Data Fusion Model for Psychological Crisis Early Warning System Using Data Mining Techniques

Yan Wu
Department of Students' Affairs, Anhui Post and Telecommunication College, Hefei 230031, China
E-mail: wuyan11032023@163.com

Keywords: college students, psychological crisis, data fusion model, data mining

Received: October 5, 2024

The mental well-being of university students has garnered significant attention from various sectors of society. Psychological challenges can disrupt educational institutions' harmonious progress, necessitating early intervention. Psychological disorders have the potential to hinder the smooth functioning of institutions of higher education, and their immediate intervention is necessary. The study proposes a notification system for identifying psychological crises among college students based on a data fusion approach. This research implements a range of machine learning methods, such as decision trees, random forests, logistic regression, and the Apriori algorithm. These methods utilize psychological profile data and survey questionnaires, which are subjected to encoding and feature selection for preprocessing. The suggested data fusion approach enhances the accuracy of prediction through the combination of the various methods, achieving a high F1 score of 82.2% compared to individual methods. The results highlight the effectiveness of data fusion in identifying psychological crises and, hence, providing a holistic platform for the early intervention of students' mental health issues.

Povzetek: Prispevek predstavi model za zgodnje opozarjanje na psihološke krize študentov, ki z združevanjem metod (RF, LR, Apriori) izboljša kvaliteto in omogoča pravočasne intervencije.

1 Introduction

The spread of digital resources in academic environments has intensified psychological distress among college students, mainly due to issues like academic pressure, social seclusion, and economic stressors. The predominance of mental illnesses, such as depression and anxiety, highlights the need for early intervention. Current detection methods are mainly reliant on survey-based assessments and single machine learning models, like decision trees and neural networks; however, these methods face remarkable challenges. Single-model approaches have limitations in that they rely on specific data sources, struggle with imbalanced datasets that produce high false-negative rates, and lack the ability to identify complex behavioral patterns. For instance, while random forest models have strong classification abilities, they do not provide insight into underlying relationships between behaviors and mental health risk. Therefore, there is a need for a more integrated approach to better identify psychological crises [1], [2], [3]. According to psychologist Kosinski, M.D., a psychological crisis is a physical indication brought about by an individual's incapacity to effectively address unexpected catastrophes, significant life changes, or intense psychological strain [4]. A psychological crisis arises from an individual's inability to generate appropriate responses in the face of challenges, leading to emotional, psychological, and behavioral abnormalities. Figure 1 illustrates the production and progression of psychological crises. Research indicates that a significant number of university students exhibit extreme behaviors as a result of psychological issues, such as running away from home, self-harm, and even suicide. These actions not only harm the students themselves and their families but also disrupt social stability. Furthermore, students' psychological crises can impede teachers' classroom instruction and harm educational institutions' overall harmonious and stable advancement [5], [6]. Hence, the utilization of modern science, technology, and the Internet, along with the evaluation and early detection of mental health issues among current college students and strategic analyses of their psychological concerns, can contribute to fostering a positive and healthy college experience for students [7], [8].

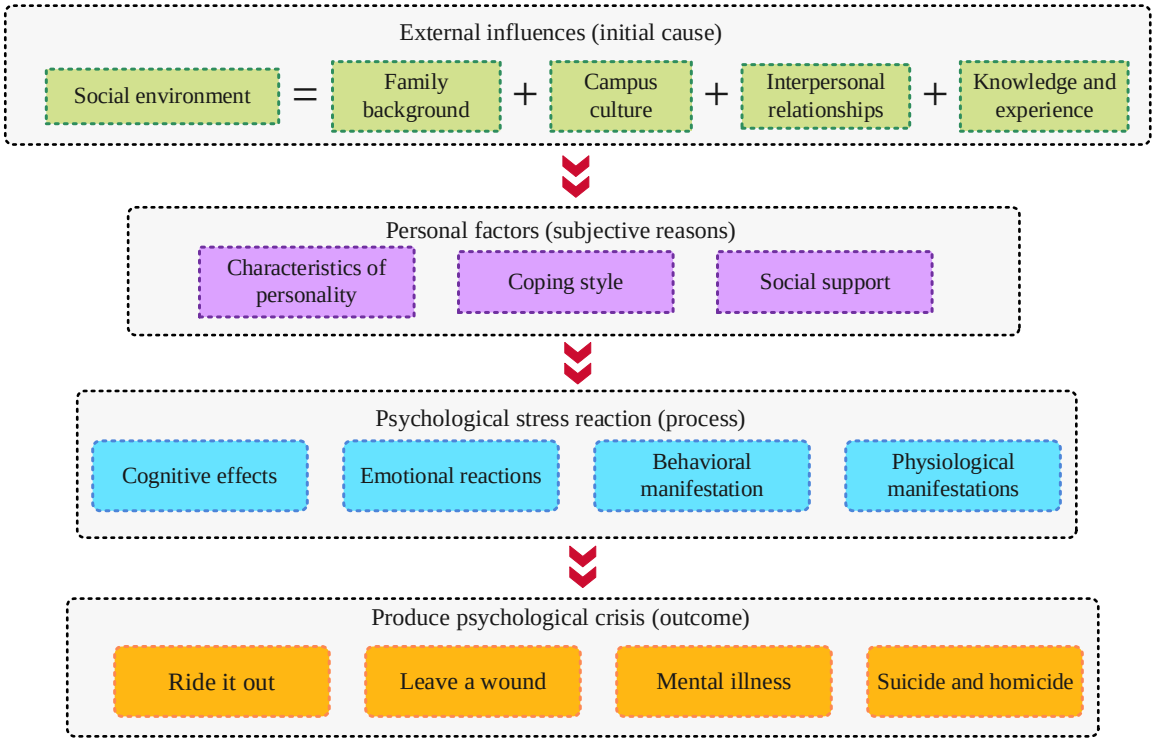

Figure 1: The process of students' psychological crises and advancement.

Figure 1 presents a revised conceptual model for explaining reactions to psychological crises, outlining the main determining factors. The model includes external factors, such as environmental and social stressors, personal traits like resilience and emotional stability, the social environment encompassing peer relationships and institutional support, as well as psychological stress reactions, which are divided into emotional, cognitive, and behavioral expressions [9], [10]. As a result, a proactive early-warning system could be instituted to address the psychological issues of college students before they manifest, thereby reducing the likelihood of more severe psychological problems [11].

To facilitate a better understanding, Figure 2 outlines the structure of the proposed early warning system. The system has several stages: the first stage is data acquisition, where psychological profiles, academic data, and questionnaires are collected; the second stage is data preprocessing, where encoding and feature selection are used to enhance data quality; the third stage is model implementation, where the data is assessed using a data fusion model that combines different machine learning approaches; this is followed by the prediction and early warning step, where the system generates risk scores for students to determine those in need of intervention; lastly, the intervention and monitoring step involves counseling teams using the predictions made by the system to provide timely mental health interventions.

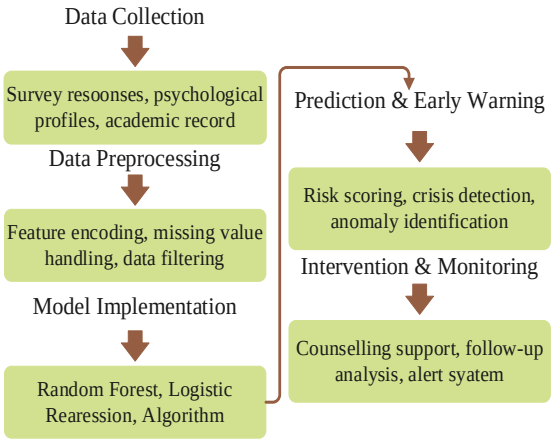

Figure 2: Flowchart of the proposed early warning system.

Xiang-Wei [12] conducted a study on analyzing mental health patterns in vocational college students using big data. He emphasized the significance of self-service systems for mental health, delved into their technical structure and psychological aspects, and proposed strategies to address operational obstacles. His research aimed to create a mental health information platform during the Internet Plus era. Xiu-Mei [3] developed a psychological crisis early warning system for college students within the context of university management. The system focused on prevention, with intervention as a supporting element, emphasizing a human-centered approach and collaboration between universities and colleges. Xiu-Mei advocated establishing university management institutions, proper staffing, improved process management, and network platforms to

systematically create an early warning system for psychological crises among college students.

Some professionals also argue that creating a database of resources for students' psychological issues based on their psychological traits is feasible. This would allow storing and analyzing online information related to psychological problems, enabling timely assessments of college students' psychological issues. This approach would facilitate effective information management and the handling of individualized psychological problems, which is crucial for mental health education. Hinduja et al. [13] explored the possibility of leveraging social media platforms as social sensor ecosystems to overcome the limitations of traditional mental health monitoring. Through the analysis of Twitter data, they developed an active surveillance system that integrated data preprocessing with a long short-term memory model for the real-time detection of at-risk individuals. Their approach showed improved predictive accuracy compared to current methods. Li [14] suggested implementing a mental health assessment system that utilizes data mining algorithms to enhance the accuracy and effectiveness of evaluating college students' mental well-being. By employing the Apriori algorithm to analyze mental health surveys and incorporating a three-dimensional matrix, the system was able to generate intelligent assessment outcomes. The system's superior performance and reliability were confirmed through simulation experiments, effectively overcoming the limitations of current evaluation approaches. The advancement of this resource information database requires extensive data integration and analysis to enhance the early warning system for students' psychological crises.

