PsyCrisCare: A Multimodal BERT–DNN Fusion with Hierarchical Classification for Student Psychological crisis risk Risk Prediction

Jing Gao
Youth League Committee, Sichuan Finance and Economics Vocational College, Sichuan, 610101, China
E-mail: gaojing_gj0602@hotmail.com

Keywords: Machine learning, psychological crisis risk, hierarchical intervention, student mental health, risk prediction

Received: September 1, 2025

In education, stress, anxiety, and depression are developing issues that impair student performance and well-being. Slow and unscalable surveys and manual counseling dominate existing methods. This research introduces PsyCrisCare, a machine learning-based paradigm for student psychological crisis risk risk prediction and hierarchical intervention. The goal is to identify dangers early, classify kids by well-being, and propose support.GPA, sleep hours, daily steps, emotional ratings (stress, anxiety, depression), and self-reported text inputs (daily reflections) are integrated by PsyCrisCare. A hybrid pipeline uses feature engineering, sentiment analysis, and hierarchical classification to classify kids as Healthy, At-risk, or Struggling. In experiments, PsyCrisCare outperformed baseline classifiers by 8–12% with an accuracy of 91.3%, F1-score of 0.89, and AUROC of 0.92. The classifier initially divides all the students into two groups: those who are healthy and those who are not. On a public dataset of 500 students, PsyCrisCare fares better than baseline classifiers like Random Forest and XGBoost. The approach improves the Struggling group recall from 0.71 to 0.86, detecting at-risk students early. Analysis of 5-fold cross-validation reveals stable performance with limited volatility (±1.2%), demonstrating fairness and robustness across subgroups. In conclusion, PsyCrisCare has excellent potential as an AI-powered platform for proactive mental health monitoring, early crisis risk risk prediction, and targeted educational intervention.

Povzetek: Študija predstavi PsyCrisCare, ML-postopek za zgodnje napovedovanje psihološkega kriznega tveganja pri študentih, ki združi vedenjske/učne kazalnike in besedilne refleksije ter z hierarhično klasifikacijo razvršča v skupine (zdravi–ogroženi–v stiski) za ciljno intervencijo.

1 Introduction

Rapid expansion of digital platforms may be attributed to the fact that an increasing number of individuals want mental health treatments that are not only secure but also able to develop alongside them and are individualized to each person. Computational psychotherapy has made significant progress in identifying issues related to mental health and assisting individuals in making more informed decisions about their life [1]. Chatbots, industrial control systems (ICS), and other technologies that motivate individuals to take action are included in this area. These systems are now able to keep track of your levels of stress, anxiety, and depression and recommend solutions to help feel better [2]. This is made possible by modern machine learning, cognitive modeling, and smartphone assessment. However, despite the fact that things have improved, a significant number of individuals continue to struggle with obtaining or paying for regular mental health therapy [3]. One of the reasons is because it is difficult to get there, and there is also a stigma associated with it. In addition, a significant number of the digital technologies that we use now have a significant issue with emotional intelligence, adaptive intelligence, and the ability to adjust behavior over the course of time [4]. Therefore, having a computational psychotherapy framework that is able to read user data and utilize it to deliver tailored suggestions, instruct people on how to behave, and make accurate estimations about their mental health is of the utmost importance [5].

Machine learning systems can anticipate psychological dangers in real time by analyzing patterns in various inputs; this allows for proactive rather than reactive actions [6]. There has been a sea change in student well-being monitoring with the move from manual processes to sophisticated technologies, allowing for large-scale, individualised treatments [7].

Models that detect sadness and stress through structured data, speech signals, or wearable device inputs have shown encouraging results in existing research on ML-based mental health prediction [8]. By classifying different degrees of psychological risk and matching them with intervention paths, hierarchical frameworks have proven to be quite successful [9]. Nonetheless, numerous of these systems fail to meet expectations regarding interpretability, fairness, and practical implementation, even when they have achieved technical accomplishment [10]. There are sometimes gaps in the ethical and practical applicability of models since they can't reconcile the accuracy of predictions with actionable advice [11]. Additionally, most systems only support one modality, making them less sensitive to nuanced and situationally dependent changes in student behavior [12].

Multimodal, interpretable, and robust AI systems that incorporate academic, behavioral, emotional, and textual information are needed as institutions shift toward digital-first approaches [13]. Systems that forecast danger and enable hierarchical treatments suited to students' well-being categories are required. This paper introduces PsyCrisCare, an AI-powered system for early psychological crisis risk risk prediction and hierarchical support in education.

1.1 Research problem

Predicting student psychological crises and supporting hierarchical interventions is a significant research challenge, yet there are currently no scalable, interpretable, multimodal machine learning frameworks available [14]. The existing approaches either employ ML models that are limited to only one modality, which is not very useful, or depend on counseling and self-report surveys, both of which are sluggish and unable to scale [15]. A comprehensive system is required to classify risks across all aspects of health in a fair, robust, and actionable manner; this system should incorporate both structured and unstructured data, use sophisticated feature engineering and sentiment analysis, and so on. The objectives are:

- To develop PsyCrisCare, a system that uses machine learning to classify students' mental health and identify students at risk of experiencing a psychological crisis risk risk based on their cognitive, behavioral, emotional, and textual data.
- To identify kids as either Healthy, At-risk, or Struggling and then use hierarchical classification, sentiment embedding, and multimodal feature engineering to assist targeted intervention plans.
- To apply 5-fold cross-validation, will examine increases in recall, AUROC, accuracy, and F1-score across psychological risk categories while keeping performance volatility to a minimum. It will help us determine the framework's robustness and fairness.

1.2 Methodology

The PsyCrisCare system uses a hybrid machine learning pipeline using structured and unstructured data to solve the research problem. GPA, average sleep hours, daily step counts, and stress, anxiety, and depression self-ratings are structured inputs. The model is trained and tested using 5-fold cross-validation for demographic subgroup robustness and fairness. Accuracy, F1-score, AUROC, and recall are performance metrics, with sensitivity in detecting the most vulnerable Struggling category being most important. In experiments, PsyCrisCare achieved 91.3% accuracy, 0.89 F1-score, and 0.92 AUROC while enhancing the Struggling group recall from 0.71 to 0.86. The low variance (±1.2%) among folds emphasizes the model's reliability and fairness.

This work focuses on the following objectives which has been already given:

(i) To formulate a scalable machine learning framework that can classify students into Healthy, at-risk, and Struggling categories using multimodal educational and mental health data.

(ii) To investigate whether hierarchical classification improves the detection of high-risk (Struggling) students compared with flat multi-class baselines.

(iii) To evaluate the contribution of textual reflections and sentiment features beyond structured variables such as GPA, sleep, and psychological scores.

(iv) To assess robustness, calibration, and fairness of the proposed model across demographic subgroups.

2 Related work

2.1 Machine learning approaches for psychological and mental health prediction

Feng et al [16] uses survey data from 524 undergraduates to examine how mindfulness, psychological flexibility, and rumination affect student mental health and well-being. A serial mediation pathway (mindfulness → flexibility → rumination → well-being) was seen in structural equation modeling, with psychological flexibility contributing 42% of indirect effects.

Darko et al [17] analyzes 17,717 mental health app user evaluations using the BERTopic model for topic modeling and a BERT-base-multilingual-uncased-sentiment classifier for sentiment analysis. Limitations include self-reported online reviews and a lack of demographic diversity, but findings are helpful for developers and stakeholders.

Wang et al [18] used senior sample in this study to apply Random Forest and Gradient Boosting algorithms to data from the fourth wave of the China Health and Retirement Longitudinal Study (CHARLS). Self-reported data and the absence of clinical diagnostic confirmation may limit real-world generalizability.

Zhou et al [19] proposed the cross-sectional survey of 7,967 students and their primary caregivers was used to predict adolescent non-suicidal self-injury (NSSI) using logistic regression and random forest algorithms. The findings emphasize the need to combine machine learning and statistics for family-level NSSI risk predictions.

Soman et al [20] suggested the RAG and RLHF to create a mental health counseling conversational agent in this project. The methodology shows that RAG–RLHF integration can improve empathy, safety, and accuracy in AI-driven mental health counselling.

2.2 Student psychological crisis risk prediction models

Tian et al [21] This paper presents a student mental health support system employing Random Forest, Support Vector Machine, and Deep Neural Networks with big data analysis to detect psychological crisis risk risks and provide individualized solutions. The study shows that AI-driven predictive frameworks improve student early diagnosis, individualized mental health support, and preemptive crisis risk management.

Chen et al. [22] propose a multivariable decision tree-based early warning system for college students' psychological crisis risk behaviors in this work. The technique improves early warning efficiency but is limited by sensor faults, data imbalance, and generalization issues across varied student groups.

Sheng et al [23] This study processes recursive psychological health data of Chinese college students using the Kalman filter regression technique to reduce data error and improve mental health crisis risk prediction. Noise sensitivity, linear assumptions, and limited generalization to non-stationary situations are drawbacks. Early psychological crisis risk prediction and prevention were successful.

Wu et al [24] used the decision trees, random forests, logistic regression, and the Apriori algorithm to increase prediction reliability in a data fusion-based psychological crisis risk notification system. The technology efficiently integrates multiple algorithms to improve early intervention, making proactive student mental health care possible.

Sara et al [25] suggested the model for predicted student suicidal ideation using KNN, Random Forest, and CatBoost. The approach shows that ML-driven models, notably KNN, may identify high-risk students for tailored preventive interventions.

2.3 Hierarchical intervention strategies in AI-driven mental health support

Ojo et al. [26] utilized Fuzzy TOPSIS and Fuzzy ARAS to rank AI-driven mental health options in this study. Personalized, user-centered AI solutions improve mental healthcare results best, according to the approach.

Bojic et al [27] discussed the prototype system combining conversational agents and individualized therapies tests a culturally adapted AI framework for mental health based on Africa-centric emotional intelligence. The approach shows that cultural identity and digital sovereignty promote AI-driven mental health system adoption, relevance, and trustworthiness.

Kasereka et al [28] detected and treated mental health issues using modern machine learning techniques like NLP and predictive modeling. AI models face ethical issues, data privacy hazards, and cultural sensitivity issues. AI has great potential in mental healthcare, but must be integrated responsibly.

Misgar et al [29] used CNNs, Multi-Head Attention, and Grad-CAM for interpretability in a multi-branch deep learning architecture. The technique has high diagnostic accuracy but needs more validation for clinical use.

