021). Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. *IEEE Access,* 9, 120043–120065.

Deng, A., and Hooi, B. (2021). Graph neural network-based anomaly detection in multivariate time series. *Proceedings of the AAAI Conference on Artificial Intelligence,* 35(5), 4027–4035.

Firkey, M. K., Sheinfil, A. Z., and Woolf-King, S. E. (2022). Substance use, sexual behavior, and general well-being of US college students during the COVID-19 pandemic: A brief report. *Journal of American College Health,* 70(8), 2270–2275.

Guo, J. (2022). Deep learning approach to text analysis for human emotion detection from big data. *Journal of Intelligent Systems,* 31(1), 113–126.

Hemtanon, S., Aekwarangkoon, S., and Kittiphattanabawon, N. (2022). Proactive depression detection from Facebook text and behavior data. *International Journal of Electrical and Computer Engineering,* 12(5), 5027–5035.

Hofmann, L., Lau, S., and Kirchebner, J. (2022). Advantages of machine learning in forensic psychiatric research—Uncovering the complexities of aggressive behavior in schizophrenia. *Applied Sciences,* 12(2), Article 819.

Ikegwu, A. C., Nweke, H. F., Anikwe, C. V., Alo, U. R., and Okonkwo, O. R. (2022). Big data analytics for data-driven industry: A review of data sources, tools, challenges, solutions, and research directions. *Cluster Computing,* 25(5), 3343–3387.

Islam, M. M., Hassan, S., Akter, S., Jibon, F. A., and Sahidullah, M. (2024). A comprehensive review of predictive analytics models for mental illness using machine learning algorithms. *Healthcare Analytics*, Article 100350.

Jeong, S., Aymerich-Franch, L., Arias, K., Alghowinem, S., Lapedriza, A., Picard, R., and Breazeal, C. (2023). Deploying a robotic positive psychology coach to improve college students' psychological well-being. *User Modeling and User-Adapted Interaction,* 33(2), 571–615.

Khoo, L. S., Lim, M. K., Chong, C. Y., and McNaney, R. (2024). Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches. *Sensors,* 24(2), 348.

Kurcer, M. A., Erdogan, Z., and Cakir Kardes, V. (2022). The effect of the COVID-19 pandemic on health anxiety and cyberchondria levels of university students. *Perspectives in Psychiatric Care,* 58(1), 132–140.

Li, M., Zhang, J., Song, J., Li, Z., and Lu, S. (2022). A clinical-oriented non-severe depression diagnosis method based on cognitive behavior of emotional conflict. *IEEE Transactions on Computational Social Systems,* 10(1), 131–141.

Lindquist, E. G., Villarosa-Hurlocker, M. C., Raposa, E. B., Pearson, M. R., and Bravo, A. J. (2023). Fear of negative evaluation and suicidal ideation among college students: The moderating role of impulsivity-like traits. *Journal of American College Health*, 71(2), 396–402.

Lui, J. C., Sagar-Ouriaghli, I., and Brown, J. S. (2024). Barriers and facilitators to help-seeking for common mental disorders among university students: A systematic review. *Journal of American College Health,* 72(8), 2605–2613.

McBride, K., and Philippou, C. (2022). "Big results require big ambitions": Big data, data analytics and accounting in masters courses. *Accounting Research Journal,* 35(1), 71–100.

Nti, I. K., Quarcoo, J. A., Aning, J., and Fosu, G. K. (2022). A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. *Big Data Mining and Analytics,* 5(2), 81–97.

Pallavicini, F., Pepe, A., and Mantovani, F. (2022). The effects of playing video games on stress, anxiety, depression, loneliness, and gaming disorder during the early stages of the COVID-19 pandemic: A PRISMA systematic review. *Cyberpsychology, Behavior, and Social Networking,* 25(6), 334–354.

Quadir, B., Chen, N. S., and Isaias, P. (2022). Analyzing the educational goals, problems and techniques used in educational big data research from 2010 to 2018. *Interactive Learning Environments,* 30(8), 1539–1555.

Rotanov, A., Karshiyev, Z., Sharapova, D., and Shernazarov, F. (2023). Diagnosis of depressive and suicidal spectrum disorders in students of a secondary special education institution. *Science and Innovation,* 2(D11), 309–315.

Sajno, E., Bartolotta, S., Tuena, C., Cipresso, P., Pedroli, E., and Riva, G. (2023). Machine learning in biosignals processing for mental health: A narrative review. *Frontiers in Psychology,* 13, 1066317.

Scotta, A. V., Cortez, M. V., and Miranda, A. R. (2022). Insomnia is associated with worry, cognitive avoidance and low academic engagement in Argentinian university students during the COVID-19 social isolation. *Psychology, Health and Medicine,* 27(1), 199–214.

Vieira, S., Liang, X., Guiomar, R., and Mechelli, A. (2022). Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. *Clinical Psychology Review,* 97, 102193.

Xia, L., Huang, C., Xu, Y., Dai, P., Bo, L., Zhang, X., and Chen, T. (2023). Multi-behavior graph neural networks for recommender systems. *IEEE Transactions on Knowledge and Data Engineering,* 35(11), 11462–11475.

Yang, Z., and Ge, Z. (2022). On paradigm of industrial big data analytics: From evolution to revolution. *IEEE Transactions on Industrial Informatics,* 18(12), 8373–8388.

Zhong, Y., Chen, L., Dan, C., and Rezaeipanah, A. (2022). A systematic survey of data mining and big data analysis in internet of things. *The Journal of Supercomputing,* 78(17), 18405–18453.

Zhu, J., Goyal, S., Verma, C., Răboacă, M., and Mihaltan, T. C. (2022). Machine learning human behavior detection mechanism based on Python architecture. *Mathematics,* 10(17), 3159.

Hansong Dong obtained his Ph.D. in Business Administration (2023) from the University of San Carlos. Presently, he is the head of the Comprehensive Department of Economics and Management at Sichuan University of Architectural Technology. He has been hired as a member of the expert database and has given multiple technical lectures on integrating student management and software engineering. Dr. Dong has reviewed nearly 100 manuscripts and proposals. He has contributed to over 20 academic papers published across business administration, software engineering, and information security.

Huang Jiao obtained a master's degree in Software Engineering from the University of Electronic Science and Technology of China (2014). She currently serves as the head of the Student Management Department at the International School of Technical Education, Sichuan University of Architectural Technology, and has published over 20 papers across software engineering, information security, and student management.