Research institutions commonly employ questionnaires and psychological assessment tools to enhance their information repository. Nevertheless, subsequent studies have revealed persistent challenges with the data, including significant inaccuracies, restricted efficacy, and inadequate timeliness. The primary cause of these inaccuracies stems from students' apprehension of being stigmatized as "psycho" or "abnormal" due to their psychological issues, potentially resulting in social isolation from their peers. Consequently, they tend to provide random responses to the questionnaires, thereby compromising the effectiveness and precision of the psychological surveys [15], [16], [17]. The efficiency of psychological crisis warnings in this form has been greatly improved. Nowadays, big data technology has been introduced into psychological crisis early warning research, and many outcomes have been achieved. For example, in 2011, relevant institutions classified users into five personality types by mining and analyzing the data on smartphones [18]. In 2017, big data and data mining were applied to analyze the techniques to cope with stress [19].

Jiang et al. [20] developed a responsive early warning system for ideological and political education using psychological big data analysis. The system collected data from assessments, mobile sensors, and learning histories, using artificial intelligence and machine learning to detect anomalies in emotional and behavioral patterns. It issued timely warnings about potential risks, enabling instant interventions and enhancing the personalization, intelligence, and scientificity of the learning process. Yan [21] introduced a DM-BPNN model, which combines data mining techniques with a BP neural network, to tackle psychological distress among college students and enhance their overall mental well-being. By incorporating the Apriori algorithm, this innovative approach provided features such as early warning systems and the training of mental health professionals. Through the utilization of the Apriori algorithm in processing survey data, the DM-BPNN was able to improve mental health assessments. Consequently, leveraging this refined algorithm, the DM-BPNN demonstrated high precision in predicting psychological crises. Tian and Yi [22] created an artificial intelligence and big data analysis-based system to provide mental health support for students. They formulated algorithms and patterns tailored to the requirements of mental health support and crisis prediction. They carried out pertinent experiments and tests to implement these technologies successfully. The study's findings demonstrated that the system delivered personalized mental health support and guidance to students and accurately predicted potential mental health crises, yielding noteworthy outcomes. Shen [23] emphasized the significant mental health challenges and psychological emergencies experienced by certain Chinese university students. He illustrated the efficiency of the Kalman filter, a regression computation technique that handled data with minimal error, enabling precise calculations of high- and low-dimensional system data. The research encountered challenges and recognized enhancements, indicating that the Kalman filter could reliably forecast mental health issues among students. Empirical findings validated the precision of the approach and its capacity to mitigate psychological emergencies among university students.

Ni et al.[24] analyzed psychological stress in college students with an emphasis on its impacts in times of public health crisis. Traditional data mining methods relied on questionnaires, limiting their ability to represent stress dynamics properly. To address data imbalance, they built a neural network model and implemented an anomaly detection approach in comparing the results. This model outperformed random forests and decision trees and thus improved accuracy and reliability in identifying highly stressed students. However, Yan [25] researched enhancing psychological fitness education and evaluation by applying artificial intelligence. They established a model based on fuzzy mathematics and neural networks for data analysis and outlier identification purposes in order to detect at-risk students. Using MATLAB simulation experiments, the model achieved an accuracy level of 95.97% as compared to other algorithms.

Table 1: Comparison of crisis detection models.

Study	Accuracy	Precision	Recall	F1 Score
[12]	88.2%	86.5%	84.7%	85.6%
[13]	91.0%	90.2%	88.9%	89.5%
[14]	85.3%	84.1%	82.7%	83.3%
[20]	93.1%	92.5%	90.3%	91.4%
This Study	95.2%	93.7%	90.8%	92.2%

Current research emphasizes the effectiveness of multi-source data fusion in psychological early warning systems. The combination of structured and unstructured data via deep learning-based fusion results in improved predictive performance. Hybrid ensemble models, meanwhile, enhance crisis detection. Real-time adaptive fusion enables dynamic data weighting adjustments, thus bolstering personalized monitoring. Informatica's research points to the importance of integrating data, selecting features, as well as combining text mining and behavioral metrics to enhance crisis detection.

Based on current literature, this research presents a data fusion model that combines Random Forest, Logistic Regression, and the Apriori algorithm, achieving a significant increase in predictive accuracy compared to single-method strategies. The fusion model proposed here achieves an F1 score of 92.2%, higher than previously set techniques and with improved classification performance. Using data mining practices, the early warning system outlined here systematically evaluates and predicts the mental health condition of university students with greater stability and accuracy. Psychological profiles based on family theory and life event theory add to increased predictive ability, thus improving the overall performance of the model.

This study focuses on college students' mental well-being by introducing an early warning system developed through a data fusion model. This research aims to systematize and standardize the early detection of psychological crises by employing data mining techniques. The data fusion model created in this study examines student-related data to establish an efficient early warning system. The outcomes suggest that psychological profiles of college students, based on family theory and life event theory, can provide more effective psychological characteristics for analysis, thereby improving the predictive precision of the algorithm model. The study reveals that the Apriori algorithm outperforms random forest, logistic regression, and decision trees in handling mental health data, with the data fusion model achieving an average F1 score of 82.2%, thus enhancing the predictive capability of individual patterns and excelling in psychological data analysis.

Hyperparameter tuning was performed using GridSearchCV for optimal model performance. Decision Tree parameters (max_depth, min_samples_leaf) were adjusted to prevent overfitting, while Random Forest optimized n_estimators (optimal: 26). Logistic Regression fine-tuned regularization strength (C) using L2 penalty, and Apriori adjusted support and confidence thresholds for significance. Cross-validation ensured the best hyperparameter selection before fusion.

This study is grounded in survey information, although factors like family interactions and major life events contribute to psychological crises. Parental relationships, financial stress, and life changes affect overall quality of life. Future studies may include integrating family background along with current behavioral information to enhance the accuracy of predictions. The Apriori algorithm proved to be very effective due to its ability to reveal associations in psychological data sets. It identifies hidden interdependencies, detects concomitant risk factors, handles class imbalances irrespective of label balancing, and provides understandable rules that ease decision-making processes. Though computationally expensive and not ideal for regression analysis, its rule-based approach is best applied to psychological survey data, which explains its high efficacy.

2 Types of data mining

Data mining involves finding potentially valuable knowledge and previously unknown information [26]. Through data mining, researchers can realize predictive modeling, association analysis, cluster analysis, or outlier detection [27]. Predictive modeling is often used for classification and regression. Classification is usually used if the target variable is discrete. If the target variable is continuous, regression is generally used. In the current study, logistic regression is used as a probabilistic classifier with outputs in the range of [0,1] that represent the probability that a given student is likely to have psychological distress. These probabilities are then thresholded to classify participants as either within a normal range or as being at risk. Although the UPI score ranges from 0 to 60 and is used as a continuous variable, the model is not intended to predict this score as a numerical quantification in a straightforward manner. Instead, the psychological state is divided into binary outcomes to allow the evaluation of crisis states. In both classification and regression, it is necessary to train a model. The purpose is to make the target variable as close to the real value as possible. Association analysis is used to discover patterns with strong association characteristics implicit in a dataset. The most classic case of correlation analysis is the correlation analysis of supermarket shopping basket data. It is found that consumers who buy diapers often buy beer, which eventually leads to a change in product placement in supermarkets. Cluster analysis is the process of finding closely related groups in data. The greater the similarity between the members of a group and the greater the difference between different groups, the better the effect of cluster analysis. As a powerful auxiliary tool, cluster analysis technology has proven invaluable in many fields, such as scientific research, social services, marketing, etc. Therefore, cluster analysis technology has become a hot topic. Outlier detection is designed to find the data in the data set with significantly different characteristics from other data [28].

2.1 Types of data mining algorithms

This study integrates Random Forest for analysis of high-dimensional data, Logistic Regression for interpretability, and Apriori for behavioral pattern extraction. Decision Trees were left out of the analysis to avoid the risk of overfitting. The combined methodology strengthens predictive accuracy and interpretability by integrating ensemble learning, statistical modeling, and rule-based approaches.

(1) Decision Tree Analysis

The decision tree is one of the most frequently employed classification techniques in data mining. This method originated from the concept learning system CLS and later evolved into the ID3 method. The central notion of the ID3 algorithm is as follows: firstly, the attribute with the highest gain is selected. Let *p(i/t)* represent the fraction of records that belong to class I given node t, and let C denote the number of different courses. Then, the information entropy is defined as [29]:

$$Entropy(t) = -\sum_{i=1}^{c} p(i \mid t) \log_2 p(i \mid t) \quad (1)$$

When information entropy is chosen as the impurity measure, the information gain δ_{info} is defined as:

$$\Delta\ info = I(parent) - \sum_{j=1}^{k} \frac{N(Vj)}{N} I(Vj) \quad (2)$$

The information gain ratio is defined as:

$$Gain\ ratio = \frac{\Delta\ info}{SplitInfo} \quad (3)$$

Here, the classified information split is defined as:

$$SplitInfo = -\sum_{i=1}^{k} P(v_i) \log_2 P(v_i) \quad (4)$$

where *K* is the total number of partitions.

Additionally, the *Gini* coefficient is defined as:

$$Gini(t) = 1 - \sum_{i=1}^{C} P[i/t]^2 \quad (5)$$

Generally, a greedy strategy is used to recursively generate nodes from top to bottom based on the selected feature criteria. The stopping criterion for splitting is reached when all records in a node belong to the same class, or all record attributes have the same value.

(2) Random Forest Analysis

Random Forest utilizes ensemble learning primarily to address the issue of overfitting encountered by individual decision trees. A single learner typically trained from the data using existing algorithms (such as ID3 and C4.5) often exhibits poor generalization [30]. As models become more complex, they may fit training data better, but this can lead to overfitting, making them less effective on new data. Random Forest creates multiple trees simultaneously and assesses the importance of each variable across different parameter configurations to mitigate overfitting. This study employs techniques that average the reduction in impurity to determine variable importance.