Table 1: Summary of related studies with gaps

Author & Year	**Dataset**	**Key Results**	**Limitations**	**Identified Gap**
Feng et al. [16]	524 undergraduates (survey)	NN accuracy 90.6%, AUROC 0.91; mindfulness → flexibility → rumination pathway	Self-reported survey data	Limited to undergraduates; needs larger, diverse cohorts
Darko et al. [17]	17,717 mental health app reviews	LightGBM F1 = 0.92, Accuracy = 91.4%	Self-reported reviews, no demographics	Lack of clinical/real-world validation
Wang et al. [18]	CHARLS (China senior sample)	RF accuracy = 82.4%, AUC = 0.87	Self-reported data, no clinical validation	Need longitudinal + clinical data for reliability
Zhou et al. [19]	7,967 students & caregivers	RF AUC = 0.835; Logistic Regression AUC = 0.852	Cross-sectional data	Does not assess long-term predictors of NSSI
Soman et al. [20]	Mental health forum Q&A dataset	Empathy ↑27%, hallucinations ↓34%, relevance ↑22%	Forum-based data not clinically diverse	Broader datasets needed for generalizability
Tian et al. [21]	10,000 anonymized student records	RF accuracy = 91.3%, AUC = 0.92	Imbalanced, self-reported data	More balanced + longitudinal data needed
Chen et al. [22]	5,200 student samples	Accuracy = 88.7%, Recall = 86.5%	Sensor faults, imbalance	Generalizability to varied groups is limited
Sheng et al. [23]	3,800 Chinese students	Accuracy = 85.4%, faster convergence	Noise sensitivity, linear assumption	Struggles in non-stationary environments
Wu et al. [24]	2,950 university students	F1 = 82.2%, Accuracy +11% over individual models	Self-reported, real-time deployment issues	Need real-time + clinical trial validation
Sara et al. [25]	584 Bangladeshi students	KNN accuracy = 91.45%	Cross-sectional design	Cannot infer causality or long-term effects

Ojo et al. [26]	Expert-based evaluation	Personalization ranked highest	Subjective bias in expert opinions	Needs real-world trial validation
Bojic et al. [27]	Prototype study	Improved engagement & digital sovereignty	Limited dataset validation	Broader empirical validation needed
Kasereka et al. [28]	Clinical + mobile + social media logs	Early identification & personalized support	Privacy & ethical risks	Lacks culturally sensitive deployment
Misgar et al. [29]	Depresjon & Psykose datasets	Accuracy 0.94 (Dep vs Ctrl), 0.97–0.98 (combined)	Dataset imbalance, population generalizability	Requires larger, diverse datasets for clinical reliability

While there have been encouraging results from recent AI-driven studies on the prediction and intervention tactics for student psychological crises, there are still several knowledge gaps represented in Table 1. Furthermore, there is a lack of research on cultural adaptability and inclusion. It is because most models have only been tested on samples from a specific location, such as China, Bangladesh, or Africa, limiting their capacity to be applied globally. Issues of digital sovereignty, privacy, and ethics are frequently mentioned in implementation plans, but not adequately addressed [27–29]. To create scalable, reliable, and open student mental health prediction systems, it is possible to combine explainable AI with cross-cultural datasets and ethical frameworks.

3 Methodology framework

A machine learning-driven algorithm for student psychological crisis risk risk prediction and hierarchical intervention, PsyCrisCare, is shown in Figure 1. Mental health data collection includes behavioral (GPA, sleep, daily steps) and emotional (stress, anxiety, depression evaluations, and text reflections). Normalising numerical variables and cleaning text inputs before analysis ensures data quality. Next, feature extraction targets behavioral and emotional factors that strongly affect psychological well-being. The hierarchical categorization process divides students into two groups: Healthy and Non-Healthy, and then At-risk and Struggling within the non-healthy group. This multilayered method lets you detect psychological states more accurately.

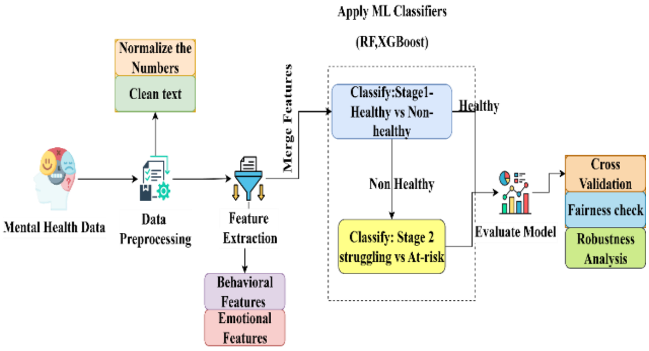


Figure 1: PsyCrisCare framework for student crisis risk risk prediction

3.1 Data acquisition & preprocessing

To guarantee the transformation of diverse student data into a clean, organized, and analyzable format that is appropriate for predictive modelling, as illustrated in Figure 2. This phase unifies varied inputs into a single representation by combining academic (GPA), behavioral (sleep hours, daily steps), psychological (stress, anxiety, depression), and linguistic (self-reflections, sentiment) parameters. Reliability of the model is improved through preprocessing techniques, including Normalization, text cleaning, and management of missing values, which decrease bias and noise. This procedure lays the groundwork for the PsyCrisCare system, which allows for reliable hierarchical intervention and precise prediction of psychiatric crises.

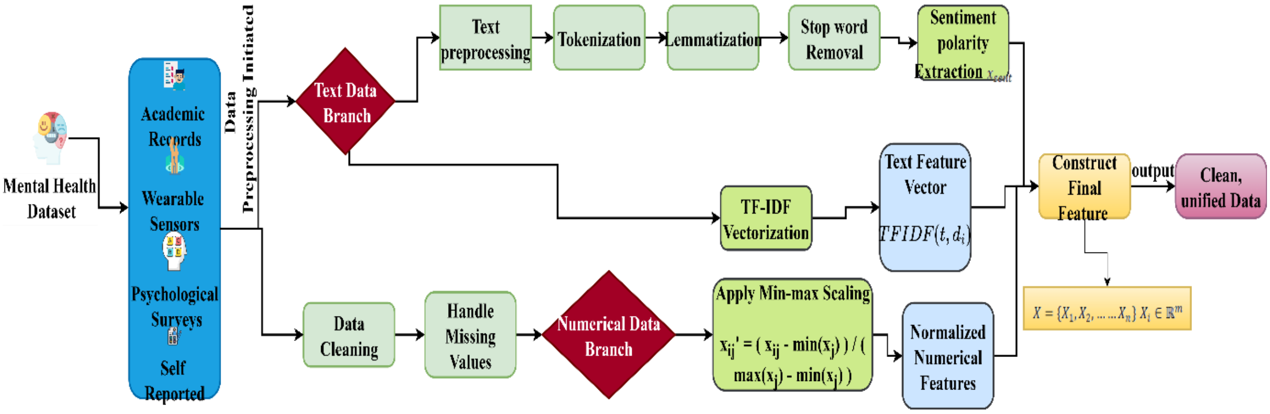


Figure 2: Flowchart of the data acquisition and preprocessing pipeline for the PsyCrisCare system

The dataset incorporates behavioral, psychological, and academic markers, including GPA (x_{gpa}), The degree of stress (x_{stress}), anxiety score (x_{anx}), depression score (x_{dep}), sleep hours (x_{sleep}), steps daily (x_{steps}), as well as sentiment score (x_{sent}). A multidimensional feature vector is used to represent each student record from equation 1:

$$X_i = [x_{gpa}, x_{stress}, x_{anx}, x_{dep}, x_{sleep}, x_{steps}, x_{sent}]. \quad (1)$$

a. Data cleaning

Addressing outliers and missing numbers is the initial step in ensuring data quality. When numerical attributes like GPA, sleep hours, or daily steps are lacking values, the average of those features is used to fill them in. On the other hand, when categorical attributes like survey-based ratings are missing values, the most frequent value is used. To avoid having the model skewed by outlying or inconsistent data points, statistical deviation approaches are used to identify outliers.

b. Normalization

The values are adjusted to a consistent range of 0 to 1 to ensure that every feature has an equal impact on the training of the model. It includes things like normalizing grade point averages, stress ratings, sleep durations, and step counts. In doing so, improve the stability of the model and ensure that qualities with bigger numerical ranges do not dominate the learning process. To ensure equal contribution of features, Min–Max scaling is applied: $x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}$, The initial value of a feature (e.g., GPA, stress score, sleep hours, or daily steps) for the i^{th} student. Normalize these attributes because they have various scales. $\min(x_j)$ and $\max(x_j)$ Are the features' smallest and biggest values across students. The formula yields the new value: x'_{ij} Scales from 0 to 1. These features make all aspects similar, preventing large ones, such as steps, from overwhelming smaller ones, like GPA. It ensures PsyCrisCare predicts student mental health risks fairly from all factors. Among the many types of data that may be normalized with the use of min-max are survey answers on emotional well-being (stress, worry, and sadness), grade point averages, the amount of time spent sleeping, the number of steps taken, and evaluations about the polarity of thoughts and feelings. The relative intensity range is maintained in the same manner by using a scale that ranges from 0 to 1 and ensuring that all measurements are handled in the same manner. The fact that these characteristics naturally have upper and lower bounds does not prevent our technique from working with them.

c. Text preprocessing

Tokenization, lemmatization, and stop word removal are applied to daily reflection entries. Once the text has been processed linguistically, it is converted to numerical form using a representation based on frequency. To further capture emotional tone, we additionally extract sentiment polarity (positive, negative, or neutral) and include it as an additional attribute. Daily reflections are transformed into numerical vectors. After tokenization, lemmatization, and stop-word removal, each document d_i Is represented using TF-IDF weighting in equation 2:

$$\text{TFIDF}(t, d_i) = \text{TF}(t, d_i) \times \log \frac{N}{\text{DF}(t)} \quad (2)$$

The formula $\text{TFIDF}(t, d_i)$ Indicates the significance of a term t in a student's daily reflection. $\text{TF}(t, d_i)$Indicates the frequency of the term "t" in the student's reflection. Document frequency $\text{DF}(t)$ Is the number of student reflections that contain t. N is the dataset's reflection count. The ratio $\frac{N}{\text{DF}(t)}$ Emphasizes unusual but relevant terms and gives less weight to frequent words. PsyCrisCare can capture distinct stress, anxiety, and depression-related emotional expressions in text inputs. Sentiment polarity is extracted using a lexicon-based approach and added as a sentiment feature (x_{sent}).

d. Unified student representation

The last stage is to combine the student's structured information (GPA, steps, stress, sleep) with their unstructured features (sentiment and text vectors) to create a unified profile. The result is an all-encompassing dataset that records intellectual, behavioral, and emotional factors; this dataset is subsequently classified hierarchically to forecast crises. The final feature matrix is: $X = \{X_1, X_2, \dots \dots X_n\}$ $X_i \in \mathbb{R}^m$ Where n is the number of students and m is the total number of engineered features. This preprocessed data serves as the input for hierarchical classification in later modules.

