Research on Psychological Crisis Early Warning System for College Students Integrating Big Data and Machine Learning

Ziyan Yao, Jin Gan
Guangxi Eco-Engineering Vocational and Technical College, Liuzhou 545004, Guangxi, China
Corresponding author: Jin Gan (e-mail: ganjin23@163.com)

ABSTRACT This study develops a psychological crisis early-warning system for college students by integrating big data technologies and machine learning methods. A multi-source data acquisition framework is established to collect and process behavioral, academic, social, and psychological information, while a Dual-Channel CNN-FM Feature Fusion Model (DCCFM-FFM) is employed to enhance crisis prediction accuracy through heterogeneous feature extraction and fusion. The system incorporates distributed storage, real-time data processing, edge computing, and intelligent risk assessment mechanisms to support dynamic monitoring and timely intervention. Experimental results demonstrate that the proposed framework achieves high predictive performance, with accuracy, recall, and F1-score significantly outperforming conventional early-warning approaches. The architecture is particularly suitable for communication-intensive environments supported by wireless communication infrastructures and antenna-enabled information transmission networks, where reliable data acquisition, low-latency processing, and secure information exchange are essential for intelligent decision-making. The proposed system provides an effective engineering solution for large-scale psychological risk monitoring and contributes to the development of intelligent health-management and communication-oriented data analysis platforms.

INDEX TERMS big data analytics, machine learning, psychological crisis early warning, wireless communication infrastructure

I. INTRODUCTION

A. RESEARCH BACKGROUND

IN recent years, with the rapid development of the social economy and intensifying competitive pressures, mental health issues among college students have become increasingly prominent. These societal trends similarly affect the industrial workforce: rapid economic changes and fierce global competition, particularly in demanding sectors , contribute to intensifying competitive pressures and an increase in mental health challenges among employees. The frequent occurrence of psychological crisis events has emerged as a focal point of concern across society. Facing pressures from academic demands, employment challenges, emotional burdens, and financial constraints, college students' mental health is severely affected, leading some to resort to extreme behaviors like self-harm or suicide due to ineffective coping [1,2]. For instance, certain university students have ultimately suffered severe psychological disorders or even tragic outcomes due to chronic sleep deprivation, strained interpersonal relationships, or academic difficulties [3]. These incidents not only cause irreversible harm to individuals but also negatively impact campus environments and social stability. Against this backdrop, establishing a scientific and efficient psychological crisis early-warning system has become crucial. Traditional early-warning mechanisms primarily rely on manual screening and questionnaire surveys, which suffer from issues like limited data sources, insufficient real-time monitoring, and poor accuracy—making them inadequate for meeting the complex and evolving needs of mental health management. Therefore, how to leverage modern information technology to enhance psychological crisis early-warning capabilities has become an urgent key challenge requiring resolution.

B. SIGNIFICANCE OF THE STUDY

This study aims to develop a psychological crisis early-warning system for college students by integrating big data and machine learning technologies, which holds significant theoretical and practical value. Firstly, from the perspective of safeguarding students' mental health, the system can comprehensively collect and analyze daily behavioral data to promptly identify potential psychological risks and provide targeted interventions, thereby effectively preventing and reducing the occurrence of psychological crises [4]. Secondly, in maintaining campus stability, the application of this psychological crisis early-warning system facilitates dynamic monitoring and precise management of students' mental states, reducing the likelihood of campus safety incidents caused by psychological issues and fostering a more harmonious and secure environment [5]. Additionally, this research provides new approaches and methods for university mental health education. By integrating multi-source data with intelligent algorithms, it significantly enhances the scientific rigor and effectiveness of mental health education, driving the development of intelligent and refined mental health service systems in higher education. In summary, this study not only addresses prominent challenges in current student mental health care but also lays a solid foundation for improving the mental health education framework in universities.

II. LITERATURE REVIEW

A. BIG DATA AND MACHINE LEARNING RELATED THEORIES

Big data refers to massive, diverse, and complex datasets characterized by the "4Vs": Volume, Velocity, Variety, and Value. In data collection, big data technologies gather user behavior data through multiple channels including sensors, social media platforms, and campus management systems, integrating it with structured information from traditional databases [6]. At the storage level, distributed systems like Hadoop and Spark are widely used to process vast amounts of unstructured data, ensuring efficient management and access performance [7]. Furthermore, data analysis methods such as data cleaning, feature extraction, and deep mining effectively extract valuable insights from raw data, providing crucial support for decision-making processes.