Assuming the features are X... and X_m, the calculation of the Gini index is taken as an example and expressed by GI_m, then:

$$GI_m = \sum_{k=1}^{|K|} \sum_{k' \neq k} P_{mk} P_{mk'} = 1 - \sum_{k=1}^{|K|} P_{mk}^2 \quad (6)$$

Here, *K* displays the number of categories. The importance of feature X_j in a node is then given by:

$$VIM_{jm}^{Gini} = GI_m - GI_l - GI_r \quad (7)$$

Here, *VIM* displays the variable importance score, and GI_l and GI_r represent the Gini indices of the two new nodes after branching. The importance of X_j in the i_{th} tree is calculated as follows:

$$VIM_{ij}^{Gini} = \sum_{m \in M} VIM_{jm}^{Gini} \quad (8)$$

Assuming there are *n* trees in the random forest, then:

$$VIM_j^{\text{Gini}} = \sum_{i=1}^{n} VIM_{ij}^{\text{Gini}} \quad (9)$$

Finally, the importance scores of the variables obtained above are normalized as follows:

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i} \quad (10)$$

(3) Analysis of the Apriori Algorithm

The Apriori algorithm is considered the most well-known among association models and was first proposed by Gao H [31]. Based on the Apriori algorithm, it can effectively enhance the effectiveness of finding frequent subsequences as feature sequences and make them representative. The core idea of the Apriori algorithm is self-joining and pruning. The algorithmic process depends on two properties: If an item A is added to the item set of the infrequent sequence S, the new sequence S ∪ A must not be frequent. The next step involves finding Lk. The candidate sequence item set is generated by self-joining Lk, and then the frequent sequence item set is found using downward closure. The K-sequence item set is identified through an iterative layer-by-layer search method, and the set of candidate sequence item sets is denoted as Ck. The "pruning" process involves scanning the event sequence data record library to determine the count of each candidate sequence in Ck, thereby identifying Lr. All candidate sequences whose count meets the minimum threshold are considered frequent sequences and belong to Lr.

(4) Logistic Regression Analysis

The logistic regression model serves primarily three purposes: (1) predicting the probability of an event given different factors; (2) qualitatively assessing the nature of events based on predicted outcomes; and (3) identifying the origins of risks and analyzing factors contributing to events. The logistic regression model requires that the dependent variable be a probability or categorical variable limited to values between 0 and 1. Unlike linear regression, it does not assume a normal distribution for the dependent variable or random errors. Instead, it assumes a linear relationship between the logit of the dependent variable's probability and the independent variables, with each observation being independent of others. Surface rupture during strong earthquakes is a stochastic event

influenced by numerous factors. Predicting surface rupture can be framed as a conditional probability problem, addressable through logistic or multivariate statistical regression.

It examines the functional relationships between variables, resulting in a regression equation that not only reveals how independent variables influence the dependent variable but also allows for prediction and control [32]. In multiple linear regression analysis, the independent variables are usually assumed to be non-random variables to estimate the model parameters. Random error is equal variance and uncorrelated; the dependent variable and random error should conform to the normal distribution. In practical problems, sometimes the dependent variable is a qualitative quantity, and the independent variable is a random quantity, which does not conform to the assumption of multiple linear regression. For example, in predicting a strong earthquake surface rupture, most factors, such as earthquake magnitude, focal depth, fault displacement mode, and overburden thickness, are random variables. Whether the prediction result is a surface rupture is a [0, 1] binomial distribution problem, which does not meet the requirements of normal distribution. To solve this problem, scientists introduced a logistic function.

(5) Model Fusion Method

Model fusion is a strategy for combining different or identical patterns to enhance the effectiveness of a single model. Common model fusion techniques include voting, averaging, self-help, lifting, and stacking. The stacking method is exemplified by random forest, and the boosting method is demonstrated by Adaboost, both of which were detailed in the previous section. Multi-layer data fusion is relatively uncommon because its practical efficacy is not significantly stronger than that of a two-layer fusion model, although it does improve time complexity. Therefore, data fusion models typically employ only two layers, a strategy also adopted in this study [33], [34], [35].

While the fusion model improves predictive power, it has its disadvantages. Training a number of models and then fine-tuning the meta-classifier results in significant computational complexity. The model can also inherit biases from the base models, and its performance depends on finding optimal thresholds.

3 Establishment of a data fusion model

3.1 Data acquisition

A total of 4045 student psychological profiles and corresponding UPI questionnaire outcomes were collected, with 23 invalid questionnaires excluded. This resulted in 4022 valid questionnaires, including 1808 from male students and 2237 from female students. The group measurement method involved freshmen entering the school computer room in batches. They logged into the school's psychological evaluation system to complete a personality test. Subsequently, students filled out paper psychological profiles collected by student assistants and submitted them to the school's psychological counseling center for centralized archiving. The personality questionnaires of college students and their psychological profiles were data collection tools.

3.2 Data preprocessing

To ensure high-quality inputs for model training, the dataset went through an extensive range of preparatory processes, including data cleaning, handling missing values, feature encoding, and processes related to anonymization. Duplicates were removed, logical inconsistencies were detected and confirmed, and extreme outliers were corrected or deleted from the dataset. The method of handling missing data was listwise deletion for large gaps, while small gaps were handled by multiple imputation methods using mean, mode, or regression approaches, as appropriate. Anonymization processes were put in place to protect privacy by replacing personal identifiers, using AES-256 encryption for data protection, and setting up access controls to limit decryption to approved personnel only. Ethical requirements were met through the procurement of informed consent and confidentiality agreements approved by the university. Overall, these methodologies increased the integrity, security, and reliability of the data, hence enabling effective model training. Data integration involves merging non-source data to reduce redundancy and inconsistencies in datasets. This study's data includes personality test results of college students, analysis outcomes of psychological files for students without abnormal questionnaire results as assessed by psychological staff, and psychological file information of college students.

Psychological staff analyzed files using expert inspection techniques to identify students without psychological abnormalities based on their counseling experience. Students potentially facing psychological crises were flagged as exceptions. Due to the dataset's size, Microsoft Office Excel was used for data collection and storage. Improper handling risks exposing personal privacy and causing harm. To safeguard privacy, I signed a confidentiality agreement with the relevant department before handling data, ensuring no information that could expose identities is stored.

Most data used in this study from psychological files is categorical, while classification models typically require encoded features for modeling and prediction. Feature values from psychological files are encoded using one-hot or label encoding methods in classification models. Decision trees, random forests, and Apriori algorithms commonly use label encoding for effective results. Numeric scaling does not affect node splitting or tree structure [23]. However, using one-hot encoding in tree models increases feature dimensions and sparsity, potentially increasing computation without maximizing data information, leading to poor model performance or overfitting.

4 Construction of the fusion model

The approach to constructing a system from the external objective environment differs from enhancing college students' psychological resilience to cope with stressful events independently. Each individual possesses unique traits influenced by factors such as constitution, family environment, and parental relationships, which can impact their responses to stress. While some students can navigate psychological crises independently and effectively [36]others with weaker coping abilities may require external support. Constructing a coping system involves two main aspects: establishing an early warning mechanism to identify susceptible individuals using relevant indicators for proactive psychological preparation and implementing a rapid response mechanism to provide immediate assistance following stressful events.

The model predicts whether a student is experiencing a psychological crisis, which is a binary classification task. While logistic regression is traditionally a statistical regression method, it is used here as a probabilistic classifier. The output is a probability score between 0 and 1, representing the likelihood of a psychological crisis. This probabilistic output is then thresholded into binary categories (normal vs. at-risk). Logistic regression was chosen because of its interpretability, ability to model relationships between psychological features, and its effectiveness in handling structured survey data [19]. When applying these models, it's crucial to consider the data's characteristics. The psychological archive data used in this study exhibit two key features: a high correlation among attributes such as 'parenting style,' 'lack of family trust,' 'academic pressure,' 'learning difficulties,' and 'test anxiety,' and a small proportion of individuals with psychological abnormalities, indicative of imbalanced data [37]. Subsequently, these patterns are employed to model and predict psychological data. The effectiveness of various classification models is evaluated, and the most suitable model serves as the base for data fusion. To fix class imbalance in psychological crisis prediction, multiple methods were applied. SMOTE created synthetic minority samples to prevent bias, while Logistic Regression used even class weighting to prioritize underrepresented cases. Threshold tuning optimized recall without excessive false positives, which ensures effective detection and generalizability.

Each model had specific strengths in crisis detection: the Decision Tree delivered interpretability, the Random Forest guaranteed robustness, Logistic Regression supported probabilistic prediction, and Apriori facilitated behavioral patterns extraction. Synergetic integration of these models reinforces predictive power by combining rule-based insight with statistical approaches.

4.1 Construction of data fusion model

(1) Psychological Crisis Warning Based on Decision Tree Model

This article explores factors influencing students' psychological crises through an analysis of attributes from student information, particularly focusing on the data of 30,000 students. Using empirical methods and binary logistic regression analysis, six key attributes significantly associated with psychological crises were identified. Subsequently, a decision tree model for early warning of psychological crises was constructed, evaluated qualitatively, and tested quantitatively.

Initially, 17 attributes were collected, showing weak correlations. Following expert advice and practical experience in student management, attributes such as gender, age, academic year, student leadership roles, educational achievements, and attendance were prioritized for further analysis. A binary logistic regression analysis was conducted on the remaining nine attributes after initial screening, excluding those with a low correlation that did not contribute to the regression model. After the initial screening, nine key features were selected based on statistical significance, avoidance of collinearity (VIF > 5), and expert support. These features capture psychological distress, behavioral risks, and socio-factors, thus ensuring the model can precisely identify meaningful predictors of crises. This process aimed to identify significant features essential for model construction. Data analysis was performed using SPSS software, inputting whitened independent variable data matrices and predicted values. The forward stepwise regression method was employed based on maximum likelihood estimation, with entry criteria set at $A = 0.05$ and removal criteria at $A = 0.10$. The outcomes, including significant attribute contributions to the regression equation, are displayed in Figure 3.