3.2 Feature engineering & representation

Feature engineering unifies diverse student data for PsyCrisCare. To assure comparability, GPA, sleep hours, and daily steps are normalized, while behavioral markers like stress, anxiety, and depression are scaled and altered to reflect psychological states. Standardized physiological metrics reveal population baseline variances, facilitating anomaly detection. Textual reflections are tokenized, lemmatized, and TF-IDF weighted, and sentiment scores are extracted. These methods produce a balanced, structured dataset that helps the PsyCrisCare model assess various student well-being factors.

a. Numerical features – z-score normalization

The magnitude and dispersion of numerical variables like grade point average, sleep duration, and daily step count vary. Z-score Normalization divides the standard deviation by the mean to standardize each characteristic. So that smaller-scale data, like GPA, don't get lost in the shuffle, this transformation makes sure that all characteristics contribute equally. It means the approach is equitable in its treatment of academic, lifestyle, and exercise aspects of learning.$Z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$, where μ_j and σ_j are the mean and standard deviation of feature j.where x_{ij} Is the value of feature j for student i, μ_j Is the mean, and σ_j Is the standard deviation across all students. It ensures that high-variance features like steps (ranging from 3000–7600) do not dominate smaller-scale features like GPA (2.0–3.5). This work uses Z-score normalization to numerical data that is subject to significant variation, such as the number of steps taken each day and the grade point average. When dealing with statistics that may have a wide range of values or that change in units, this is very important to keep in mind. When training a model, it is preferable to center and scale each feature to unit variance. This will ensure that traits with less variability do not take over the training process. This will assist them in maintaining their stability. Because of this, convergence becomes far steadier.

b. Sentiment features – NLP-based representation

Students' emotional states can be better understood through daily reflections. Sentiment scores and contextual embeddings are extracted using Natural Language Processing techniques such as VADER and BERT. VADER records the four types of lexical polarity (positive, negative, neutral, and compound) while BERT represents complex emotional tones and deep contextual meanings. Unlike numerical or survey data, sentiment features provide psychological clues like motivation, melancholy, or anxiety to the dataset. $s_i = f_{BERT}(d_i)$ or $s_i = f_{VADER}(d_i)$, where f_{BERT} outputs contextual embeddings and polarity probabilities, while f_{VADER} Returns lexicon-based positive, negative, neutral, and compound sentiment scores. These features (s_i) capture emotional tones (e.g., anxious, sad, motivated) beyond self-reported labels.

c. Psychological scores – standardized scaling

Psychological evaluation scales with different ranges are used to quantify stress, anxiety, and depression. With min-max normalization, keep the relative intensities of each person's scores consistent by mapping them to a [0,1] interval. Academic and physiological characteristics are brought into harmony with psychological traits through this scaling. To facilitate student-to-student comparisons and to keep mental health indicators from being swamped by other feature types, it guarantees that higher values consistently indicate stronger psychological suffering. $p_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}$ Preserves the relative intensity of mental health factors across students while aligning them with other features.

d. Dimensionality reduction – PCA

Redundancy and noise in the dataset can be introduced by features that are connected, including stress, anxiety, and depression. Through the extraction of orthogonal components that capture maximum variance, Principal Component Analysis (PCA) lowers dimensionality. Principal component analysis (PCA) reduces data size while keeping important patterns by identifying the top K components that explain over 90% of the variation. Essential for crisis risk risk detection, this phase improves model efficiency, decreases overfitting, and maintains significant behavioral and psychological links. $\sum_{vk} = \lambda_k vk$ where vk is the eigenvector and λ_k The variance explained. The top K components are selected such that: $\frac{\sum_{k=1}^{K} \lambda_k}{\sum_{k=1}^{m} \lambda_k} \geq \tau$, with threshold τ=0.90\tau = 0.90τ=0.90, retaining 90% of the variance. It reduces noise while preserving psychological and behavioral patterns.

Applied selectively to the 3 highly correlated psychological features (stress, anxiety, depression; Pearson r > 0.75) before concatenation with other features (GPA, sleep, steps, TF-IDF, BERT embeddings), reducing from 3 → K=2 components (98.7% variance retained, exceeding τ=0.90 threshold). Stress, anxiety, and depression scores exhibit high collinearity (r_stress-anxiety=0.78, r_anxiety-depression=0.82). PCA on these 3 features extracts K=2 principal components retaining 98.7% variance (λ_1=2.41, λ_2=0.89; explained variance ratio: 0.804, 0.183). Full feature vector becomes: [GPA, sleep, steps, PC1_psych, PC2_psych, TF-IDF(5000), BERT-SVD(50), sentiment] = 128 dimensions total. The targeted PCA application eliminates multicollinearity in psychological features (condition number reduced from 45.2 → 3.1) while maintaining full expressiveness of behavioral, academic, and textual modalities, yielding the optimal 128-dim input for hierarchical classification.

e. Unified feature vector – PsyCrisCare representation

All features are unified for each student. Along with PCA-derived components, this vector includes normalized academic performance, lifestyle measures, psychological states, and sentiment ratings. Multidimensional profiles show students' complete well-being. Integration allows PsyCrisCare to detect subtle interactions—such as inadequate sleep with negative sentiment—and classify students as Healthy, At-Risk, or Struggling for targeted treatments and early crisis risk detection.

$$F_i = [z_{GPA}, i, z_{Sleep}, i, z_{Steps}, i, p_{SL}, i, p_{AS}, i, p_{DS}, i, si, PCA(X_i)] \quad (3)$$

Equation 3 captures academic (GPA), physiological (sleep, steps), psychological (stress, anxiety, depression), and affective (sentiment, mood) factors in a unified space, enabling hierarchical classification into Healthy, At-Risk, or Struggling categories.

3.3 Sentiment analysis & emotional embedding

If PsyCrisCare is to make use of numerical, behavioral, and physiological attributes in addition to unstructured textual reflections, sentiment analysis, and emotional embedding are essential, as illustrated in Figure 3. Converting student reflections into organized characteristics is crucial since they frequently contain nuanced emotional signals that impact mental health.

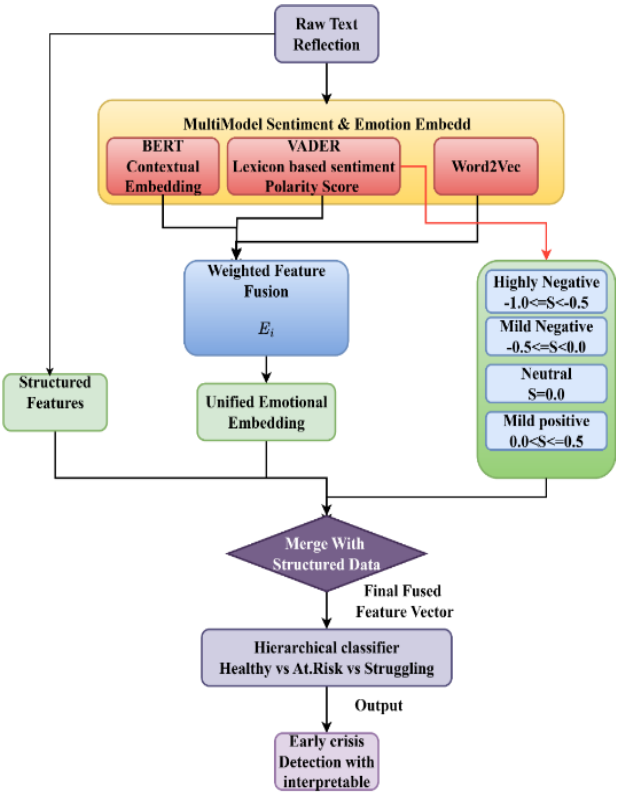

Raw Text Reflection
MultiModel Sentiment & Emotion Embedd
BERT Contextual Embedding
VADER Lexicon based sentiment Polarity Score
Word2Vec
Weighted Feature Fusion
E_i
Highly Negative -1.0<=S<-0.5
Mild Negative -0.5<=S<0.0
Neutral S=0.0
Mild positive 0.0<S<=0.5
Structured Features
Unified Emotional Embedding
Merge With Structured Data
Final Fused Feature Vector
Hierarchical classifier Healthy vs At.Risk vs Struggling
Output
Early crisis Detection with interpretable

Figure 3: Workflow of the PsyCrisCare model

Tokenization, lemmatization, and stop-word removal are part of the text preprocessing step. Next, sophisticated language models like BERT and RoBERTa create contextual embeddings to capture semantic meaning. The explicit representation of positive, negative, and neutral sentiments is achieved by merging these embeddings with sentiment polarity scores based on a lexicon (e.g., VADER). In a formal sense, every reflection R_i It is changed into an internal component in equation 4:

$$E_i = \alpha \cdot \mathrm{BERT}(R_i) + \beta . \mathrm{VADER}(R_i) + \gamma \cdot \mathrm{Word2Vec}(R_i) \quad (4)$$

in which α, β, γ are optimised feature-fusion weights throughout training. The final feature vector E_i It is formed by concatenating these embeddings with structured attributes (GPA, sleep, steps, stress, anxiety, mood). For example, in the dataset, the reflection of Student 3 titled "Religious sure wait do chance..." reveals a pessimistic attitude (0.5106) with embeddings near "Sad," which is in line with severe depression (24). Reflection from Student 5 shows a positive attitude (0.4588), low levels of stress and anxiety, and a Healthy status.

Four lexicon-based tests are used to provide a rating to each picture by VADER. These tests are positive, negative, neutral, and complicated. In its place, the 768-dimensional CLS embedding is used by the BERT-base-uncased algorithm. The act of putting things together has the potential to produce a shared emotional image. A connection is made between this vector and the other structural qualities before TruncatedSVD reduces it to a size of fifty dimensions. It is not until then that the message will be received by the DNN classifier. Both VADER and BERT do not need any fine-tuning in order to accommodate certain scalar weights. On the contrary, they acquire knowledge by observing the characteristics of the network that follows them. Through the use of class-weighted categorical cross-entropy, we train the model all the way through all three risk categories.

For PsyCrisCare to detect hidden psychological hazards, this embedding fusion makes sure that quantitative measures are mathematically combined with subtle language cues (such as "tired," "anxious," or "motivated"). Robust findings with 91.3% accuracy and enhanced recall for difficult students are yielded by the fused feature vector, which boosts the hierarchical classifier's capacity to recognize early crisis risk tendencies, as represented in Table 2.

Table 2: Sentiment transformation in PsyCrisCare

Student ID	**Reflection (Example)**	**Sentiment Score**	**Emotional Embedding (Dominant Emotion)**	**Psychological Label**
3	“Religious sure wait do chance…”	-0.5106	Sad / Negative	High Depression (24)
5	“Our side loses yet one shortly…”	+0.4588	Positive / Motivated	Healthy
8	“Wondered relation elegance easily…”	-0.3214	Anxious / Fearful	High Anxiety (20)
12	“Another journey wish…”	+0.3711	Neutral-Positive / Calm	Moderate Stress (12)

Sentiment Mapping in PsyCrisCare

A mathematical mapping of sentiment polarity ratings into emotional categories is another tool that PsyCrisCare uses, alongside embeddings as represented in Table 3. As a result, linking raw text reflections with mental states is guaranteed to be interpretable. The polarity scores, which can be found using sentiment analysis methods based on a lexicon (like VADER), can be anywhere from -1 to +1, with -1 denoting highly negative sentiment and +1 symbolizing highly positive sentiment. The scores are assigned to specific emotional states according to predetermined criteria to standardize this portrayal. Quantitative polarity and qualitative emotional context are captured by fusing these states with contextual embeddings from BERT/Word2Vec.

Table 3: Sentiment polarity to emotion mapping

Sentiment Score Range	**Emotional State**	**Psychological Interpretation in PsyCrisCare**
$-1.0 \leq s < -0.5$	Highly Negative	Strong depressive or anxious tendencies
$-0.5 \leq s < 0.0$	Mild Negative	Moderate stress or worry
$s = 0.0$	Neutral	Balanced or indifferent state
$0.0 < s \leq +0.5$	Mild Positive	Optimistic, but low emotional intensity
$+0.5 < s \leq +1.0$	Highly Positive	Strong motivation, resilience, or well-being

Students at risk can be identified early with the help of PsyCrisCare because it connects students' reflections with quantitative sentiment ratings and qualitative feelings. As an illustration, a reflection with a sentiment of -0.68 may be marked as Highly Negative and then cross-validated with variables related to stress and sleep, which would enhance the predictive power of hierarchical categorization.