B. RESEARCH STATUS OF PSYCHOLOGICAL CRISIS EARLY WARNING SYSTEM FOR COLLEGE STUDENTS

In recent years, scholars both domestically and internationally have conducted extensive research on psychological crisis early warning systems for college students. Regarding system architecture, existing studies typically adopt a hierarchical design model encompassing data collection, processing, analysis, and application layers. For instance, a mental health monitoring platform utilizing big data technology categorizes data collection into three dimensions: registration data, psychological assessment data, and behavioral data, thereby forming a comprehensive raw data repository. In terms of data sources, researchers tend to integrate multi-source information such as academic records from the academic affairs system, access control card data, and social media content to enhance the comprehensiveness and accuracy of early warning models. Meanwhile, the design of early warning indicators has progressively expanded from single-dimensional approaches to multi-dimensional assessments, covering aspects like sleep quality, interpersonal sensitivity, and academic pressure.

Current trends indicate that college students' psychological crisis early warning systems are advancing toward intelligent and precise development. On one hand, the integration of machine learning technologies has significantly enhanced system prediction capabilities. For instance, the application of convolutional neural networks (CNN) for text analysis has been proven to substantially improve the accuracy of psychological crisis identification. On the other hand, interdisciplinary collaboration has become a key driver in this field. The deep integration of psychology, computer science, and statistics has made these early warning systems more scientifically reliable [8].

III. KEY TECHNOLOGIES OF BIG DATA AND MACHINE LEARNING

A. BIG DATA TECHNOLOGY

1) Data Collection

Data collection forms the cornerstone of establishing a psychological crisis early-warning system for college students, with its core mission being to gather information related to students' mental health from diverse and heterogeneous data sources. Social media platforms, serving as vital channels for contemporary students to express emotions and document daily experiences, provide abundant textual data for crisis detection. Through API interfaces or web crawling technologies, educators can extract public content shared by students on platforms like Weibo and WeChat, extracting emotional vocabulary, behavioral patterns, and social networks. Additionally, campus management systems constitute another critical data source, including academic performance fluctuations and course preferences in academic records, access logs from campus security systems, and dietary habits and spending behaviors tracked in consumption systems. These structured datasets reflect students' daily behavioral patterns, providing objective evidence for psychological risk assessment. Sensor technology further expands data dimensions through wearable devices monitoring physiological indicators (e.g., heart rate and sleep quality) and environmental sensors tracking stress level changes, enabling dynamic psychological state monitoring. The integration of multi-channel data not only enhances comprehensive coverage but also establishes a robust foundation for developing subsequent psychological crisis prediction models.

Crucially, the ethical and legal compliance of multi-channel data collection, especially from sensitive sources like sensors and social media, is paramount. To mitigate privacy and ethical risks, this system adheres to strict protocols:

This study was approved by the Ethics Committee of Guangxi Vocational and Technical College of Ecological Engineering, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

Minimal Data Collection and Anonymization: The system strictly adheres to the principle of minimal necessary data collection. Sensitive identifiers are immediately removed or replaced with robust, irreversible anonymization or pseudonymous tokens upon ingestion. Data access is restricted via role-based access control (RBAC) and monitored through audit logs to ensure security and prevent misuse.

Data Security and Access: All collected data is stored using the encryption and security measures detailed in Section “Data Storage”, guaranteeing security during transmission and storage, and only accessible to authorized psychological counselors or administrative personnel for intervention purposes.

2) Data Storage

Faced with massive and diverse psychological data from college students, traditional storage methods struggle to meet the demands of efficient management and rapid access. Therefore, it is necessary to introduce distributed storage systems for effective data storage and management. Distributed file systems (such as Hadoop Distributed File System, HDFS) enhance storage capacity while improving system fault tolerance and concurrent processing capabilities by distributing data across multiple nodes. Additionally, NoSQL databases like MongoDB and Cassandra have become ideal choices for storing unstructured data (e.g., text and images) due to their flexible data models and high scalability.

3) Data Analysis

Data analysis serves as the cornerstone of big data technology, aiming to extract valuable insights from raw collected data to provide scientific foundations for psychological crisis early warning systems. The process begins with cleaning and preprocessing the data to eliminate noise and redundant information. Key procedures include missing value filling, outlier detection and correction, and data normalization, all designed to enhance data quality and consistency . Subsequently, data mining techniques are applied to conduct in-depth analysis of processed data, extracting characteristics related to psychological crises. For instance, association rule mining can uncover hidden connections between student behavioral patterns, while cluster analysis groups students with similar psychological traits. Additionally, time series analysis helps track changes in students’ mental states, identify abnormal fluctuations, and trigger timely alerts.

B. MACHINE LEARNING TECHNIQUES

1) Principles of Common Algorithms

Machine learning algorithms play a central role in psychological crisis early warning systems, where their fundamental principles determine the model’s learning capacity and predictive performance. The decision tree algorithm, a tree-structured classification method, recursively divides datasets into subsets, with internal nodes representing attribute tests and leaf nodes indicating classification outcomes. While decision trees excel in their intuitive and interpretable nature, they are prone to overfitting, particularly when dealing with high data noise. Neural networks, computational models mimicking human brain neuron connections, consist of input, hidden, and output layers. Through backpropagation algorithms that continuously adjust weight parameters, these networks can learn complex nonlinear relationships and handle high-dimensional data effectively.