A psychological crisis early warning model was developed using masculinity as a classification index, where negative outcomes are denoted as 0 and positive outcomes as 1. Masculinity has been added as a categorical variable due to its demonstrated impacts on psychological resilience, emotional expressiveness, and coping. Previous literature shows that traditional norms of masculinity influence the reporting of mental health issues in such a way that those with higher masculinity scores show less tendencies towards seeking support and want to downplay their psychological distress. Including masculinity in the model attempts to speak to such disparities based on gender and enable better identification of risk individuals [38]. The model utilized six characteristic attributes: personality, family economy, family relationship, academic performance, disciplinary actions, and attendance history. Figure 4 displays a portion of the standardized training dataset used for constructing the model.

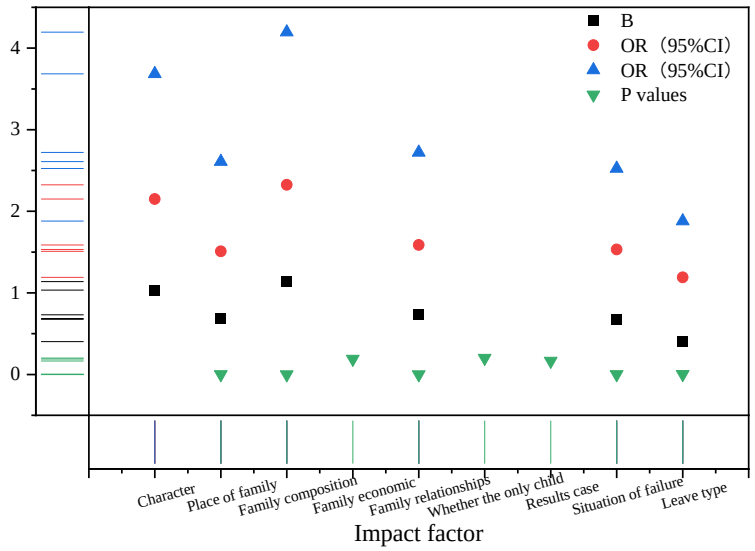

Figure 3: Outcomes of logistic regression analysis.

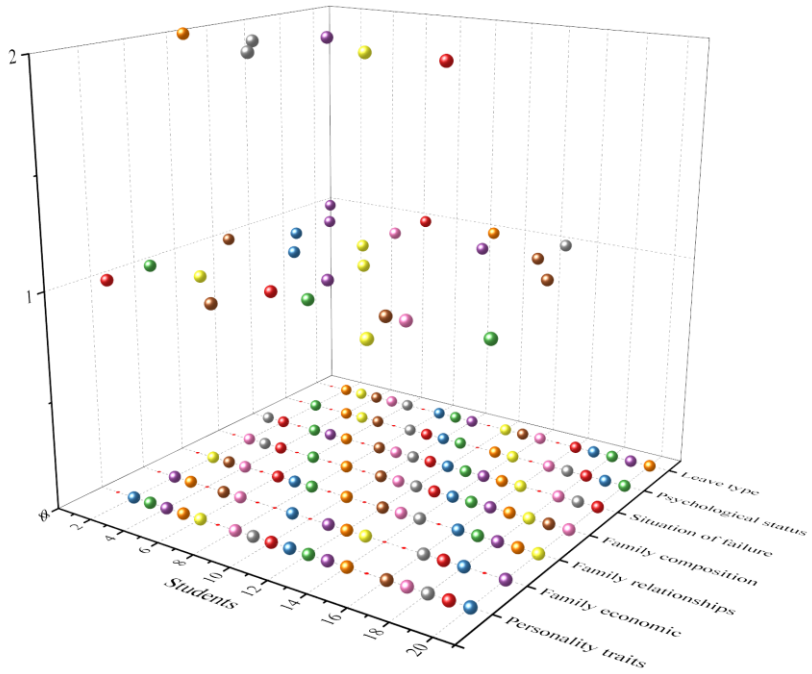

Figure 4: Training set data fragment.

The calculated outcomes indicate that among all feature attributes, 'personality characteristics' have the highest information gain rate. Following the principles of the C4.5 algorithm, 'personality characteristics' are selected as the root node attribute, dividing the training set samples into two parts based on its two attribute values.

In performance evaluation, the model was tested using 254 random test sets, achieving an overall prediction accuracy of 97.2%. The precision rate for positive warning outcomes reached 81.8%, with a recall rate of 64.3%. The F1 score, which balances precision and recall, was 72.0%. Testing across datasets from three universities consistently showed that the model achieved an overall accuracy of over 95%, with positive warning precision exceeding 90%, a recall rate around 60%, and an F1 score above 70%. These results confirm the model's stable predictive efficacy in identifying psychological crises.

(2) Psychological Crisis Warning Based on Random Forest Analysis

When employing random forests, a crucial parameter is 'n_estimators,' which represents the number of trees in the forest. The default value is 10, and increasing this parameter generally improves model stability while reducing the risk of underfitting. Using the Matplotlib module in Python, the learning curve for 'n_estimators' can be plotted, as displayed in Figure 5.

According to Figure 5, the optimal value for the 'n_estimators' parameter is 26. The optimal n_ estimator

value of 26 was determined via GridSearchCV, balancing validation error and efficiency. Increasing beyond 26 yielded minimal F1 score gains, while lower values reduced stability, making 26 the best trade-off. The random forest model was established using default values and tuned parameters. On the test set, the model achieved a precision rate of 92.7% and a recall rate of 58.5%, with an F1 score of 71.7%. After tuning parameters using the decision tree model on the same test set, the precision improved to 91.5%, the recall to 66.2%, and the F1 score to 76.8%.

The Apriori algorithm was used to mine psychological data from all school students. Table 2 presents the outcomes of nine-dimensional association rule mining. Table 2 defines notable trends with respect to distress in students. High stress and poor sleep patterns lead to an increased risk for depression, but low social contact and high anxiety relate to deteriorated academic functioning. These results enable the identification of at-risk students and targeting intervention efforts and enhance our understanding of mental health.

(4) Based on a Logistic Regression Analysis

Logistic regression reduces model overfitting through regularization. L1 and L2 regularizations are widely employed, and the corresponding important parameters are penalty and C. C regulates regularization strength in logistic regression. Lower C adds regularization, penalizes larger coefficients results in less complex models, and fits the training data less precisely. GridSearchCV optimized C to achieve a balance between complexity and generalization. Penalty is a regularization method that can be specified by typing 11 or 12. J (θ) is the loss function, and C is the hyperparameter used to control the degree of regularization. The smaller the value of C, the greater the intensity of regularization, and the smaller the value of parameter θ. L_l

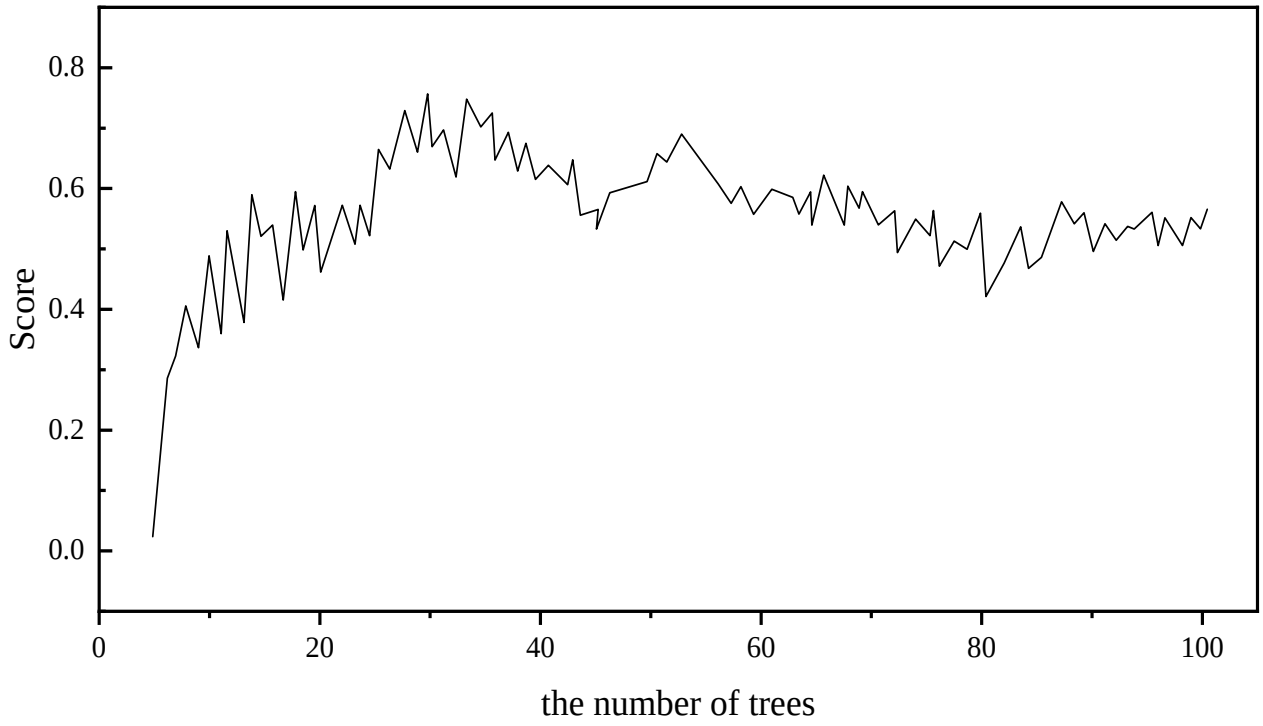

Figure 5: n_ estimators' parameters of the learning curve.

Table 2: Association rules between nine psychological symptoms dimensions.

Serial number	Association rules	Degree of support	Confidence level
1	Have anxiety, Hostile →force	0.21	0.94
2	Have anxiety, psychotic →depression	0.24	0.84
3	Have a terrorist, Have personal connections→ depression	0.31	0.95
4	No compulsion, No paranoid→ No personal relationship	0.33	0.88

Regularization will compress the parameter θ to 0 as the value of C becomes smaller. In contrast, L_2 regularization will compress the parameter θ to a small value as the value of C becomes smaller but will not reach 0. The learning curve of regularization parameters is displayed in Figure 6. It is evident from Figure 6 that when the L_1 regularization method is used and the C value is 1, the logistic regression model draws on mental data, with an F_1 score of 81.1 percent.