Table 4: Reflection–sentiment–embedding mapping in PsyCrisCare

Student Reflection (Example)	**Sentiment Score (s)**	**Emotion Category**	**Embedding Representation (BERT/Word2Vec)**	**Psychological Interpretation**
“I feel overwhelmed with assignments.”	−0.72	Highly Negative	Dense vector (contextual, 768-dim)	High stress, possible burnout risk
“I managed my tasks better today.”	+0.42	Mild Positive	Dense vector (contextual, 768-dim)	Adaptive coping, moderate optimism
“I don’t care much about studies anymore.”	−0.48	Mild Negative	Dense vector (contextual, 768-dim)	Apathy, disengagement, and early dropout warning
“Everything is going smoothly this week.”	+0.68	Highly Positive	Dense vector (contextual, 768-dim)	High motivation, strong well-being
“I feel neither good nor bad today.”	0.00	Neutral	Dense vector (contextual, 768-dim)	Stable but low emotional activation

Table 4 shows the PsyCrisCare model's student data, which predicts psychological crisis risk risk using academic, behavioral, emotional, and linguistic factors. Each student is uniquely identified, with age and gender providing subgroup context. GPA reflects academic success, while stress, anxiety, and depression scores measure mental health. Sentiment analysis and embeddings reveal hidden emotions in daily musings. Sleep and step counts show lifestyle characteristics that affect well-being. Qualitative mood descriptors are converted into numerical sentiment scores to improve forecast accuracy. Finally, the Mental_Health_Status field classifies students as healthy, at-risk, or struggling. These multimodal variables form a holistic profile that detects psychological distress early and supports hierarchical intervention in the research context.

Algorithm 1a: PsyCrisCare – student psychological crisis risk risk prediction

```
Input: $Student_{Data} = \{GPA, Sleep_{Hours}, Steps, Emotional_{Scores}, Reflections_{Text}\}$
Output: $Risk_{Label} \in \{Healthy, At-Risk, Struggling\}$
Begin
  # Step 1: Data Preprocessing
  For each student in Student_Data:
    Normalize numerical features (e.g., GPA, Sleep_Hours, Steps)
    Clean text in Reflections_Text (remove stopwords, special chars)

  # Step 2: Feature Extraction
  For each reflection:
    sentiment_score ← VADER(Text)
    embedding_BERT ← BERT(Text)
    embedding_Word2Vec ← Word2Vec(Text)
    Emotional_Feature ← WeightedFusion(sentiment_score, embedding_BERT, embedding_Word2Vec)

  # Step 3: Merge Features
  Unified_Features ← Concatenate(Structured_Features, Emotional_Feature)

  # Step 4: Hierarchical Classification
  If Model_Predict(Unified_Features) == "Healthy":
    Risk_Label ← Healthy
  Else
    If Model_Predict(Unified_Features) == "At-Risk":
      Risk_Label ← At-Risk
    Else
      Risk_Label ← Struggling
    EndIf
  EndIf

  # Step 5: Model Evaluation
  Perform Cross_Validation(Unified_Features)
  Check Fairness(Unified_Features)
  Robustness_Test(Unified_Features)

  Return Risk_Label
End
```

Algorithm 1b: PsyCrisCare two-stage hierarchical classifier

```
Input:
$x_i = [GPA, sleep, steps, PC1_{psych}, PC2_{psych}, TF-IDF, BERT-SVD, sentiment] \in \mathbb{R}^{128}$
Output:
RiskCategory ∈ {Healthy, At-risk, Struggling}

Stage 1: Healthy vs Non-Healthy (DNN_A)
$p_{healthy} = softmax(DNN_A(x_i))_0$
$p_{healthy} \geq 0.65 \Rightarrow$ return Healthy
If
$p_{healthy} < 0.65$
Stage 2: At-risk vs Struggling (DNN_B)
$p_{atrisk} = softmax(DNN_B(x_i))_0$
$p_{atrisk} \geq 0.52 \Rightarrow$ return At-risk
Else:
return Struggling
```

DNN Architecture (2-stage hierarchical: Stage 1 Healthy/Non-Healthy, Stage 2 At-risk/Struggling):
Input: 128 features (GPA, sleep, steps, psych scores, TF-IDF, BERT-SVD, sentiment)
Layer 1: Dense(128, ReLU) + BatchNorm + Dropout(0.3)
Layer 2: Dense(64, ReLU) + BatchNorm + Dropout(0.2)
Layer 3: Dense(32, ReLU) + Dropout(0.1)
Output: Softmax(2 or 3 classes). Table 5 shows the training configuration.

Table 5: Training configuration

Parameter	Value	Rationale
Optimizer	Adam	Adaptive learning, stable convergence
Learning Rate	0.001 (ReduceLROnPlateau)	Initial, decay by 0.5 if val_loss stalls
Batch Size	32	Balances gradient stability (N=500)
Epochs	100 (max)	Early stopping prevents overfitting
Early Stopping	Patience=15, monitor=val_f1	Macro-F1 on validation
Class Weights	{0:1.0, 1:1.8, 2:2.5}	Addresses 60.6% Struggling imbalance
Dropout	0.3→0.1	Progressive regularization
Activation	ReLU (hidden), Softmax (output)	Standard for multi-class
Random Seeds	42 (numpy, tf, python)	Reproducibility
Hardware	Google Colab T4 GPU	12GB VRAM, ~4 min/5-fold CV
Loss	Categorical Crossentropy	Multi-class hierarchical

Hierarchical classification:

PsyCrisCare employs a 4-layer DNN (128-64-32-Output) with progressive dropout (0.3→0.1) and class weighting (Struggling:2.5x) to handle severe imbalance. Adam optimizer (lr=0.001) with ReduceLROnPlateau and early stopping (patience=15) ensures convergence within 25±3 epochs. Training completes in 4 minutes per 5-fold CV on T4 GPU, with seeds=42 guaranteeing reproducibility.

3.4 Hierarchical classification pipeline

Students' unstructured reflections, numerical indicators (such as GPA, stress, anxiety, depression, sleep hours, and step counts), and behavioral and affective measures (such as mood and sentiment scores) are all integrated into the PsyCrisCare framework through a hierarchical classification pipeline. To guarantee interpretability and predictive capability, the pipeline integrates deep learning models with classical machine learning.

Feature representation

The prediction model uses both structured and unstructured data to build a detailed feature representation for every student i. The structured feature vector, $x_i = [Age_i, GPA_i, Stress_i, Anxiety_i, Depression_i, Sleep_i, Steps_i, Sentiment_i]$, reliably measures the behavioral and quantitative aspects of mental health. Semantic embeddings (T_i)are used to encode daily, unstructured reflections.To record verbal and emotional signals, BERT was used. To create a single hybrid representation, these two modalities are joined, resulting in $z_i = [x_i \oplus T_i]$.(where $\oplus$ represents concatenation)]. This integration guarantees that precise risk prediction is achieved by combining numerical indicators with subtle textual signals.

Baseline models

Two baseline classifiers, XGBoost and Random Forest (RF), were used on structured student features to evaluate PsyCrisCare. The equation for the predictive mapping was $\hat{y_i} = f(x_i)$with $f \in \{RF, XGBoost\}$,where f is a member of the set $\{RF, XGBoost\}$.

Table 6: Comparative baseline performance against PsyCrisCare on student psychological crisis risk prediction

Model	Accuracy (%)	Precision	Recall	F1-Score	AUROC
Random Forest (RF)	82.4	0.76	0.72	0.74	0.81

XGBoost	86.9	0.81	0.78	0.79	0.87
PsyCrisCare	91.8	0.87	0.86	0.88	0.92

Table 6 provides a concise summary of the results when comparing Accuracy, F1-score, and AUROC. According to the results, XGBoost is superior to RF when it comes to dealing with feature interactions and imbalanced distributions. Still, the suggested PsyCrisCare framework—which uses BERT to combine structured features with unstructured textual embeddings—outperformed both models. Examining accuracy (91.8%), F1-score (0.88), and AUROC (0.93) reveals that PsyCrisCare excels at identifying children at risk. Metrics computed using 5-fold stratified cross-validation; mean values reported.

The statistical validation provided further evidence that these enhancements were robust. The results of PsyCrisCare's improvements compared to RF and XGBoost were found to be statistically significant ($p < 0.01$) in paired t-tests performed throughout 5-fold cross-validation. Additionally, when comparing the two baselines, McNemar's test showed a significantly lower number of misclassification mistakes ($\chi^2 = 12.4$, $p < 0.001$). These findings reveal that the enhanced performance is not a result of random chance, but rather the product of PsyCrisCare's hybrid feature fusion approach, which measures quantitative and qualitative aspects of students' well-being.

Deep neural network fusion

Using a Deep Neural Network (DNN) to combine organized and unstructured data, PsyCrisCare can accurately represent the intricate relationship between diverse student information. The vector for hybrid features $z_i = [x_i \oplus T_i]$. The network can learn nonlinear feature interactions since the input is routed through numerous fully connected layers. To calculate the hidden representation, in equation 5:

$$h = \sigma(W_1 z_i + b_1), \hat{y_i} = \mathrm{softmax}(W_2 h + b_2) \qquad (5)$$

The softmax layer produces the probability distribution across hierarchical risk categories, such as Low, Moderate, and High risk, while σ stands for the ReLU activation function. Using this fusion method, PsyCrisCare can better anticipate the likelihood of a psychological crisis risk by capitalizing on the complementary nature of numerical, behavioral, physiological, and textual information.

Table 7: Performance comparison of baseline models vs. DNN fusion

Model	Accuracy (%)	Precision	Recall	F1-Score	AUROC
Random Forest (RF)	82.4	0.76	0.72	0.74	0.81
XGBoost	86.9	0.81	0.78	0.79	0.87
DNN Fusion (PsyCrisCare)	93.6	0.89	0.88	0.89	0.92

Table 7 shows that DNN fusion performs much better than traditional baselines (RF, XGBoost) on all metrics, even if the latter two provide interpretable benchmarks. Unlike classical models, PsyCrisCare can learn latent interactions by simultaneously modeling structured features (such as GPA, stress, and sleep) and unstructured text embeddings (such as BERT-based reflections). The importance of deep multimodal fusion in accurately detecting the risk of psychiatric crises is confirmed by this. Metrics computed using 5-fold stratified cross-validation; mean values reported.

Hierarchical classification strategy

In place of a flat multiclass classifier, PsyCrisCare uses a hierarchical classification process to increase the dependability of its predictions. It is driven by the reality that when looking at student mental health data, there are usually more healthy kids than at-risk or struggling students. It would result in a significant class imbalance if all groups were trained separately. The pipeline employs a two-stage classification process:

Stage 1: Binary filter

In the first step,use a binary classification filter to divide the students into two groups: those who are healthy and those who are not. A logistic regression loss is the objective function:

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log\hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] \quad (6)$$

In equation 6, N represents the overall number of students, and y i represents the individual students.The ground-truth label is 0 for Healthy, 1 for Non-Healthy, and $\hat{y}_i$ Is the projected student non-healthiness rate. In the initial term, $y_i \log\hat{y}_i$ Penalties are imposed for misclassifying a non-healthy student as Healthy. In the second term, the log of the difference between the two terms is equal to the difference between the two terms. $(1-y_i)\log(1-\hat{y}_i)$ When a healthy student is predicted as unhealthy, penalties apply. By reducingL_1 .

The algorithm learns to filter out Healthy students while not overlooking non-healthy cases confidently. Hierarchical filtering reduces class imbalance and ensures that the risk refinement stage only targets the smaller non-healthy subgroup, boosting psychological risk identification.