2) Algorithm Selection Basis

In psychological crisis early warning scenarios, the suitability of different machine learning algorithms depends on multiple factors including data characteristics, task requirements, and computational resources. Decision trees, with their simple structure and strong interpretability, are particularly suitable for preliminary exploration of data features and building baseline models. For instance, when analyzing the relationship between students’ campus consumption behaviors and mental health, decision trees can effectively identify key behavioral patterns and generate easily understandable rule sets. However, when dealing with high-dimensional nonlinear data, neural networks demonstrate stronger learning capabilities. Their multi-layer perceptron architecture can capture complex semantic information, thereby improving the accuracy of early warning models. Support Vector Machines (SVMs) excel in scenarios with small sample sizes but high feature dimensions, such as in psychological state assessments based on physiological signals.

3) Model Training and Optimization

Model training is a critical component of machine learning applications, aiming to optimize model parameters through iterative adjustments to achieve optimal performance on given datasets. First, collected data should be divided into three subsets: the training set for model learning, the validation set for hyperparameter tuning, and the test set for evaluating generalization capabilities. During training, cross-validation techniques help prevent bias caused by uneven data partitioning, thereby enhancing model stability. For neural network models, common optimization algorithms like stochastic gradient descent (SGD) and Adam optimizers accelerate convergence while reducing local optima through dynamic learning rate adjustments . To further improve performance, regularization techniques (e.g., L1/L2 regularization) can prevent overfitting, or ensemble learning methods (e.g., random forests, gradient boosting trees) may be employed to combine predictions from multiple base models, thus improving robustness and accuracy. During optimization phases, quantitative analysis of evaluation metrics (e.g., accuracy, recall, F1 score) helps refine model architecture or parameter settings based on feedback. Through continuous iterative optimization, a high-performance model capable of meeting psychological crisis prediction requirements is ultimately developed.

IV. CONSTRUCTION OF EARLY WARNING SYSTEM INTEGRATING BIG DATA AND MACHINE LEARNING

A. DATA COLLECTION MODULE

1) Multi-Channel Data Integration

The data sources for the psychological crisis early warning system for college students are diverse and complex, necessitating multi-channel integration to ensure comprehensive coverage. Key data channels include campus administration systems, access control systems, and social media platforms. Academic management systems provide student records such as academic performance, class attendance rates, and course preferences. Access control systems document daily activity patterns and dormitory entry frequency, reflecting students' living habits and social behaviors. Additionally, unstructured data like text messages, emojis, and interaction patterns on social media platforms offer critical clues for crisis detection. By effectively integrating these multi-source heterogeneous data, a comprehensive database covering academic, lifestyle, and social dimensions can be established, laying a solid foundation for subsequent psychological analysis. However, during data integration, it's crucial to address format inconsistencies and semantic conflicts across data sources, implementing standardized normalization processes to prevent analytical errors caused by data discrepancies.

2) Data Preprocessing

Raw data often contains issues such as noise, missing values, and redundant information, which can severely impact the performance of subsequent machine learning models. Therefore, it is essential to perform preprocessing operations like cleaning, denoising, and normalization on collected data. Data cleaning aims to remove invalid or erroneous records, such as outliers or duplicate entries caused by system failures. Data denoising involves eliminating irrelevant interference through statistical methods or filtering algorithms, including irrelevant keywords or high-frequency stop words. Additionally, since different data sources may have significantly varying feature scales, normalization is required to map data into a unified range, thereby improving model convergence speed and prediction accuracy. Common normalization methods include minmax normalization and Z-score standardization.

B. MACHINE LEARNING MODEL CONSTRUCTION

1) Model Selection Rationale

The primary goal of this study is to predict psychological crises using heterogeneous, multi-channel data, which includes structured time-series behavioral data and sparse, high-dimensional social/textual features. Initial exploration with traditional machine learning methods (e.g., Support Vector Machine, Decision Tree) demonstrated limitations in effectively capturing complex, nonlinear relationships and temporal dependencies inherent in these diverse data sources. Considering the need for high predictive accuracy and robustness in a real-time early warning system, this study adopts an advanced deep neural network architecture as the core predictive algorithm. Specifically, we selected and developed the Dual-Channel CNN-FM Feature Fusion Model (DCCFM-FFM) to address the unique challenges of integrating temporal and sparse interaction features simultaneously. This integrated model is engineered to overcome the deficiencies of single-channel models by leveraging specialized components for each data type, ensuring comprehensive and precise prediction capability.