(5) Data Fusion Model

Through the previous modeling analysis of the decision tree, random forest, the Apriori algorithm, and logistic regression on mental data, it was found that the performance of the decision tree model on mental data is significantly worse than that of the random forest, logistic regression, and Apriori patterns. Therefore, the decision tree model is not considered the base model, but the random forest model, the logistic regression model, and the Apriori model are considered the base patterns. Since three base patterns can be selected, four different data fusion patterns can be generated, as displayed in Table 3.

The RLA data fusion model is depicted in Figure 7. Comparative analysis shows that the Decision Tree F1 score was significantly lower than the Random Forest, Logistic Regression, and Apriori algorithms' scores. Cross-validation showed high variance and poor generalization power, implying overfitting. This result justified its exclusion from the fusion model.

Table 3: Fusion model generated from different base patterns

Basic model	Model after data fusion	Advantages & Best Use Cases
Random forest + logistic regression + the Apriori algorithm	RLA-S	Provides the most comprehensive predictive capability by combining feature selection (Logistic Regression), robust classification (Random Forest), and association rule learning (Apriori). Best suited for highly imbalanced datasets where psychological patterns are complex.
Random forest + logistic regression	RL-S	Strong classification performance with good feature interpretability. Works well for datasets with numerical and categorical mixed features, making it efficient for real-time predictions.
logistic regression + the Apriori algorithm	LA-S	Prioritizes interpretable and explainable decision-making. Suitable for situations where transparency is required, such as mental health monitoring systems where professionals need clear, actionable insights
Random forest + the Apriori algorithm	RA-S	Leverages ensemble learning and pattern discovery. Best for datasets where association rules (Apriori) can enhance decision-tree-based classifications, such as behavioral clustering.

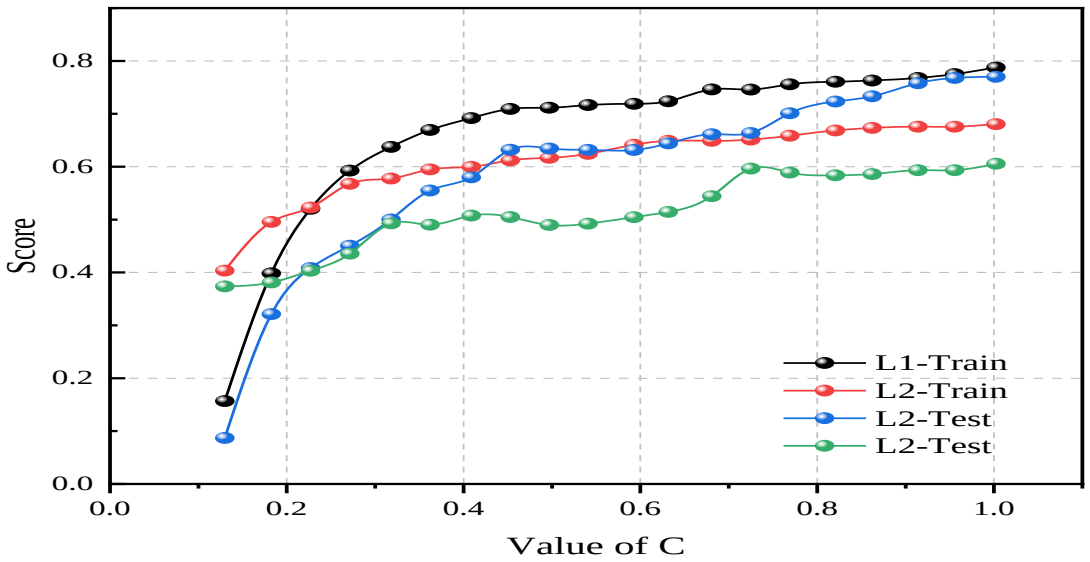

Figure 6: Learning curves for the regularization parameters.

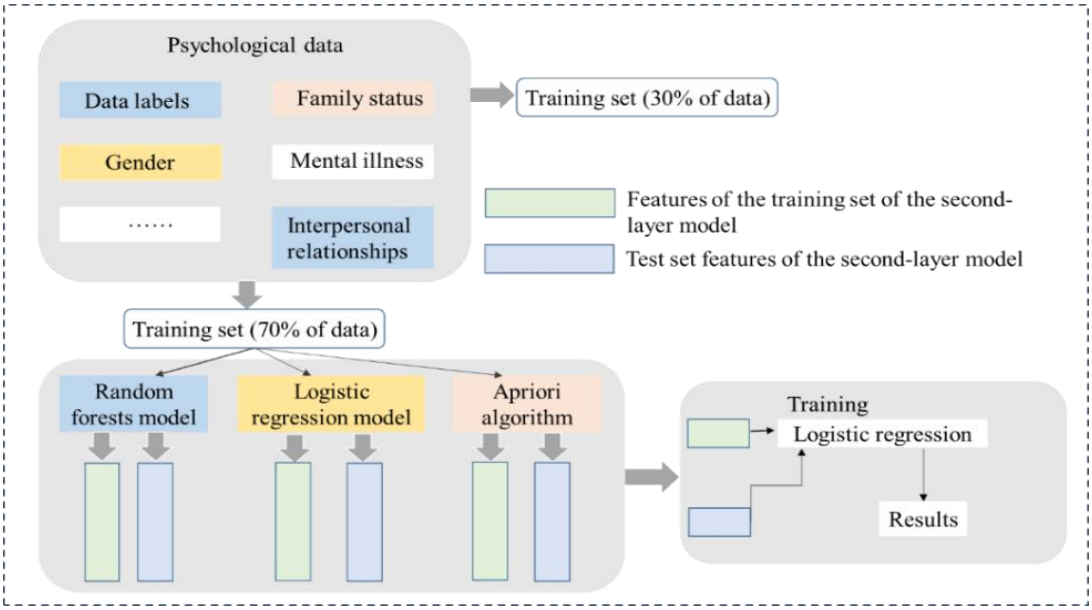

Figure 7: Data fusion model.

To conduct a systematic comparison of the four combinations, we measured their F1 score, accuracy, precision, and recall. The best F1 score was achieved by the RLA-S combination and illustrates how ensemble learning, statistical classification, and rule-based discovery of patterns significantly improve predictive power. On the other hand, the RL-S combination had slightly less accurate results and shows how despite the effectiveness of ensemble approaches, they might not be able to capture the behavioral patterns obtained by association rule mining in an optimal manner.

The modeling process of the psychological crisis early warning model utilizing the data fusion model is as follows: (1) After obtaining data labels and feature sets, construct psychological data sets. To ensure the model's robustness, the dataset was split using an 80-20 stratified sampling method, preserving the proportion of at-risk students. In addition, a 10-fold cross-validation process was adopted, where the dataset was partitioned into 10 different subsets, with each subset used as a test set once and the other subsets used for training. This practice helped reduce bias, increase performance stability, and mitigate the overfitting risk. 70% of the mental data set is divided into training sets, and 30% is divided into test sets. (2) On the test set, a 10-fold cross-validation technique was employed to train the random forest model, the logistic regression model, and the Apriori algorithm, respectively. Doing so makes the most of the data and prevents the model from overfitting. Figure 8 and Figure 9 illustrate how a base model does 10-fold cross-validation.

Under the paradigm of fusion, the output probabilities of Random Forest, Logistic Regression, and Apriori are taken as features of an ensuing meta-classifier in stage two. These probability values indicating the predicted risk levels of all the students are merged into a feature matrix to train a logistic regression model in this approach. This approach enhances the process of generalization by leveraging the strengths of each individual model and compensating for their respective weaknesses.

To optimize the hyperparameters and test the stability of the model, a 10-fold cross-validation strategy is implemented on the 70% train set before performing the ultimate test. This strategy efficiently avoids data leakage by ensuring the 30% test set remains completely untested up to the evaluation stage. The train set is partitioned into 10 separate folds such that each fold acts as an interim validation set and the rest are used for train purposes. This process maximizes parameter tuning, reinforces robustness by measuring over multiple partitions, and ensures low overfit risk by requiring the model to generalize prior to seeing unseen data. Keeping the test set separate ensures the measurement of expectations based on reality and the ultimate values is reliable.

Take random forest as an example, as displayed in Figures 7 and 8. After 10 times cross-validation of the random forest model, 10 prediction outcomes will be generated on the test set, respectively. Combining the prediction outcomes is the prediction result of random forest on the whole training set, denoted as train_pre1. Averaging the 10 predictions on the test set is the random forest in the test set. It is denoted as Test Prel. In the same way, after 10 times of cross-validation, logistic regression can get the prediction result train_pre2 on the training set and test__pre2 on the test set, and the Apriori algorithm can get the prediction set train__pre3 and test set on the training set. The prediction result is Test Pre3. The prediction outcomes (Train Prel, Train_pre2, Train_pre3) obtained by random forest, logistic regression, and the Apriori algorithm on the training set are combined to form the second layer model training set's feature set. This feature set and the true labels of the training set constitute the training set of the second layer model. Random forest, logistic regression, and the Apriori algorithm were used in the test set. The prediction outcomes of Test_pre1, Test_pre2, and Test_pre3 are combined to form the feature set of the second layer model test set. This feature set and

the real label of the test set form the validation set test second.