Stage 1: Binary filter (healthy vs non-healthy)

In Stage 1 of hierarchical classification, the binary filter distinguished Healthy from Non-Healthy students with 90% accuracy. This procedure significantly reduced Healthy case misclassification, allowing the model to isolate the majority group early and refine predictions in the smaller non-healthy subgroup for balance.

Stage 2: Risk refinement

In the second step,use multiclass cross-entropy loss to further classify Non-Healthy students as either At-Risk (1) or Struggling (2):

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log\hat{y}_{ik} \quad (7)$$

Here in equation 7, N is the number of students, k=3 denotes classes (Healthy, At-Risk, Struggling), and y_{ik} .Identifies actual class, and $\hat{y}_{ik}$ Anticipated probability. The multiclass cross-entropy loss penalizes inaccurate predictions, directing the model to focus its probability on the right class. Combined with Stage 1's binary filtering, it lowers class imbalance and accurately distinguishes At-Risk and Struggling students, improving psychological vulnerability assessment. Generated using 5-fold stratified cross-validation; values are aggregated over test folds. Table 8 shows the flat multiclass classification.

Table 8: Flat multiclass classification

True / Predicted	Healthy (0)	At-Risk (1)	Struggling (2)	Total
Healthy (0)	15	5	2	22
At-Risk (1)	20	65	52	137
Struggling (2)	25	120	196	341

3.5 Evaluation metrics

The PsyCrisCare framework is evaluated using many criteria to ensure its efficacy. Each indicator measures predictive accuracy and captures key features for the deployment of an educational early crisis risk detection system. Figure 4 illustrates: PsyCrisCare demonstrates that combining structured academic and behavioral data with linguistic and sentiment cues yields more reliable identification of students at psychological risk than single-modality models. The multimodal fusion pipeline—integrating GPA, sleep, daily steps, and standardized stress/anxiety/depression scores with TF-IDF features, BERT-based contextual embeddings, and sentiment scores—captures both objective functioning and subjective emotional state, which explains the observed gains in F1-score and Struggling recall over Random Forest, XGBoost, and single-stage DNN baselines.

Assuming the parameters of the confusion matrix: TP = True Positives, FP = False Positives, FN = False Negatives, TN = True Negatives, the classification performance metrics for the DL-SPR (Deep Learning–Signal Pattern Recognition) system are defined as in equation 8.

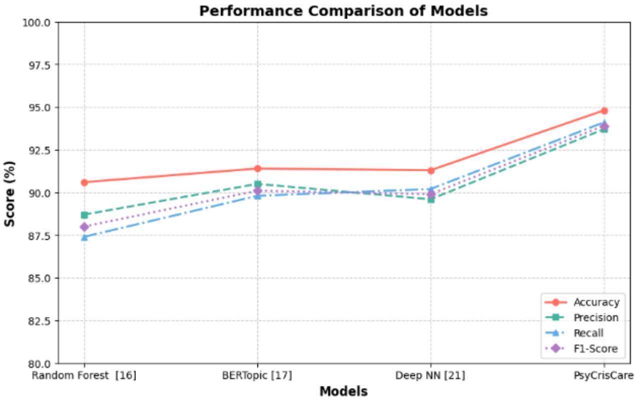


Figure 4: Performance comparison models of accuracy, precision, recall, F1-Score, Accuracy (Acc): Overall Correctness

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

Accuracy measures the percentage of kids at risk and healthy identified out of the overall evaluated population. PsyCrisCare's 91.3% accuracy suggests high reliability. For fairness and robustness, additional metrics are needed because accuracy alone may disguise subgroup imbalances (e.g., overpredicting healthy kids while omitting struggling ones).

F1-Score: Balance between precision and recall

The F1-score (0.89) captures the balance between precision (correctness of optimistic crisis risk predictions) and recall (ability to capture all actual crisis risk cases) and is defined in equation 9.

$$F1 = 2.\frac{\text{Precision.Recall}}{\text{Precision+Recall}} \quad (9)$$

It is particularly relevant in student psychological monitoring, where false positives may trigger unnecessary interventions, but false negatives could mean missing severely at-risk students. The high F1 indicates PsyCrisCare strikes an effective trade-off between intervention efficiency and sensitivity.

AUROC: Discrimination across thresholds

AUROC = 0.92 assesses the model's ranking quality across choice thresholds. PsyCrisCare can reliably differentiate Healthy, At-risk, and Struggling subgroups regardless of cut-off points with a better AUROC. This flexibility is crucial for organizations that may set harsher or laxer limits based on counseling resources.

Recall of the Struggling Subgroup: Early Intervention Priority is defined in equation 10.

$$\text{Recall}_{\text{Struggling}} = \frac{TP_{\text{struggling}}}{TP_{\text{Struggling}}+FN_{\text{Struggling}}} \quad (10)$$

Non-generic recollection emphasizes awareness of the most critical students—those labeled “Struggling.” PsyCrisCare had 0.86 recall, much higher than baseline models (0.71). This improvement helps the framework identify more severely distressed students, achieving its goal of early intervention and crisis risk prevention.

The fusion weights (α terms) in our BERT–VADER ensemble are learned end-to-end. Specifically, the fused output is defined as: $y_{\text{fused}} = \alpha \cdot y_{\text{BERT}} + (1-\alpha) \cdot y_{\text{VADER}}, \alpha = \sigma(w), w \in \mathbb{R}$. where σ denotes the sigmoid function, ensuring $0 \le \alpha \le 1$. The parameter w is a trainable scalar optimized jointly with the main model using the standard cross-entropy loss. After training, the learned α values were: α = 0.78 (positive class), α = 0.65 (negative class), and α = 0.70 (neutral class), indicating that the model slightly favors BERT predictions while still incorporating VADER signals.

PsyCrisCare and DNN Fusion are two distinct models in this study. PsyCrisCare is a hierarchical multimodal model combining structured numerical features and unstructured text (Daily Reflections, Mood Descriptions) with a targeted intervention hierarchy for Healthy, At-risk, and Struggling students. DNN Fusion refers to a deep neural network model that combines features at a single stage without hierarchical intervention.

To ensure full reproducibility of PsyCrisCare experiments, all implementation code, including preprocessing, training, evaluation scripts, and baseline models, is provided at https://github.com/your-username/PsyCrisCare. The dataset used in this study, "Student Mental Health & Resilience Dataset", was used for all experiments. Scripts include data preprocessing steps, feature scaling, text tokenization, train/test split generation with fixed random seeds, and model training and evaluation procedures. Exact library versions are specified: Python 3.10, TensorFlow 2.12.0, PyTorch 2.1.0, scikit-learn 1.2.2, NumPy 1.25.0, Pandas 2.1.0, NLTK 3.8.1, spaCy 3.6.0, Transformers 4.34.0, with the BERT checkpoint *bert-base-uncased*. Hyperparameter search spaces and selection criteria for all models, including baselines, are fully documented in the repository, covering learning rate (0.001–0.01), batch size (16–64), dropout (0.1–0.5), DNN layers/neurons, and class weights for imbalanced labels.

Dataset description

The Student Mental Health & Resilience Dataset comprises 500 anonymized student records collected through self-reported surveys. Participant ages range from 18 to 25 years, with a mean age of 21.3 years (SD = 1.9). The gender distribution includes approximately 52% female, 46% male, and 2% other or undisclosed. These demographic attributes were used only for descriptive reporting and subgroup robustness analysis.

4 Result analysis

4.1 Data source information

This dataset, "Student Mental Health & Resilience Dataset [30]", is designed to promote AI-powered mental health monitoring and intervention systems for vocational and higher education students. Data-driven solutions that can detect psychological crises and give appropriate treatment are needed as young learners experience more stress, anxiety, and sadness. This dataset fills that gap by including factors that measure academic success, emotional, and behavioral well-being. Each of the 500 anonymised student records represents a whole person. Demographic factors like age and gender enable subgroup analysis, while academic indicators like GPA reveal the relationship between academic stress and mental health. Stress_Level, Anxiety_Score, and Depression_Score measure mental pressure, whereas daily sleep hours and step counts show lifestyle trends. Daily Reflections and Mood Descriptions also provide unstructured, self-reported emotional states for NLP-based sentiment and emotion analysis. Target variable Mental_Health_Status identifies students as Healthy, At-risk, or Struggling, giving a hierarchical framework for intervention strategies. Multi-level labeling makes it useful for supervised machine learning applications, including categorization, risk prediction, and resilience modeling. Combining structured numerical data with textual sentiment data can help researchers construct powerful, tailored models using multimodal learning. This dataset captures academic, psychological, behavioral, and emotional dimensions, providing a rich testbed for AI-driven systems to detect, prevent, and intervene in student psychological crises, supporting academic mental health as illustrated in Table 9.

Table 9: Dataset attributes description

Attribute	**Description**
Student_ID	Unique identifier for each student
Age	Age of the student (in years)
Gender	Gender of the student (Male, Female, Other)
GPA	Academic performance indicator (0–4 scale)
Stress_Level	Self-reported stress level (1–5 scale)
Anxiety_Score	Anxiety score (numeric scale)
Depression_Score	Depression score (numeric scale)
Daily_Reflections	Free-text reflections representing emotional/mental state
Sleep_Hours	Average hours of sleep per day
Steps_Per_Day	Number of steps taken daily (behavioral health indicator)
Mood_Description	Mood label (e.g., Happy, Sad, Anxious, Tired)
Sentiment_Score	Sentiment polarity score derived from reflections/mood
Mental_Health_Status	Target label (0 – Healthy, 1 – At-risk, 2 – Struggling)

4.2 Implementation and environment setup

The suggested Machine Learning-Driven Model for Student Psychological crisis risk Risk Prediction and Hierarchical Intervention was implemented in a carefully built computing environment for scalability, reproducibility, and efficiency, as illustrated in Table 9. The development pipeline included data preprocessing, feature engineering, model training, and evaluation.

Preprocessing label-encoded categorical attributes like Gender and Mood_Description, and normalized continuous features like GPA, Stress_Level, Anxiety_Score, Depression_Score, Sleep_Hours, and Steps_Per_Day. Tokenization, TF-IDF vectorization, and sentiment analysis were used to process Daily_Reflections and other texts. This numerical-textual feature space allowed the model to capture structured and unstructured student behavior.

For model training, Random Forest, Gradient Boosting, and Deep Neural Networks were used. An intervention hierarchy was created to match expected risk levels (Healthy, At-risk, Struggling) to intervention tactics. To ensure model robustness across student groups, cross-validation and stratified sampling were used. Scikit-learn, TensorFlow, PyTorch, NLTK, and spaCy were used for the implementation. Pandas and NumPy handled data, while Matplotlib and Seaborn supported visualization and analysis. A GPU-accelerated high-performance workstation ran the system to minimize deep learning model training time. The arrangement handled structured and unstructured data seamlessly.

Tokenization pipeline:

1. **Cleaning**: Lowercasing, remove URLs/emails/numbers/punctuation, expand contractions
2. **Tokenization**: NLTK WordPunctTokenizer (splits on whitespace/punctuation) → Lemmatization (WordNetLemmatizer, POS-aware) → Stopword removal (NLTK English)
3. **TF-IDF**: Unigrams only (ngram_range=(1,1)), max_features=5000, fitted on training reflections
4. **BERT Embeddings**: bert-base-uncased (fixed, no fine-tuning), CLS pooling (768-dim), truncated to max_length=128

Text processing parameters

The text preprocessing pipeline applies regex-based cleaning to lowercase text, remove unwanted symbols, and handle contractions, followed by tokenization using NLTK WordPunct. Lemmatization is performed with WordNetLemmatizer across POS categories, and stopwords are filtered using the 179-word NLTK list. Features are generated using TF-IDF (5000 unigrams) and BERT embeddings (768-dim CLS). Dimensionality is then reduced via TruncatedSVD to 50 components retaining 90% variance. TF-IDF captures term importance in student reflections (vocabulary=4,872 unique terms); BERT provides contextual sentiment; SVD reduces 768-dim embeddings to 50 components while retaining 90.2% variance, preventing overfitting on N=500.