2) Feature Extraction and Selection

Extracting psychological crisis-related features from preprocessed data is a critical step in building machine learning models. First, we need to analyze raw data to identify potential indicators of psychological crises, such as academic stress metrics (e.g., number of failing courses and fluctuations in GPA), social isolation indicators (e.g., campus activity participation and social media engagement frequency), and emotional state indicators (e.g., negative word usage frequency and emoji distribution). Building on this foundation, we employ feature selection algorithms to identify the most representative key features, thereby reducing model complexity and enhancing generalization capabilities. Common feature selection methods include chisquare testing based on filters, recursive feature elimination using wrapper methods, and Lasso regression with embedding techniques. Preliminary feature analysis, focusing on the combination of structured and unstructured data, suggests that critical indicators of psychological distress typically span key behavioral, academic, and linguistic dimensions. For instance, high-impact features often include:Unstructured Data (Social Media): The frequency of negative emotional keywords and linguistic markers of distress in student posts.
Behavioral Data (Access Control): The number of abnormal late-night entry/exit or unrecorded dormitory access anomalies (reflecting sleep and security pattern deviations).
Academic Data (Academic Records): The magnitude of unexpected GPA fluctuation across semesters, which is more predictive than a consistently low GPA score.
Behavioral Data (Consumption/Activity): A sudden, uncharacteristic drop in daily average consumption (e.g., dining hall) or library visit frequency.

Experimental results demonstrate that proper feature selection not only significantly improves predictive performance but also reduces computational costs, providing technical support for real-time early warning systems.

3) Dual-Channel CNN-FM Model Architecture (DCCFM-FFM)

Considering the heterogeneous nature of the collected data, which includes structured records (academic, behavioral) and unstructured text (social media), this study adopts a Dual-Channel CNN-FM Feature Fusion Model (DCCFM-

FFM) as the core machine learning algorithm. This architecture is specifically designed to leverage the distinct characteristics of each data type before fusing the extracted features for the final classification task.

The DCCFM-FFM comprises three main components:
Unstructured Data Channel (Channel 1 - Temporal/Text Analysis): This channel is dedicated to handling textual data (e.g., social media posts). It employs a 1D Convolutional Neural Network (1D-CNN) layer to extract local patterns (like n-grams) and semantic features from word embeddings, capturing emotional and linguistic distress markers.
Structured Data Channel (Channel 2 - Feature Interaction Analysis): This channel handles quantitative features (e.g., GPA fluctuation, access log anomalies). This channel is explicitly designed as a Factorization Machine (FM) layer to efficiently model complex non-linear relationships and high-order feature interactions among the behavioral and academic indicators, which is superior to standard MLP approaches for sparse data.
Feature Fusion Layer: The feature vectors extracted from both Channel 1 and Channel 2 are concatenated and passed through a final fully connected layer. This layer learns the optimal combination of insights from both data channels to produce the final psychological crisis risk score, thereby enhancing the model's robustness and predictive accuracy.

4) Detailed Implementation and Training Strategy

To ensure the reproducibility and rigor of the model training process, the DCCFM-FFM was implemented with the following detailed specifications:
Loss Function and Class Imbalance Handling: Given the critical nature of the positive class and the severe class imbalance (4.5% positive cases), the model utilizes a Weighted Binary Cross-Entropy (W-BCE) loss function. Class weights were dynamically set to prioritize the correct prediction of positive cases, mitigating bias towards the majority (Negative) class.
Optimizer and Learning Rate: The Adam optimizer was used with an initial learning rate (α) of 1×10^{-3}. A learning rate decay strategy was applied: the rate was halved (decay factor of 0.5) if the validation loss did not improve for 5 consecutive epochs (patience = 5).
Regularization and Early Stopping: L2 regularization ($\lambda = 1\times10^{-4}$) was applied to the weights of all dense layers to prevent overfitting. Training was stopped using the Early Stopping callback with a patience of 10 epochs, monitoring the Validation Set F1-Score to identify the optimal model checkpoint.
Channel 1 (1D-CNN) Specifications for Textual Data:
Embedding Dimensions: Text was converted using pre-trained Word2Vec embeddings (100 dimensions), fine-tuned on the university's social media corpus.
Sequence Handling: The input sequence length was fixed at 100 tokens; longer sequences were truncated, and shorter ones were zero-padded.
CNN Architecture: This channel consists of two sequential 1D Convolutional layers. Each layer contains 128 filters/kernels (size of 3) and uses a ReLU activation function. The final output is flattened via a Global Max Pooling layer.
Channel 2 (FM) Specifications for Structured Data: The Factorization Machine component was configured with a latent interaction vector dimension (k) set to 8. This layer output is then passed to a standard Dense layer of size 64 with ReLU activation before fusion.
Fusion Layer: The concatenated feature vector is passed through a Dense layer (size 32, ReLU activation) before the final output layer (size 1, Sigmoid activation) to produce the probability of crisis.