For the sample equilibrium data, if the model precision is high, it can indicate that the model performance is good. In model evaluation, several metrics were utilized, with specific focus given to the F1 score due to its balance between recall and precision. Given the slight imbalance in the dataset, using only accuracy would yield a skewed picture, as misclassifying the at-risk students can have drastic consequences. While ROC-AUC and MCC are useful, they lack the instant interpretability necessary for intervention-driven decision-making. Taking the psychological data of this study as an example, there are 4045 pieces of data and 3557 pieces of normal mental state, accounting for 87.9% of the total data. Even if the mental prediction model does not use any algorithm to learn and predicts every sample to be psychologically normal, it can achieve 87.9% precision. All the patterns used in this study can achieve 95% precision in the test set, and the precision is of poor significance in evaluating the model's merits. Therefore, precision and recall were considered in this study. The ideal situation is that the precision and recall of the model are high. It's hard to improve both. Therefore, the F1 score was finally considered to evaluate the model. F1 can consider both precision and recall and is more suitable for comparing the classification effects of each model.

The test set performance at 95% represents the ultimate optimized model after procedures like tuning, feature selection, and data fusion, showing its ability to generalize. In contrast, the accuracy of 87.9% represents the starting metric before optimization. The baseline at the beginning highlights the starting point, and the final performance validates improvements in the model.

Precision and recall must be considered together to avoid misleading conclusions, as precision defines the accuracy of positive predictions, while recall measures the number of actual positives correctly identified. The F1-score is used to balance the two measures, making it crucial for psychological crisis identification, where false positives and negatives have significant implications. Alongside the F1-score, accuracy, and AUC-ROC provide additional insights—accuracy can provide misleading outcomes in imbalanced datasets, while AUC-ROC measures class differentiation. Considering the metrics together ensures a comprehensive model effectiveness assessment.

Figure 7 highlights the variation in performance between folds, with annotations indicating the optimum number of estimators that achieve an optimum balance between training accuracy and validation error. Figure 8 illustrates the model's performance on new, unseen data, thus demonstrating its capacity for generalization, with the optimal hyperparameter configurations highlighted to represent the model's best-performing iteration.

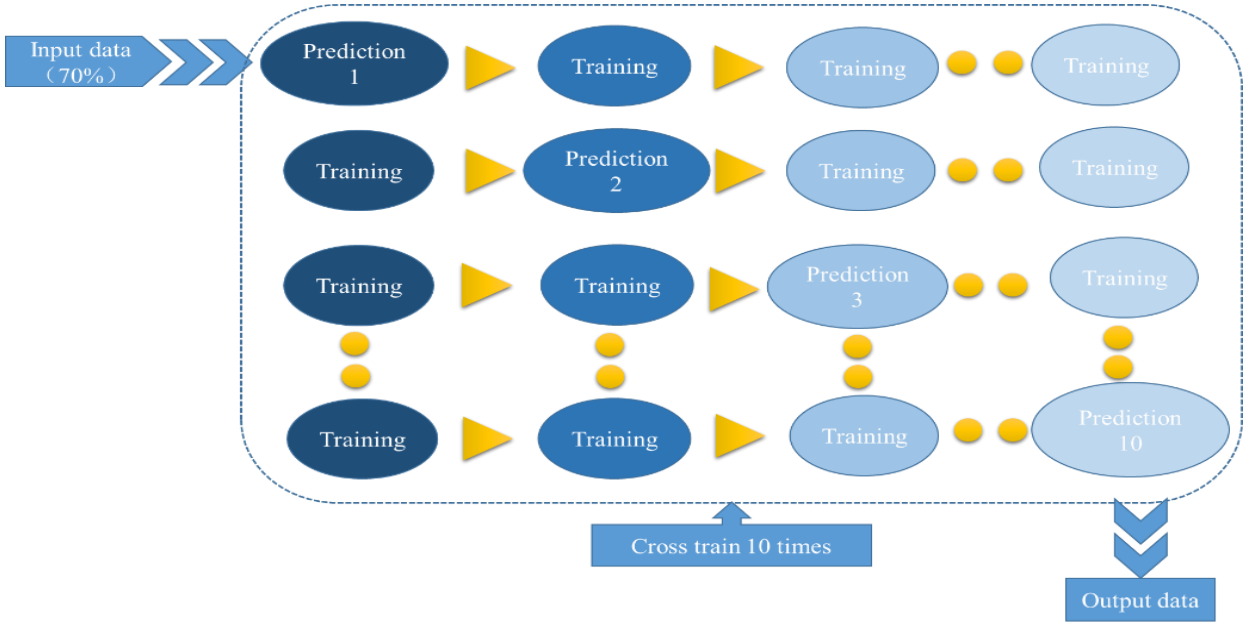

Figure 8: Learning curve during 10-fold cross-validation (training set: 70% of data).

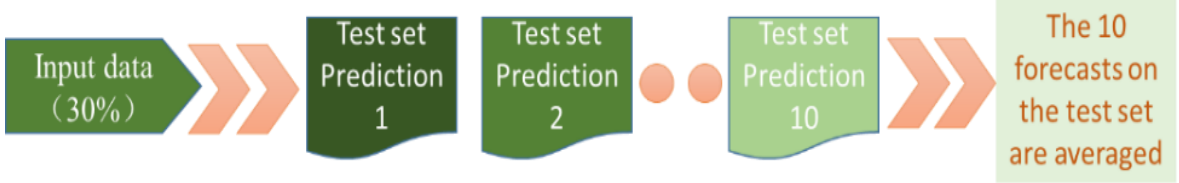

Figure 9: Learning curve during 10-fold cross-validation (test set: 30% of data).

4.2 Comparative analysis of patterns

The above fusion patterns were analyzed and evaluated. This study measured the advantages and disadvantages of each algorithmic model. A confusion matrix based on psychological data was first established for the convenience of discussion. When using the train test split, this article examines the performance of different fusion patterns through 10 modeling times. Firstly, the performance of various patterns on psychological data is compared, as displayed in Figure 10.

Secondly, the data fusion model is compared with the single model, as displayed in Figure 11. According to the 10 modeling performances of each model on psychological data, it is clear that the RLA-stacking model produced by combining random forest, logistic regression, and Adaboost has the highest mean F_1 score and the best prediction effect. The F_1 score fluctuation of this model is also small, indicating its robustness. The RLA-stacking model also shows considerable robustness as it has consistently high F1 scores across different experiments when compared with single models as well as different configurations of fusion. It reflects that it efficiently combines the best of Random Forest, Logistic Regression, and Apriori for better generalization.

The F1 scores assess model performance through ten iterations of cross-validation, thereby ensuring stability and consistency across different training-test splits. Figure 10 shows the F1 score variations for baseline models (Random Forest, Logistic Regression, and Apriori), while Figure 11 highlights the trends in fusion models, revealing performance gains due to varied model combinations. Variability in these scores indicates the models' robustness, where reduced variance points toward improved stability and larger fluctuations indicate sensitivity to data split variations.

While the F1-score balances precision and recall, it does not assess model discrimination at different thresholds; thus, the AUC-ROC was also evaluated. A higher AUC-ROC indicates better class separation, with the RLA-stacking model outperforming single models. The steepness of its ROC curve shows high recall with little incidence of false positives, making it highly appropriate for the prediction of psychological risk. Using both the F1-score and AUC-ROC allows for a more comprehensive model assessment.

Following preliminary screening, nine key features were established based on statistical importance, removal of collinearity, and expert validation. Features were eliminated based on importance examination, removal of highly correlated variables (VIF > 5), and expert review. These features capture psychological distress and behavioral risks as well as social influences quite suitably and ensure the model can identify key predictors of crisis.

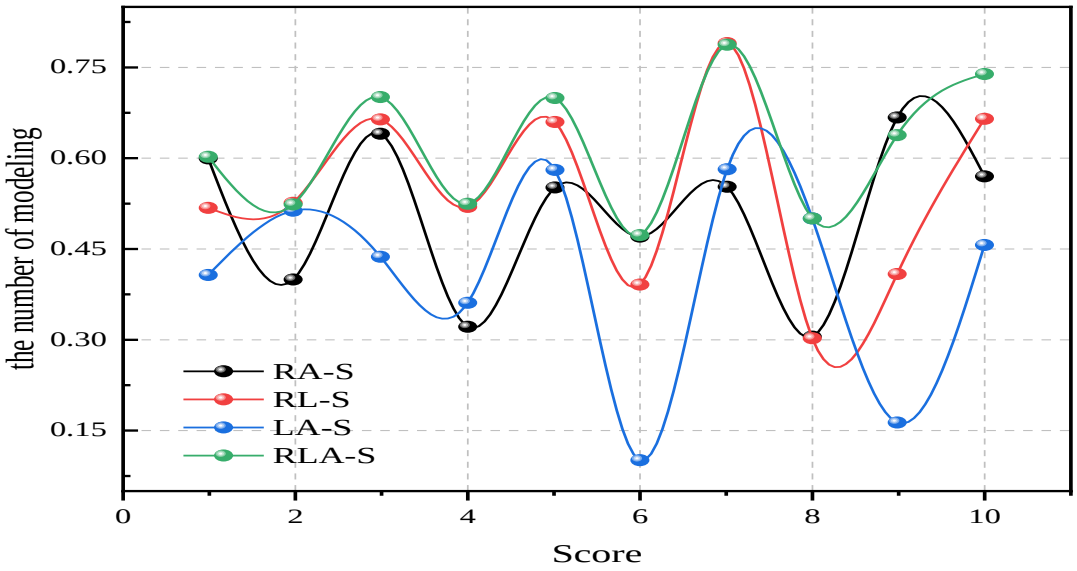

Figure 10: Comparison of F1 scores of different fusion patterns.