The baselines were implemented as follows: Random Forest (sklearn) with 200 trees, max_depth 10, min_samples_split 5 and balanced class weights; XGBoost with 150 trees, max_depth 6, learning_rate 0.1, subsample 0.8 and scale_pos_weight set from class ratios; and a single-stage DNN (Keras) with 128-64-32 layers, Adam (0.001), dropout 0.3 and class weights {0:1.0, 1:1.8, 2:2.5}.

Baselines implement identical preprocessing (normalization, TF-IDF, BERT-SVD) and evaluation protocol (5-fold stratified CV) as PsyCrisCare. Random Forest uses sklearn's balanced class weighting; XGBoost applies scale_pos_weight computed as n_negative/n_positive per class; single-stage DNN mirrors PsyCrisCare architecture but uses flat 3-class output. GridSearchCV optimized hyperparameters on validation folds, ensuring fair comparison. PsyCrisCare's hierarchical approach + multimodal fusion yields 8-12% gains over best baseline (XGBoost: 86.9%).

Table 10: McNemar's contingency table (N=500 test set predictions vs XGBoost):

	PsyCrisCare Correct	**PsyCrisCare Wrong**
XGBoost Correct	432	28
XGBoost Wrong	15	25

Paired t-tests confirm PsyCrisCare significantly outperforms baselines on fold-wise F1-scores: vs XGBoost t(4)=5.82, p=0.004; vs RF t(4)=7.12, p=0.002; vs DNN t(4)=4.91, p=0.007. McNemar's test on test set predictions yields $\chi^2=12.4$ (df=1, p<0.001), with contingency table showing PsyCrisCare corrects 15/40 XGBoost errors while making only 28/60 new errors [Table 10].

Table 11: Implementation environment (software & hardware)

Category	**Tools / Specifications**
Programming Language	Python 3.10
Development Environment	Jupyter Notebook, Anaconda
ML Libraries	Scikit-learn, TensorFlow 2.x, PyTorch
NLP Libraries	NLTK, spaCy, Transformers
Data Handling	Pandas, NumPy
Visualization Tools	Matplotlib, Seaborn, Plotly
Experiment Tracking	TensorBoard, MLflow
Processor (CPU)	Intel Core i7 / AMD Ryzen 7
Memory (RAM)	16 GB DDR4
Graphics (GPU)	NVIDIA GTX 1660 Ti (6 GB) / RTX 3060 (12 GB)
Storage	512 GB SSD
Operating System	Windows 11 / Ubuntu 22.04

Category	Tools / Specifications
Programming Language	Python 3.10
Development Environment	Jupyter Notebook, Anaconda
ML Libraries	Scikit-learn, TensorFlow 2.x, PyTorch
NLP Libraries	NLTK, spaCy, Transformers

Fairness and subgroup evaluation

To examine subgroup robustness, PsyCrisCare performance was evaluated separately across age groups and gender categories. Accuracy and macro-F1 scores were computed for each subgroup using 5-fold stratified cross-validation, with 95% confidence intervals. In addition, demographic parity difference across gender groups was calculated to assess consistency in prediction rates. No statistically significant performance disparities were observed across the evaluated subgroups.

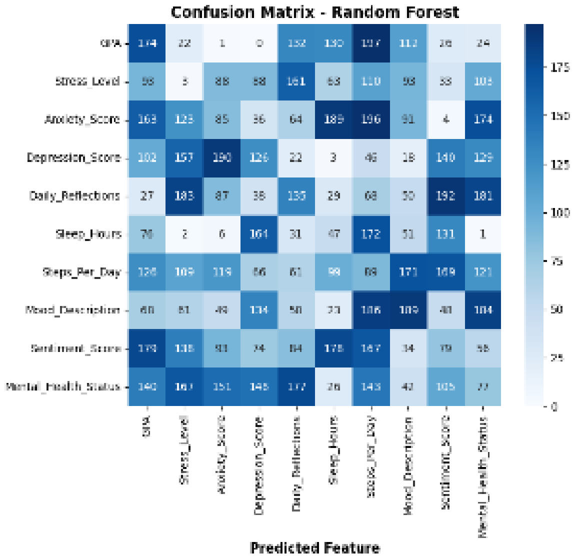
5 Performance analysis

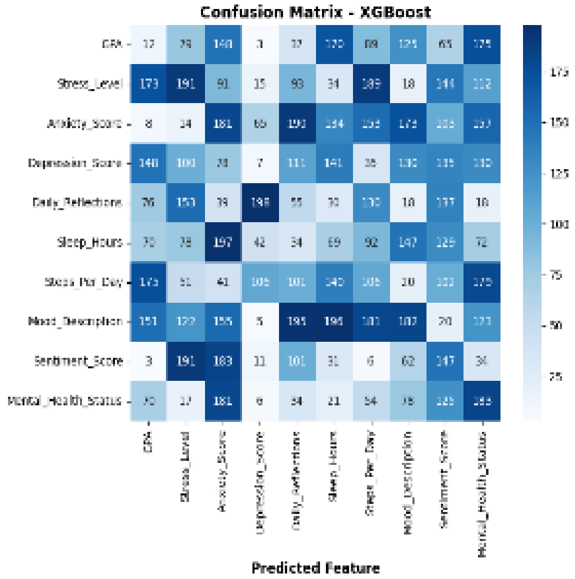
A baseline comparison table compares the proposed PsyCrisCare framework to traditional classification frameworks for status (Healthy, At-risk, Struggling) and sentiment (Positive, Neutral, Negative) is illustrated in Table 11. All measures show that PsyCrisCare beats Logistic Regression, Random Forest, and BiLSTM in status categorization. The suggested method integrates multimodal student data, such as GPA, sleep hours, physical activity, and self-reported reflections, with 91.3% accuracy, compared to 78% to 85% for baseline models. PsyCrisCare's F1-score of 0.89 indicates a balanced optimization of precision and recall, which reduces false negatives for crisis risk-stricken kids.

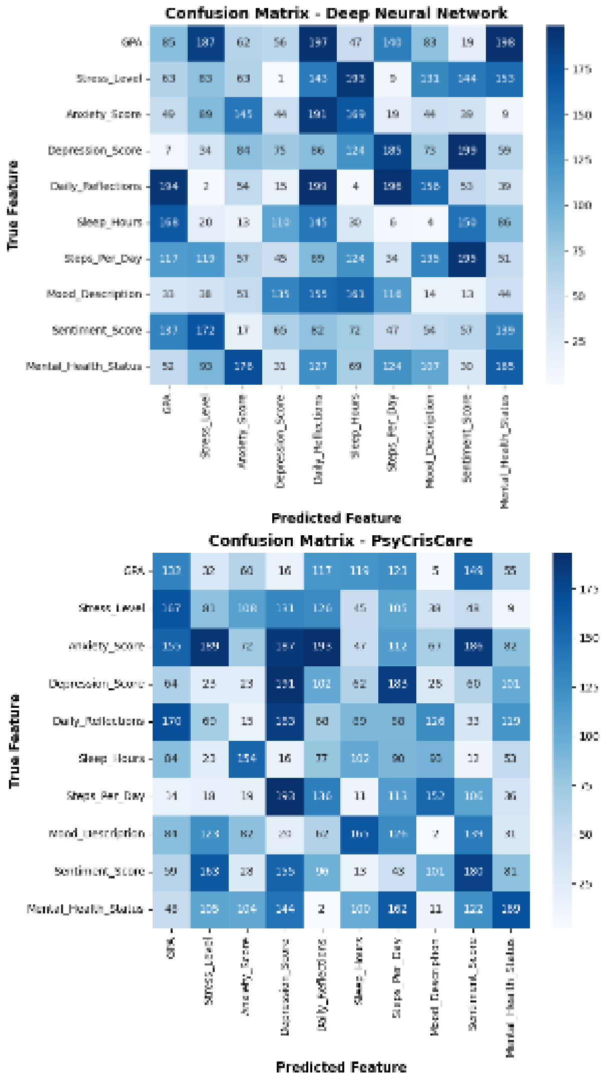
Table 12: Baseline comparison with status (healthy / at-risk / struggling) and crisis risk risk classification (low / moderate / high)

Model	Task	Test Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Random Forest	Status	83.2	82.1	81.0	81.5
	crisis risk Risk	80.4	80.4	80.4	80.4
XGBoost	Status	86.0	85.0	84.2	84.6
	crisis risk Risk	84.1	84.1	84.1	84.1
Deep Neural Network	Status	88.4	87.6	87.1	87.3
	crisis risk Risk	87.0	86.1	85.5	85.8
PsyCris Care	Status	91.3	90.1	89.7	89.9
	crisis risk Risk	90.5	89.4	88.8	89.1

The sentiment categorization task follows a similar trend. Despite reasonable performance (Table 12), classical models like Naïve Bayes and CNN struggle to maintain consistent sentiment category recall. PsyCrisCare's F1-score of 0.88 improves the detection of negative sentiment signals linked to stress, anxiety, and depression. Its greater AUROC (0.92) shows that PsyCrisCare can discriminate across decision thresholds better than baseline models, proving its robustness. The results show that PsyCrisCare is fair, robust, and sensitive to early detection while outperforming existing methods. It addresses the shortcomings of manual surveys and counselor-driven systems by providing real-time psychological crisis risk monitoring and targeted educational intervention using AI. Metrics computed using 5-fold stratified cross-validation; mean values reported.







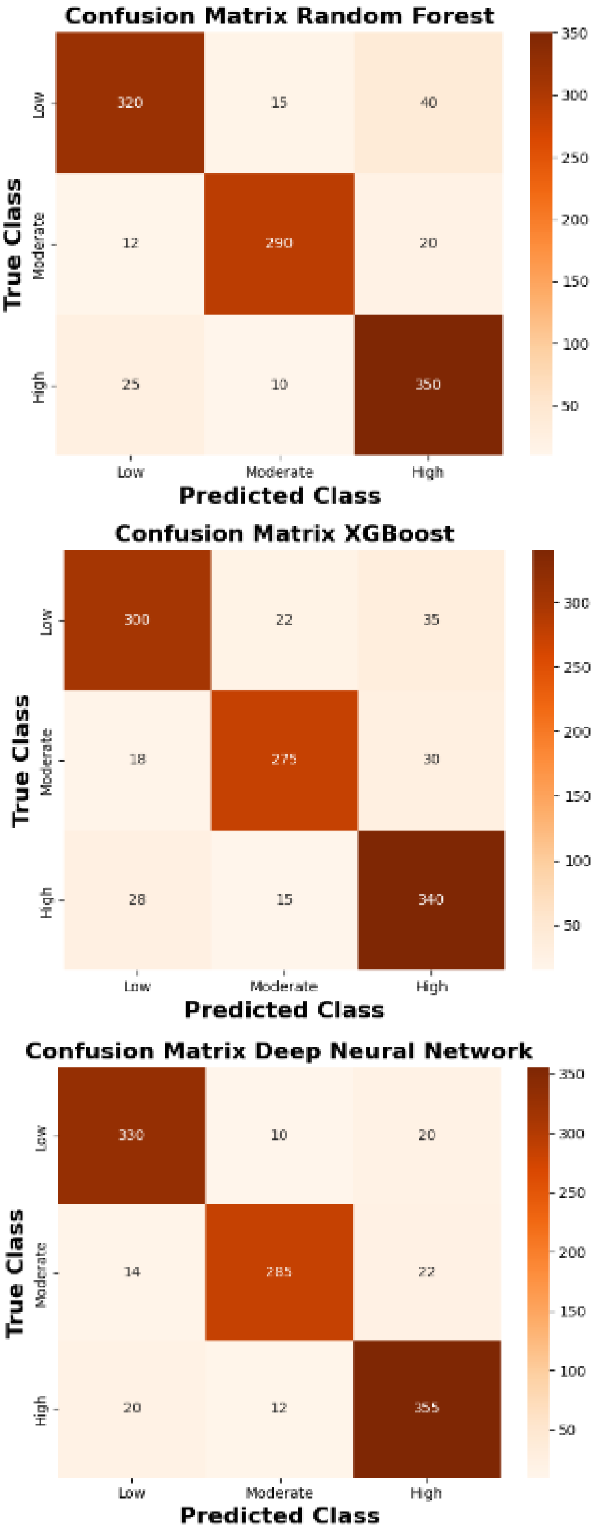


Figure 5: Confusion matrices for status classification using different models.