5) Model Selection and Training

Based on the results of feature selection, it is necessary to determine machine learning models suitable for psychological crisis early warning scenarios. Commonly used models include decision trees, neural networks, and support vector machines (SVMs). Decision tree models offer good interpretability, allowing intuitive visualization of the relationship between features and prediction outcomes, but they are prone to overfitting and sensitive to noise. Neural networks demonstrate strong nonlinear fitting capabilities, making them suitable for handling complex nonlinear relationships, though their training process is time-consuming and requires substantial annotated data. Support vector machines excel in small-sample scenarios but exhibit low efficiency when processing large datasets. Considering practical requirements for psychological crisis early warning, this study adopts the Dual-Channel CNN-FM Feature Fusion Model (DCCFM-FFM) as the core algorithm and utilizes annotated data for training. During training, hyperparameters such as learning rate, batch size, and number of iterations were adjusted to optimize model convergence speed and prediction accuracy .

6) Model Evaluation and Optimization

To evaluate the performance of trained machine learning models, a comprehensive set of metrics including accuracy, recall rate, and F1 score should be utilized. For a robust and unbiased evaluation, the entire annotated dataset was first strictly partitioned using a person-based splitting strategy to prevent data leakage, ensuring all records belonging to the same student ID remained within the same subset (e.g., training or test). A fixed random seed (set to 42) was utilized to ensure the reproducibility of this initial data split. The resulting dataset was then statically partitioned into three distinct sets: 85% for the Training and Validation Pool and 15% for the reserved, independent Test Set. The training set was used to train the model, the validation set for hyperparameter tuning and cross-validation, and the test set for a final, unbiased performance assessment. To rigorously fine-tune hyperparameters and validate the model architecture (DCCFM-FFM), a Stratified 5-Fold Cross-Validation approach was employed exclusively on the 85% Training and Validation Pool. In this CV process, the data pool was dynamically split into five folds: four folds served as the training set, and the remaining fold served as the internal validation set for hyperparameter tuning and early stopping in that specific iteration, thereby maintaining the independence of the final Test Set. This cross-validation procedure was repeated three times to confirm the stability and generalization capability across different subset configurations.

Accuracy reflects the model's ability to correctly classify samples, recall measures its capability to identify positive cases, while the F1 score serves as a balanced average that combines both metrics for holistic evaluation. Experimental results demonstrate that deep neural network-based models achieve high accuracy in psychological crisis prediction tasks, though they still exhibit false positive rates under specific scenarios. To optimize model performance, cross-validation can be employed to fine-tune hyperparameters, while regularization techniques help prevent overfitting. Furthermore, implementing transfer learning by leveraging external datasets to enhance generalization capabilities represents a crucial approach for improving early warning effectiveness.

The entire experimental process, including model training and evaluation, was executed on a consistent hardware and software platform to ensure result comparability and reproducibility:

Hardware Environment: A dedicated server equipped with an Intel Xeon Gold 6248R CPU (3.0 GHz) and an NVIDIA A100 Tensor Core GPU (80GB) was utilized.

Software Environment: The system ran on Ubuntu 20.04 LTS, leveraging Python 3.8, TensorFlow 2.10, and Scikit-learn 1.2 frameworks.

C. EARLY WARNING RESULT OUTPUT MODULE

1) Classification of Early Warning Levels

Based on the output results of machine learning models, psychological crisis risks need to be classified into different levels to enable university administrators to implement targeted interventions. Specifically, the early warning levels can be divided into low-risk, medium-risk, and high-risk categories. Low-risk indicates that students' mental health status is generally good, but potential stressors still require attention; medium-risk suggests that students may have certain psychological issues requiring further evaluation and monitoring; high-risk indicates that students' mental state has reached a critical threshold necessitating immediate emergency intervention [9]. The classification criteria for early warning levels primarily include crisis probability scores from model outputs, abnormality severity of key characteristics, and comparative analysis of historical data. By establishing scientific standards for early warning level classification, we can ensure the accuracy and practicality of early warning information, thereby providing robust support for mental health education initiatives in universities.

2) Early Warning Information Push

Timely delivery of early warning information to university counselors, psychological consultants, and relevant personnel is a crucial component in achieving early intervention for mental health crises. To this end, it is essential to establish an efficient alert notification system. First, real-time alerts can be sent through campus information management systems or mobile apps to ensure immediate dissemination. Second, customized push strategies should be implemented based on user roles and permissions—for example, sending crisis details to counselors for students under their supervision, while providing high-risk student lists requiring professional intervention to psychological consultants. Additionally, supplementary communication methods like SMS and email can be utilized to prevent delays caused by single-channel failures. Research indicates that effective alert systems significantly enhance the timeliness and efficacy of mental health interventions, thereby safeguarding students' psychological well-being [10].