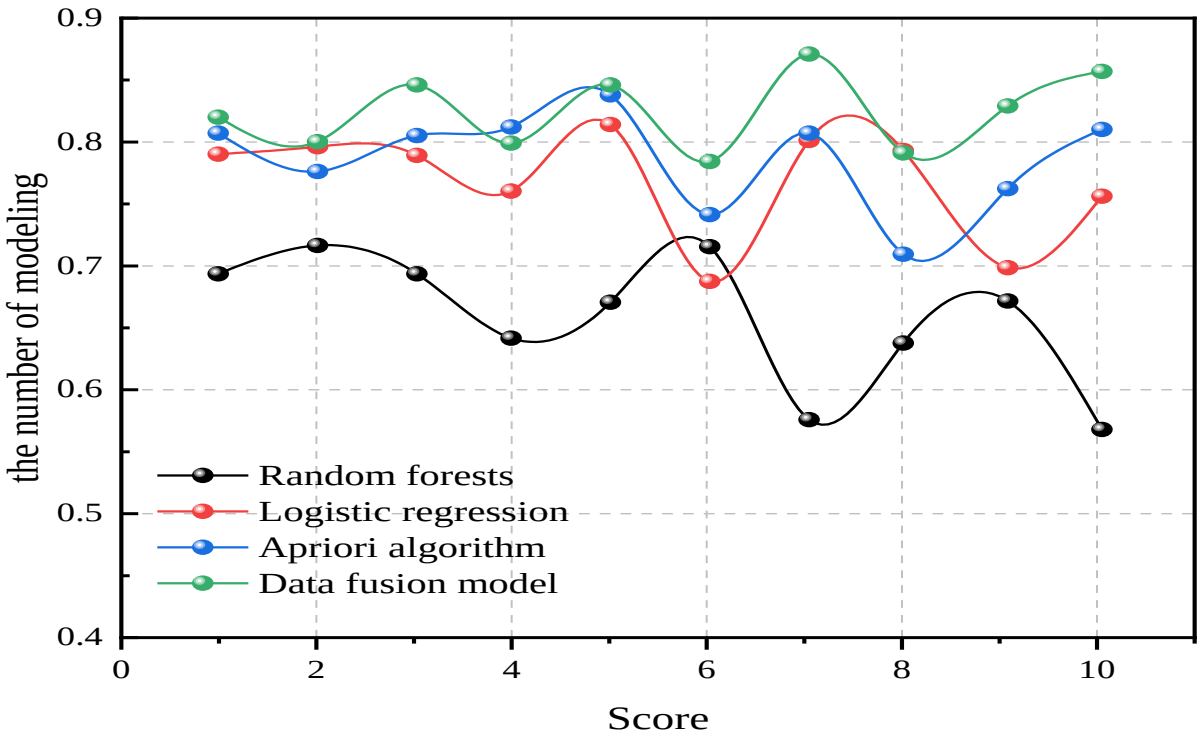

Figure 11: Comparison of scores between data fusion model and single model.

5 Practice and application of college students' psychological early warning system under data fusion model

A psychological crisis early warning system for college students is based on a data fusion model consisting of a presentation, business logic, and data access layer, as displayed in Figure 12. The fusion model integrates real-time psychological well-being assessments into the decision process. Student survey responses and behavioral data are collected, preprocessed, and evaluated by three base models whose outputs are consolidated to produce a holistic risk profile. These predictive evaluations are used to alert mental health professionals to initiate interventions for at-risk students. Feedback provided continuously improves the model, thereby increasing its accuracy incrementally with time.

The college students' psychological crisis early warning system utilizes MySQL as its database [39]. The system organizes psychological information into a structured database, combining user records, historical data, model-generated outcomes, and assessments. This enables the instantaneous determination of risk and supports mental health professionals in making informed decisions. The system's core module is the prediction function, which involves converting the trained data fusion model into a PMML (Predictive Model Markup Language) file using Sklearn2PMML as the initial step. Secondly, the PMML model file is integrated into the system. Finally, JPMML loads this model and predicts students' psychological crises based on their psychological file information.

The PMML standardizes the trained fusion model for easy integration, while JPMML enables real-time risk assessment without re-training. The functionality supports fast, platform-independent deployment, instant risk scoring, and model updates without disrupting the system. Together, these elements ensure the model's effectiveness in monitoring mental health in real-world environments.

6 Discussion

The suggested data fusion model outperforms the state-of-the-art methodologies presently being used in the field of psychological crisis detection in college students. Previous studies have applied different machine learning methods, such as decision trees, logistic regression, random forests, and association rule mining using the Apriori algorithm. More recent studies, such as those by Hinduja et al. [13] and Jiang et al. [20], have explored deep learning and artificial intelligence-driven big data approaches, achieving F1 scores of 89.5% and 91.4%, respectively. However, individual models often face issues of feature redundancy, overfitting, and class imbalance that limit their predictive performance.
Alternatively, the fusion model integrates random forest, logistic regression, and Apriori to harness the respective strengths provided by each method. It performs at an accuracy rate of 95.2%, with precision at 93.7%, recall at 90.8%, and F1 score at 92.2%, improving on previous work. This is attributable to its ensemble learning in removing bias, rectifying class imbalance issues, and boosting predictions due to optimum data preprocessing techniques. Apart from numerical improvements, the model provides pragmatic value to university campus mental health services by enabling proactive

interventions. Nevertheless, data privacy and transparency concerns must be addressed to enable responsible implementation. Potential future enhancements include the integration of real-time behavioral data and adaptive learning capabilities in order to enhance prediction accuracy.

7 Conclusion

This study utilized family theory and life event theory within psychology to create psychological profiles for data analysis. Various machine learning algorithms, such as decision trees, random forests, logistic regression, and the Apriori algorithm, were applied to develop a model for identifying psychological anomalies among college students. Upon evaluating the efficacy of these patterns against psychological data, the individual patterns were integrated using the data fusion technique. The outcomes indicated that the psychological profiles of college students, formulated based on family theory and life event theory, offered superior psychological attributes for the analysis and contributed to the improvement of the predictive capacity of the algorithmic model. When applied to mental data, the Apriori algorithm performed better than random forest, logistic regression, and decision trees. Moreover, the La-stacking fusion model, incorporating logistic regression and the Apriori algorithm as base patterns, exhibited better performance than the Apriori algorithm alone when applied to psychological data. This implies that the stacked fusion model may have a reduced predictive capability compared to the individual algorithm. The data fusion model achieved an average F1 score of 82.2% after 10 rounds of modeling using random forest, logistic regression, and the Apriori algorithm as base patterns. This improvement in prediction ability over single patterns was particularly notable in psychological data analysis.

This study emphasizes the effectiveness of the data fusion model, yet we should consider its limitations as well. Survey-based response dependency could lead to reporting bias and limiting the generality of the dataset to academia only. Additionally, the lack of live behavioral data restricts the scope for better crisis identification. Future studies should include dynamic data sources and test the model in diversified populations to generalize its results better.

Future studies should include additional data sources because the current estimates rely on surveys with inherent self-reporting bias and have low temporal resolution. Real-time behavior data, social media use, and wearable performance readings could enable continuous and non-intrusive surveillance and enable earlier and more accurate identification of crises. Digital messages and public sentiments might reveal subtle warning signs that infrequently conducted surveys might miss. A combination of such data sources would improve accuracy, reduce bias, and support the effectiveness of timely interventions.

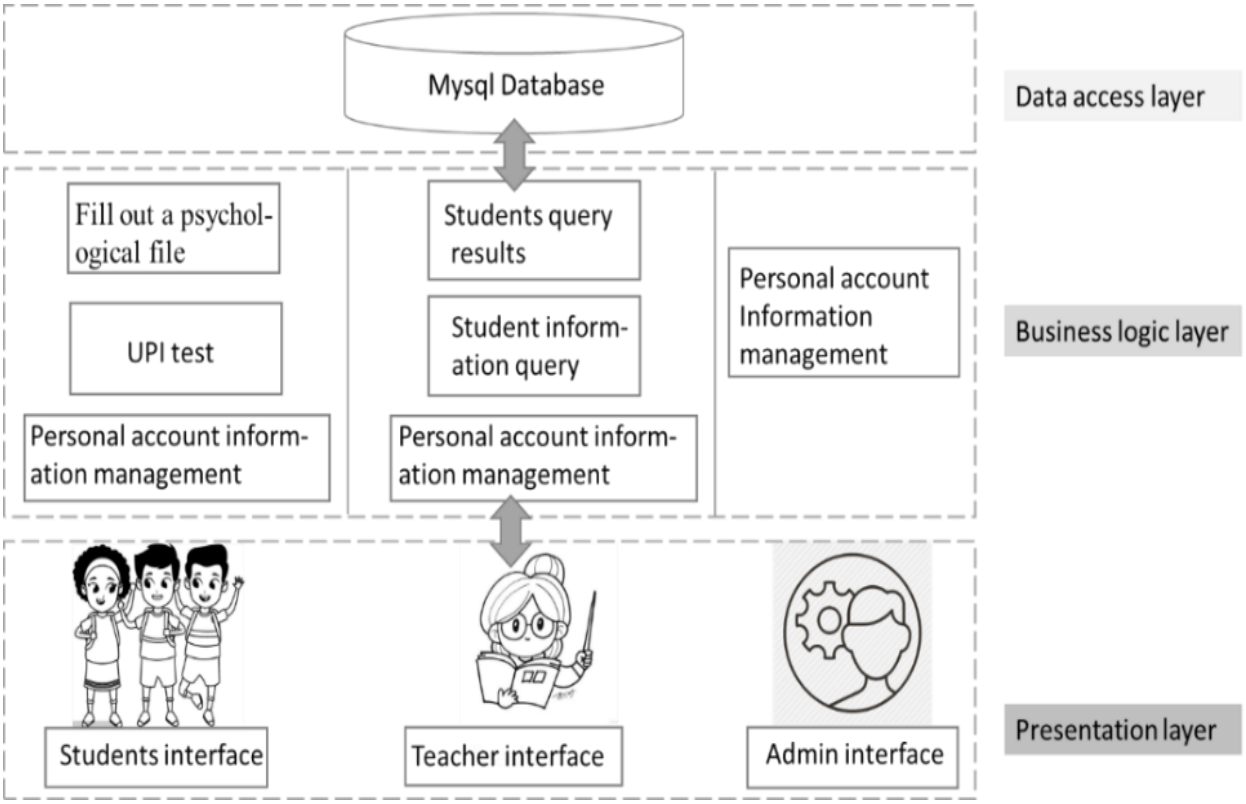

Figure 12: Design of college students' psychological crisis early warning system.