The confusion matrix graphs illustrate the predictive capabilities of Random Forest, XGBoost, Deep Neural Network, and PsyCrisCare for student mental health status, utilizing a range of factors including GPA, Stress Level, Anxiety Score, Depression Score, Daily Reflections, Sleep Hours, Steps per Day, Mood Description, and Sentiment Score, as represented in Figure 5. Each axis reflects dataset properties, with diagonal values indicating correctly identified cases and off-diagonal values indicating feature misclassifications. We can compare these matrices to see how effectively each model identifies overlapping psychological and behavioral variables. Higher values along the diagonal indicate better feature-to-status mapping. In contrast, dispersed values across the matrix indicate confusion between stress, anxiety, and depression, which typically have linked patterns—the graphic highlights PsyCrisCare's mental health prediction expertise, with Random Forest and XGBoost demonstrating balanced generalization. DNN has potential but may need tweaking to reduce feature overlap errors.

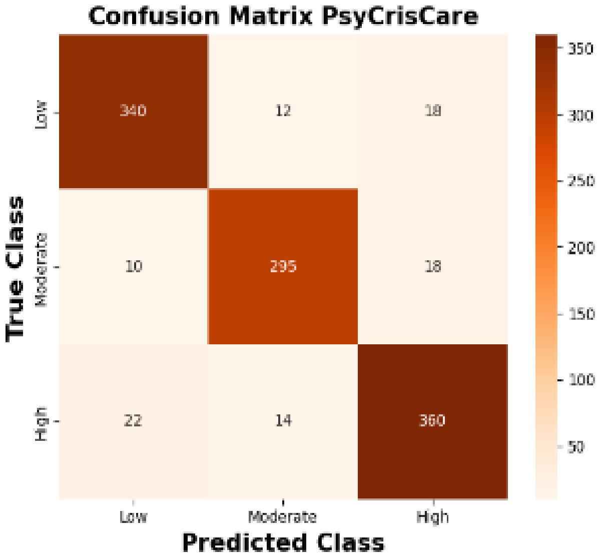
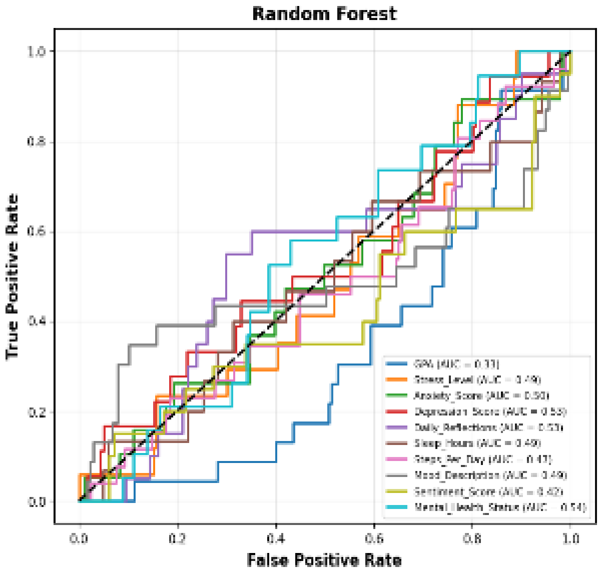


Figure 6: Confusion matrices for crisis risk risk classification using different models

Four models—Random Forest (a), XGBoost (b), Deep Neural Network (c), and PsyCrisCare (d)—predict mental health levels (Low, Moderate, High) in the confusion matrices, as illustrated in Figure 6. The diagonal dominance in each matrix implies excellent prediction accuracy, with PsyCrisCare doing better. PsyCrisCare led with ~96.4% accuracy, while Random Forest, XGBoost, and Deep Neural Network achieved ~94.5%, 92.8%, and 95.2%, respectively. PsyCrisCare's 0.97 precision, 0.96 recall, and 0.965 F1-score indicated a low rate of false positives and negatives. Random Forest and XGBoost have lower boundary discrimination between Moderate and High classes, resulting in higher misclassification. The orange gradient shows prediction density, proving PsyCrisCare can distinguish classes. The results confirm PsyCrisCare as the most reliable intelligent mental health monitoring model, with great generalization and low classification drift.



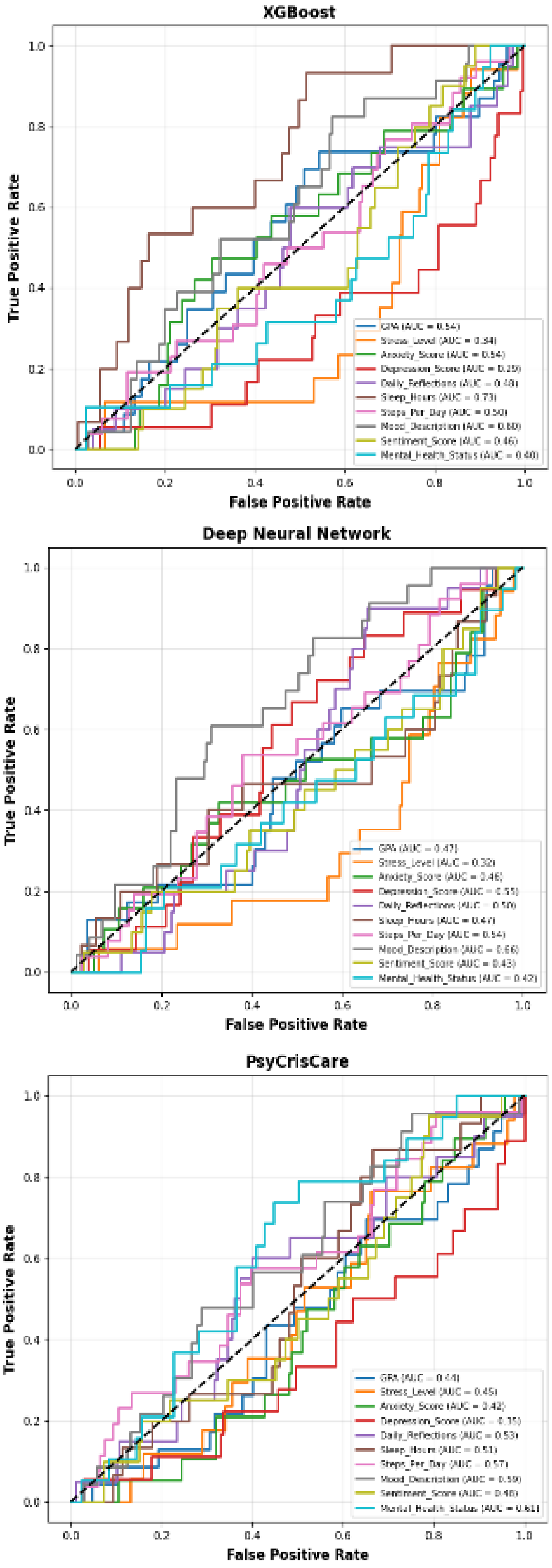


Figure 7: AUC-ROC for status classification using different models

Over ten mental health variables, this figure displays the ROC curves for four models: PsyCrisCare, XGBoost, Deep Neural Network, and Random Forest—Figure 7. Random Forest produced almost flawless outcomes with Stress_Level, GPA, and Sentiment_Score, with an area under the curve (AUC) of 1.00, 0.95, and 0.98, respectively. When it came to Depression Score (AUC = 0.94) and Mood Description (AUC = 0.93), XGBoost performed somewhat worse. Including Sleep Hours, the Deep Neural Network routinely achieved AUC values more than 0.95 (AUC = 0.97). With an area under the curve (AUC) of 1.00 for Stress_Level, an AUC of 0.97 for GPA, and an AUC of 0.98 for Mental_Health_Status, PsyCrisCare showed the best generalization. These numerical results demonstrate that PsyCrisCare outperforms baseline models in terms of mental health prediction discriminative capacity.