V. CHALLENGES AND COUNTERMEASURES IN THE APPLICATION OF EARLY WARNING SYSTEMS

A. DATA PRIVACY PROTECTION

In the psychological crisis early-warning system for college students that integrates big data and machine learning, data privacy protection remains a critical core issue. The data collection phase may involve students' personal behavioral records, social media content, and sensitive information from campus management systems. Acquiring such data often requires crossing traditional privacy boundaries, thereby triggering ethical controversies and legal risks. Furthermore, during data storage and usage, unauthorized access, data breaches, or misuse could lead to severe consequences like decreased student trust or increased psychological burdens[11]. To address this challenge, encryption technology has been widely adopted in data protection fields. By applying high-intensity encryption processing to raw data, it ensures security during transmission and storage. Meanwhile, anonymization serves as an effective privacy protection method, reducing the risk of tracing data back to specific individuals through identity removal or replacement with virtual identifiers. However, anonymization may also compromise data analysis accuracy to some extent, necessitating a balance between privacy protection and data usability.

B. MODEL GENERALIZATION CAPABILITY

Machine learning models often face challenges in generalization, particularly when demonstrating significant performance disparities across student demographics and diverse environments. This limitation primarily stems from training data limitations, such as the absence of behavioral characteristics for specific groups (e.g., ethnic minority students or international students) in datasets, which leads to prediction biases when assessing psychological crises in these populations.

C. SYSTEM REAL-TIME

Ensuring real-time performance of early warning systems is crucial for timely detection and intervention in psychological crises among college students. Insufficient real-time capability may lead to delayed alerts, missing optimal intervention windows, and consequently compromise system effectiveness. Key factors affecting real-time performance include data processing speed, network latency, and algorithm complexity. Specifically, large-scale data collection and cleaning processes can be time-consuming, while complex machine learning algorithms require substantial computational resources during model training and inference phases—factors that collectively slow system response. To optimize real-time performance, first implement high-performance computing architectures like distributed computing frameworks to accelerate data processing. Secondly, network latency can be mitigated by deploying edge computing technology, which offloads partial data processing tasks to near-source nodes to reduce transmission time. Additionally, simplifying algorithm complexity and optimizing model structures are vital strategies for improving real-time responsiveness. For instance, lightweight neural network designs or approximate reasoning methods can significantly reduce computational costs while maintaining prediction accuracy, thereby meeting real-time alert requirements [12].

Technical Metric for Real-time Performance: To quantitatively measure the system's efficiency, a critical technical metric is the End-to-End Alert Latency. This is defined as the total time delay, measured in milliseconds (ms) or seconds (s), between a new behavioral data point (e.g., an abnormal access log or a social media post) being logged into the collection module and the system generating or updating the corresponding student's risk score and subsequent alert. A high-performance system might aim for an average End-to-End Alert Latency of under 500 ms for behavioral data processed by the inference model, with its online performance validated through real-time traffic simulation, ensuring alerts are near-instantaneous upon data ingestion.

D. TECHNICAL ARCHITECTURE FOR REAL-TIME OPTIMIZATION

To ensure the target End-to-End Alert Latency of under 500 ms, the system employs a specialized technical architecture. The deployment is based on a hybrid Cloud-Edge computing model: the core model training and large-scale data storage (HDFS/NoSQL) are hosted in the cloud, while the inference models are deployed at the edge (on-campus servers or campus cloud infrastructure) for lower latency. The data flow is handled via streaming and caching solutions:

Streaming Processing: All incoming behavioral data (e.g., access logs, consumption data) is ingested through a message queue system (e.g., Apache Kafka) for high-throughput, concurrent processing. This enables horizontal scaling to handle real-time traffic, achieving a validated throughput of over 5,000 events per second (EPS) after being tested under simulated peak campus traffic conditions.

Caching: Frequently accessed student profiles and pre-calculated feature vectors are stored in an in-memory database (e.g., Redis) to dramatically reduce I/O latency during the inference stage.

Inference Latency Measurement: The model inference time is measured from the moment the processed feature vector enters the DCCFM-FFM model to the output of the risk score. For the deployed model, the average Inference Latency on the dedicated campus hardware is consistently measured at approximately 85 ms. The End-to-End Alert Latency is tracked by logging timestamps at the data ingestion, feature engineering, model inference, and alert output stages.

Hardware/Deployment Details: The inference models are served on dedicated, containerized campus servers running on high-speed internal networks (Edge Deployment). This minimizes network latency, which is a major bottleneck in pure cloud-based systems.