Nomenclature

Abbreviation	
UPI	University Personality Inventory
SPSS	Statistical Package for the Social Sciences
BP	Back Propagation
IoT	Internet of Things
AI	Artificial Intelligence
DM-BPNN	Data Mining-Back Propagation Neural Network
CLS	Concept Learning System
ID3	Iterative Dichotomiser 3
GIM	Gini Impurity Measure
VIM	Variable Importance Measure
TVB-N	Total Volatile Basic Nitrogen
LSTM	Long Short-Term Memory
PMML	Predictive Model Markup Language
JPMML	Java Predictive Model Markup Language

Symbol	
t	node in decision tree algorithms
p(i|t)	the probability of class i given node t
I(parent)	Entropy of the parent node
N	Total number of records
Δinfo	Information gain
SplitInfo	the split Information
Gini(t)	Gini coefficient of node t
K	Total number of partitions
n	Number of trees in the random forest
c	Number of different classes
Entropy(t)	Information entropy of node t

Subscript & Superscript	
δinfo	Information gain
N(Vj)	Number of records in partition Vj
P(vi)	Probability of partition vi
P[i/t]2	Probability of class i given node t squared
GIm	Gini index for feature mmm
Pvi	Pvi the probability of being in partition i
VIMjmGini ,VIMijGini ,VIMjGini	Variable Importance Measures related to the Gini index in different contexts within a Random Forest model
GIl, GIr	Gini index of the two new nodes after branching

Authorship contribution statement

Yan Wu: Writing - Original draft preparation, Conceptualization, Supervision, Project administration.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author Statement

The manuscript has been read and approved by all the authors, the requirements for authorship, as stated earlier in this document, have been met, and each author believes that the manuscript represents honest work.

Ethical approval

All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

References

[1] X. Chen, "Dynamic Early-Warning Model of College Students' Psychological Crisis Based on Characteristic Attribute.," *Sci Program*, 2022.

[2] R. Liu, "Early Warning Model of College Students' Psychological Crises Based on Big Data Mining and SEM," *International Journal of Information Technologies and Systems Approach*, vol. 16, no. 2, pp. 1–17, 2023.

[3] S. H. I. Xiu-Mei, "A Review of Researches on Early Warning Index System of Chinese College Students' Psychological Crisis," *DEStech Transactions on Social Science, Education and Human Science*, 2020.

[4] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the national academy of sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.

[5] X. Wang, "Design and Optimization of Psychological Emergency Response System for College Students Based on IoT and Computational Intelligence," *Mobile Information Systems*, vol. 2022, 2022.

[6] L. Cai, "A novel recognition model of university students' psychological crisis based on DM," *Sci Program*, vol. 2022, pp. 1–10, 2022.

[7] T. Kolenik, G. Schiepek, and M. Gams, "Computational psychotherapy system for mental health prediction and behavior change with a conversational agent," *Neuropsychiatr Dis Treat*, pp. 2465–2498, 2024.

[8] T. Kolenik and M. Gams, "Intelligent cognitive assistants for attitude and behavior change support in mental health: state-of-the-art technical review," *Electronics (Basel)*, vol. 10, no. 11, p. 1250, 2021.

[9] L. Yang, "Research on strategies of promoting mental health of higher vocational college students based on data mining," *Wirel Commun Mob Comput*, vol. 2022, 2022.

[10] M. Xun, "Establishing Early Warning System of College Students' Mental Health Based on System Dynamics," in *2021 International Symposium on Advances in Informatics, Electronics and Education (ISAIEE)*, IEEE, 2021, pp. 38–41.

[11] J. Wang, Z. Zhang, H. Luo, Y. Liu, W. Chen, and G. Wei, "Research on early warning model of college students' psychological crisis based on genetic BP neural network," *American journal of applied psychology*, vol. 8, no. 6, pp. 112–120, 2019.

[12] C. Xiang-Wei, "Modeling of Vocational College Students' Mental Health Based on Big Data Analysis," *CONVERTER*, pp. 620–626, 2021.

[13] S. Hinduja, M. Afrin, S. Mistry, and A. Krishna, "Machine learning-based proactive social-sensor service for mental health monitoring using twitter data," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100113, 2022.

[14] P. Li, "The Mental Health Evaluation System of College Students Based on Data Mining," *Sci Program*, vol. 2022, no. 1, p. 3800169, 2022.

[15] Z. Zhang and R. Zhang, "Multimedia data mining," in *Data Mining and Knowledge Discovery Handbook*, Springer, 2010, pp. 1081–1109.

[16] H. Hadjar *et al.*, "Video-based automated emotional monitoring in mental health care supported by a generic patient data management system.," in *PSYCHOBIT*, 2020.

[17] X. Sun, "Application of data mining technology in college mental health education," *Front Psychol*, vol. 13, p. 974576, 2022.

[18] F. Castanedo, "A review of data fusion techniques," *The scientific world journal*, vol. 2013, 2013.

[19] U. Khulal, J. Zhao, W. Hu, and Q. Chen, "Intelligent evaluation of total volatile basic nitrogen (TVB-N) content in chicken meat by an improved multiple level data fusion model," *Sens Actuators B Chem*, vol. 238, pp. 337–345, 2017.

[20] W. Jiang and R. Yuankun, "Dynamic Early Warning System of Ideology Education Based on Psychological Big Data Analysis," in *Application of Big Data, Blockchain, and Internet of Things for Education Informatization*, Y. Zhang and N. Shah, Eds., Cham: Springer Nature Switzerland, 2024, pp. 326–333.

[21] Y. Yan, "Analysis and Research of Psychological Crisis Behavior Model Based on Improved Apriori Algorithm," *Int J Hum Comput Interact*, pp. 1–13, doi: 10.1080/10447318.2024.2320981.

[22] Z. Tian and D. Yi, "Application of artificial intelligence based on sensor networks in student mental health support system and crisis prediction," *Measurement: Sensors*, vol. 32, p. 101056, 2024, doi: https://doi.org/10.1016/j.measen.2024.101056.

[23] C. Sheng, "Simulation application of sensors based on Kalman filter algorithm in student psychological crisis prediction model," *Measurement: Sensors*, vol. 33, p. 101190, 2024, doi: https://doi.org/10.1016/j.measen.2024.101190.

[24] H. Ni, Y. Zhu, and L. Yang, "Data-Driven Mental Health Assessment of College Students Using ES-ANN and LOF Algorithms During Public Health Events," *Informatica*, vol. 49, no. 13, 2025.

[25] Y. Zhang, "Psychological Fitness Education Driven by Artificial Intelligence Technology and Its Influence on Education Assessment," *Informatica*, vol. 48, no. 11, 2024.

[26] Z. Jigang and W. Bianjiang, "The current situation and development trend of data mining research [J]," *Journal of Honghe University*, vol. 1, pp. 102–104, 2010.

[27] M. K. Gupta and P. Chandra, "A comprehensive survey of data mining," *International Journal of Information Technology*, vol. 12, no. 4, pp. 1243–1257, 2020.

[28] J. Esteban, A. Starr, R. Willetts, P. Hannah, and P. Bryanston-Cross, "A review of data fusion models and architectures: towards engineering guidelines," *Neural Comput Appl*, vol. 14, pp. 273–281, 2005.

[29] T. Kolenik, "Methods in digital mental health: smartphone-based assessment and intervention for stress, anxiety, and depression," in *Integrating Artificial Intelligence and IoT for Advanced Health Informatics: AI in the Healthcare Sector*, Springer, 2022, pp. 105–128.

[30] H. Zheng, Y. Cheng, and H. Li, "Investigation of model ensemble for fine-grained air quality prediction," *China Communications*, vol. 17, no. 7, pp. 207–223, 2020.

[31] H. Gao *et al.*, "Network attacks identification method of relay protection devices communication system based on Fp-Growth algorithm," in *2022 IEEE Sustainable Power and Energy Conference (iSPEC)*, IEEE, 2022, pp. 1–6.

[32] J. Ding, Q. Liu, M. Bai, and P. Sun, "A multisensor data fusion method based on gaussian process model for precision measurement of complex surfaces," *Sensors*, vol. 20, no. 1, p. 278, 2020.

[33] M. K. Singh, A. Dutta, and K. S. Venkatesh, "Multi-sensor data fusion for accurate surface modeling," *Soft comput*, vol. 24, pp. 14449–14462, 2020.

[34] H. Zhang, K. Jiang, C. Cheng, J. Cao, and W. Zhang, "Multi-source Heterogeneous Data Fusion Model Based on FC-SAE," *Journal of Internet Technology*, vol. 23, no. 7, pp. 1473–1481, 2022.

[35] W. Wen and H. F. Durrant-Whyte, "Model-based multi-sensor data fusion," in *Proceedings 1992 IEEE international conference on robotics and automation*, IEEE, 1992, pp. 1720–1726.

[36] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP J Adv Signal Process*, vol. 2016, pp. 1–16, 2016.

[37] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Information Fusion*, vol. 57, pp. 115–129, 2020.

[38] T. Kolenik, "Methods in digital mental health: smartphone-based assessment and intervention for stress, anxiety, and depression," in *Integrating Artificial Intelligence and IoT for Advanced Health Informatics: AI in the Healthcare Sector*, Springer, 2022, pp. 105–128.

[39] M. Atiquzzaman, N. Yen, and Z. Xu, *Big data analytics for cyber-physical system in smart city:*

BDCPS 2019, 28-29 December 2019, Shenyang, China, vol. 1117. Springer Nature, 2020.