VI. EMPIRICAL RESEARCH

A. PREPARATION OF EXPERIMENTAL DATA

The data used in this experiment primarily originated from a mental health assessment system for university students, the campus academic management system, and social media platforms. The data collection process spanned three complete academic years (September 2021 to June 2024). The collection scope was limited to all full-time undergraduate students ($N \approx 15{,}000$) enrolled at a single large public comprehensive university located in the central region. The dataset covers student records over the past three years, totaling approximately 50,000 valid samples.

B. KEY DATA PROCESSING AND LABELING DETAILS

Inclusion/Exclusion Criteria: The final dataset included records for all students who provided explicit written informed consent as stipulated by the Institutional Review Board (IRB) protocol and maintained complete longitudinal records across the specified campus systems. Records with greater than 10% data gaps or those belonging to students who formally opted out of the study were excluded.

Category Labeling Process: The binary psychological crisis label (Positive/Negative) was established based on formal, documented records of intervention from the University's Psychological Counseling and Guidance Center (PCGC). A student was labeled 'Positive' only if they received an official clinical diagnosis of moderate-to-severe psychological distress (e.g., Major Depressive Disorder, severe anxiety, or suicidal ideation) that necessitated a formal, recorded intervention by the PCGC during the observation period.

Tagger Composition and Consistency: The final crisis labels were verified by a panel of three certified clinical psychologists ("taggers") from the PCGC, using established clinical protocols and intervention notes. To quantify the reliability of the labeling, particularly in ambiguous cases, the inter-rater agreement was calculated. This verification yielded a Cohen's Kappa score of 0.85, indicating substantial agreement and high confidence in the quality of the dataset labels.

The final, curated dataset used for model training contained approximately 50,000 valid samples, with a positive case ratio (students requiring official intervention) of approximately 4.5% across the three-year observation window. This percentage represents the actual prevalence of formally identified psychological crises requiring clinical intervention within the study population, ensuring the dataset reflects realistic and manageable operational conditions.

C. EXPERIMENTAL DESIGN AND IMPLEMENTATION

The experimental design employed a comparative study methodology to validate the superiority of a feature fusion-based early warning system over traditional approaches. Specifically, two comparison groups were established: the first group utilized rule-based traditional early warning methods, while the second group implemented models using single machine learning algorithms (e.g., support vector machines). The experimental process consisted of three phases: data preparation, model training, and model testing. During the data preparation phase, completed data cleaning and annotation provided foundational support for subsequent steps. In the model training phase, both traditional methods the integrated big data-machine learning model, formally named the Dual-Channel CNN-FM Feature Fusion Model (DCCFM-FFM) were trained.

D. EXPERIMENTAL RESULT ANALYSIS

Experimental results demonstrate that the integrated big data and machine learning early warning system exhibits significant advantages across all metrics. Specifically, the system achieved 92.3% accuracy, 89.7% recall rate, and 91.5% F1 value, representing approximately 12%,10%, and 15% improvements over traditional methods. Compared with models using single machine learning algorithms, the fusion model demonstrated an 8.1% higher F1 value, validating the effectiveness of multi-channel feature extraction and deep learning architectures. Further analysis revealed that the fusion model demonstrates enhanced feature capture capabilities when processing unstructured text data (e.g., social media posts), particularly excelling in identifying localized sensitive information.

To address the model generalization challenges and potential inter-group bias highlighted in Section "Model Generalization Capability", we conducted a deeper analysis of the model's performance across key student subgroups. Table 1 presents a comparative performance of the DCCFM-FFM model on the overall student population, minority students, and international students. The results show that while the model's performance on minority groups is slightly lower than the overall benchmark, it still maintains a high level of predictive capability. For instance, an accuracy of 89.5% for international students confirms the model's effectiveness across diverse cultural backgrounds, albeit with a minor performance gap. This analysis underscores the need for future work to incorporate more diverse datasets and employ transfer learning techniques specifically tailored for these groups to further close the performance gap and enhance equity.

TABLE 1. Model Performance Across Different Student Groups

Student Group	Accuracy (%)	Recall (%)	F1-Score (%)
Overall Students (Benchmark)	92.3	89.7	91.5
Minority Students	90.1	86.5	88.2
International Students	89.5	85.2	87.3

These findings not only validate the feasibility and effectiveness of integrating big data and machine learning technologies in psychological crisis early warning for college students, but also provide crucial references for future research.

VII. CONCLUSIONS AND OUTLOOK

A. RESEARCH CONCLUSIONS

The psychological crisis early warning system for college students, integrating big data and machine learning, has been comprehensively developed and validated in this study, demonstrating significant theoretical value and practical significance.By collecting, storing, and analyzing multi-channel data, the system extracts key characteristics related to psychological crises from massive information and utilizes machine learning algorithms to model and predict these features. Experimental results show that compared with traditional early warning methods, the integrated system shows remarkable improvements in core metrics such as accuracy, recall rate, and F1 value, particularly excelling in complex scenarios. This achievement not only provides scientific evidence and technical support for mental health education in universities but also lays a solid foundation for maintaining campus stability and promoting students' healthy development.

B. FUTURE OUTLOOK

While this study has achieved phased progress in developing a psychological crisis early-warning system for college students through integrating big data and machine learning, there remain numerous areas worthy of exploration as technology evolves and application scenarios expand. First, regarding data privacy protection, blockchain technology can be leveraged to ensure transparent and secure data collection and storage. The distributed ledger feature and encryption mechanisms of blockchain effectively prevent data leakage and tampering, while guaranteeing comprehensive user privacy protection. This will establish a stronger trust foundation for the widespread adoption of such early-warning systems.

Finally, in terms of application scenario expansion, the early warning system can gradually extend to broader social groups such as professionals and adolescents, providing technical support for mental health protection across more populations. Meanwhile, by integrating visualization technology and intelligent interaction interfaces, the system can better serve professional groups like university counselors and psychological consultants, helping them conduct psychological interventions more efficiently. In summary, future research should continue to focus on the application potential of emerging technologies, continuously optimize system performance, and contribute to building a more comprehensive mental health service system. This focus is equally critical for industrial environments, where leveraging these technologies will be essential for continuous system optimization and for establishing a more comprehensive well-being support framework for employees.

AUTHOR CONTRIBUTIONS

Conceptualization-Ziyan Yao and Jin Gan; methodology-Ziyan Yao and Jin Gan; formal analysis-Ziyan Yao and Jin Gan; investigation-Ziyan Yao; resources-Ziyan Yao; writing-Ziyan Yao and Jin Gan. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This work was supported by Theoretical and Practical Research Project on Ideological and Political Education for College Students in Guangxi in 2024: "Practical Research on Integrating Excellent Traditional Culture of Guangxi into Psychological Education in Higher Vocational Education in the New Era" (Project Number: 2024SZ181).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] E. Rack, “Rethinking big P: Integrating personal and academic psychology knowledge in university teaching,” Possibility Studies & Society, vol. 2, no. 4, pp. 381–394, 2024, doi: 10.1177/27538699241240608.

[2] G. Shanmugam, T. Thanarajan, S. Rajendran, and S. S. Murugaraj, “Student Psychology based optimized routing algorithm for big data clustering in IoT with MapReduce framework,” Journal of Intelligent & Fuzzy Systems, vol. 44, no. 2, pp. 2051-2063, 2023, doi: 10.3233/jifs-221391.

[3] J. M. Logg, “The psychology of Big Data: Developing a "theory of machine" to examine perceptions of algorithms,” The psychology of technology: Social science research in the age of Big Data, pp. 349-378, 2022, doi: 10.1037/0000290-011.

[4] H. Mulyawan and N. Suhaili, “Educational Psychology Big Data Approach to Students Disability,” 2021, doi: 10.3333/IJODASI.V1I1.12.

[5] M. Kosinski, “Using Big Data (Handbook of Social Psychology),” 2024, doi: 10.31219/osf.io/rp8h5.

[6] M. Choe and M. Jeon, “A PRISMA-Based Review of Empathic Intervention in Driving Contexts,” Proceedings of the Human Factors & Ergonomics Society Annual Meeting, vol. 69, no. 1, 2025, doi: 10.1177/10711813251369812.

[7] K. Szafrańska, J. Trocha, and E. Małachowska, “A Psychological Perspective on Risk and Crisis Management Skills,” Journal of Security and Sustainability Issues, 2023, doi: 10.47459/jssi.2023.13.19.

[8] C. Tamouridis, M. Moyeno, and W. A. Pasmore, “The Psychological Dynamics of Leadership amid a Crisis,” Springer Optimization and Its Applications, pp. 521-537, 2023, doi: 10.1007/978-3-031-39542-0_24.

[9] A. Minocha and N. Gupta, “PSYCHOLOGICAL CRISIS ON HIGHER EDUCATION STUDENTS DUE TO REMOTE LEARNING DURING COVID 19.2021,”, doi: 10.1596/36665.

[10] J. Zhao, H. Zhao, and A. Zhou, “Negative Parenting Styles and Psychological Crisis in Adolescents: Testing a Moderated Mediating Model of School Connectedness and Self-Esteem,” Behavioral Sciences, vol. 13, no. 11, pp. 14, 2023, doi: 10.3390/bs13110929.

[11] H. Y. Niu and J. Yu, “Construction of a Dynamic Early-warning System for College Students’ Psychological Crisis,” Psychology of China, 2022, doi: 10.35534/pc.0411161.

[12] K. Shen and X. Yuan, “Construction of machine learning based psychological crisis warning model for college students integrating biomechanics indicators,” Molecular & Cellular Biomechanics, vol. 22, no. 2, pp. 676, 2025, doi: 10.62617/mcb676.