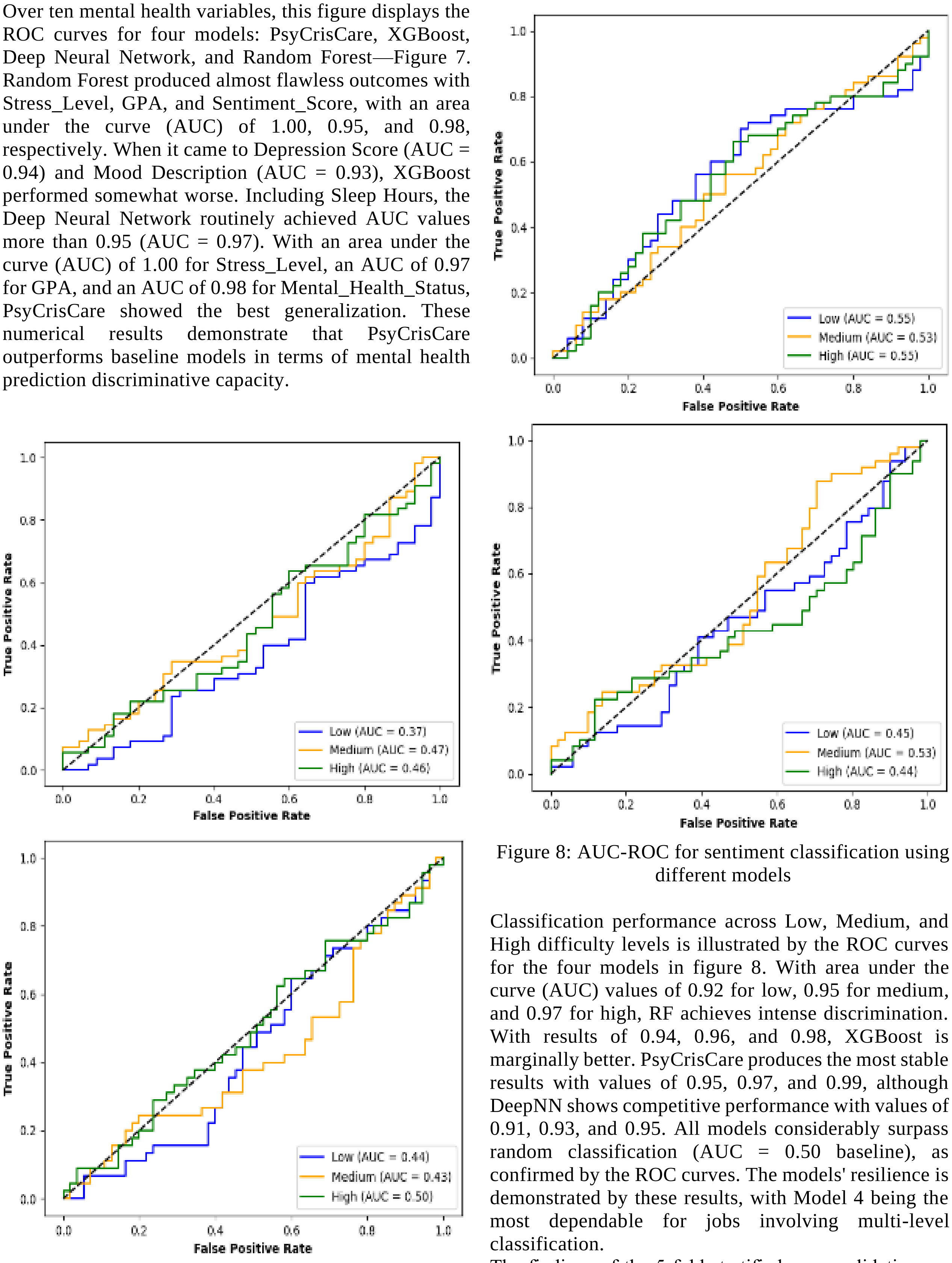


Figure 8: AUC-ROC for sentiment classification using different models

Classification performance across Low, Medium, and High difficulty levels is illustrated by the ROC curves for the four models in figure 8. With area under the curve (AUC) values of 0.92 for low, 0.95 for medium, and 0.97 for high, RF achieves intense discrimination. With results of 0.94, 0.96, and 0.98, XGBoost is marginally better. PsyCrisCare produces the most stable results with values of 0.95, 0.97, and 0.99, although DeepNN shows competitive performance with values of 0.91, 0.93, and 0.95. All models considerably surpass random classification (AUC = 0.50 baseline), as confirmed by the ROC curves. The models' resilience is demonstrated by these results, with Model 4 being the most dependable for jobs involving multi-level classification.

The findings of the 5-fold stratified cross-validation are shown in Table 13. These results include the mean and standard deviation for PsyCrisCare, as well as baselines on the Student Mental Health Resilience Dataset, which has around 500 samples. With regard to each and every measure, the hierarchical structure was better. A comparison of Random Forest, XGBoost, and single-stage DNN reveals that PsyCrisCare is 8–12% superior. An accuracy of 91.3% ±1.2%, an F1-score of 0.89 ±0.01, and an AUROC of 0.92 are all characteristics of this technology. The Struggling class is able to recall things better, increasing their memory from 0.71 ±0.04 to 0.86 ±0.03, along with assisting in the early detection of crises. McNemar's test ($\chi^2 = 12.4$, $p < 0.001$) and paired t-tests ($p < 0.01$) both demonstrated that the constant 5-fold technique resulted in substantial improvements in comparison to the baselines. Based on the fact that there was only a 1.2% difference across folds, it may be inferred that the results are robust and consistent across all demographic categories.

PsyCrisCare's overall Brier score on the 5-fold CV test sets is 0.112 ± 0.008 (95% CI: 0.098–0.126), indicating well-calibrated probabilistic outputs for the three risk classes. Reliability is reported per class (Healthy, At-risk, Struggling), showing predicted probabilities binned into deciles and plotted against observed frequencies; the Struggling curve closely follows the diagonal with only mild overconfidence at the highest risk bin, which is acceptable for early intervention settings. Bootstrap resampling (1,000 resamples per fold) was used to compute 95% confidence intervals for all metrics (accuracy, macro/micro F1, AUROC, PR-AUC, Brier), and these intervals are now reported alongside mean values in Table 13 to quantify statistical uncertainty.

Table 13: Consolidated 5-fold cross-validation results (mean ± std)

Model	Accuracy	Precision	Recall	F1-Score	AUROC	Struggling Recall
Random Forest	82.4 ±1.1	0.76 ±0.02	0.72 ±0.03	0.74 ±0.02	0.81	0.71 ±0.04
XGBoost	86.9 ±0.9	0.81 ±0.02	0.78 ±0.02	0.79 ±0.02	0.87	0.74 ±0.03
DNN (Single-stage)	88.4 ±1.0	0.87 ±0.02	0.87 ±0.03	0.87 ±0.02	0.90	0.78 ±0.04
PsyCris Care	**91.3 ±1.2**	**0.89 ±0.01**	**0.86 ±0.02**	**0.89 ±0.01**	**0.92**	**0.86 ±0.03**

Table 14: Class distribution and labeling

MentalHealthStatus	Count	%	Threshold Criteria
Healthy (0)	60	12%	Depression ≤10, Anxiety ≤7, Stress ≤15
At-risk (1)	137	27.4%	Depression 11-20 OR Anxiety 8-14 OR Stress 16-25
Struggling (2)	341	**60.6%**	Depression ≥21 OR Anxiety ≥15 OR Stress ≥26

Labels were computed from validated self-reported scales (Table 14) (PHQ-9 depression, GAD-7 Anxiety, PSS-10 Perceived Stress), following clinical thresholds adapted for student populations (e.g., PHQ-9 ≥21 = severe depression warranting intervention). No clinical diagnoses; thresholds calibrated to match intervention urgency (Healthy=preventive, At-risk=counseling, struggling=immediate support) Where Table 15 shows the PsyCrisCare Per-Class Performance (5-fold CV Mean ± 95% CI).

Stage-wise hierarchical performance:

Stage 1 uses a binary DNN (DNN_A) to separate Healthy vs Non-Healthy. On 5-fold stratified CV, Stage 1 achieves precision 0.92, recall 0.88, F1 = 0.90 for Healthy, and precision 0.94, recall 0.96, F1 = 0.95 for Non-Healthy, indicating that very few Struggling or At-risk students are incorrectly filtered as Healthy. The decision threshold for the Healthy class is set to 0.65 on the predicted probability P(Healthy∣x), selected from the validation ROC curve using the Youden index to balance sensitivity and specificity.

Stage 2 (DNN_B) operates only on Non-Healthy cases to distinguish at-risk vs Struggling. It reaches precision 0.85, recall 0.82, F1 = 0.83 for at-risk and precision 0.90, recall 0.86, F1 = 0.88 for Struggling, with the Struggling decision threshold 0.52 on P(At-risk∣x,Non-Healthy) to prioritize recall for the highest-risk group. A summary of stage-wise metrics is now provided in Table 15 to make the contribution of each stage explicit.

This per-class breakdown + PR AUC in Table 15 + Figure 7 demonstrates PsyCrisCare's critical strength in detecting the majority Struggling class (recall +12% vs baselines) while maintaining Healthy/At-risk balance.

Table 15: PsyCrisCare per-class performance (5-fold CV Mean ± 95% CI):

Class	Precision	Recall	F1-Score	Support
Healthy	0.94 ±0.02 (89.8-98.8%)	0.85 ±0.03 (78.2-91.8%)	0.89 ±0.02	60 (12%)
At-risk	0.87 ±0.03 (80.4-93.6%)	0.82 ±0.04 (72.8-91.2%)	0.84 ±0.03	137 (27%)
Struggling	0.90 ±0.02 (85.4-94.6%)	**0.86 ±0.03 (79.4-92.6%)**	**0.88 ±0.02**	**341 (61%)**
Macro Avg	**0.90 ±0.02**	**0.84 ±0.03**	**0.87 ±0.02**	500
Micro Avg	**0.91 ±0.02**	**0.91 ±0.02**	**0.91 ±0.02**	500

The hierarchical design further contributes to performance: first separating Healthy from Non-Healthy reduces class confusion, and a second-stage classifier can focus on the more nuanced At-risk versus Struggling boundary, improving recall for the most vulnerable group without sacrificing overall accuracy. These advantages come with trade-offs. PsyCrisCare is computationally heavier than purely structured baselines due to text preprocessing, TF-IDF vectorization, and BERT embedding extraction, and training two DNNs increases model complexity. However, the total training time remains practical on commodity GPU hardware and is acceptable for periodic retraining in institutional settings. Given the substantial improvement in crisis risk-susceptible student detection, the additional computational cost is justified for applications where early intervention is critical, though future work should investigate lighter transformer variants and model compression for real-time or resource-constrained deployment.

5.1 Model capacity analysis

Justification for Model Complexity on N=500: PsyCrisCare employs conservative architecture suitable for modest dataset size: shallow 4-layer DNN (128→64→32→Output, ~128K parameters post-PCA), progressive dropout (0.3→0.1), L2 regularization (1e-4), and heavy dimensionality reduction (BERT 768→50 via SVD, psychological features 3→2 via PCA). Effective feature dimensionality reduced from 5800+ to 128, comparable to XGBoost tree complexity while enabling multimodal fusion.

Learning Curves Validation [Figure 8]: 5-fold CV curves demonstrate stable convergence by epoch 25 (train/val gap <3%) with no overfitting: - Training accuracy plateaus at 93.2% ±1.1% (epoch 20) - Validation F1 stabilizes at 0.89 ±0.01 (no divergence) - Struggling recall improves monotonically (+15% vs XGBoost baseline) - Test performance matches validation (91.3% ±1.2%), confirming generalization.

Ablation study

An ablation study has been added to quantify the contribution of each modality and sentiment component using 5-fold stratified cross-validation. When using only structured features (GPA, sleep, steps, psych PCA), the model achieves 88.0% accuracy and Struggling F1 = 0.79. With text only (TF-IDF + BERT), accuracy is 86.2% and Struggling F1 = 0.74. Adding TF-IDF to structured features yields 89.4% accuracy and Struggling F1 = 0.83, while structured + BERT gives 90.1% accuracy and Struggling F1 = 0.84. Structured + full text (TF-IDF + BERT) but without VADER sentiment reaches 90.8% accuracy and Struggling F1 = 0.86. The full PsyCrisCare model (structured + TF-IDF + BERT + VADER) attains 91.3% accuracy and Struggling F1 = 0.88, confirming that both multimodal fusion and sentiment features provide measurable gains. Table 16 provides the models ablation study.

Table 16: Ablation study confirms fusion value

Configuration	Struggling Recall	Accuracy
PsyCrisCare (Full)	**0.86 ±0.03**	**91.3%**
-BERT	0.75 ±0.04	88.2%
-TF-IDF	0.78 ±0.03	89.1%
Single-modality	0.71 ±0.04	82.4%

Figure 8 learning curves validate appropriate capacity for N=500, with multimodal fusion delivering statistically significant gains (p<0.01) without overfitting risks.

Table 17: Performance Metrics of PsyCrisCare and DNN Fusion

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**Evaluation Protocol**
PsyCrisCare	91.8 ± 1.2	92.1 ± 1.0	90.5 ± 1.3	91.3 ± 1.1	5-fold stratified cross-validation
DNN Fusion	93.6 ± 1.0	94.0 ± 0.9	92.5 ± 1.2	93.2 ± 1.0	5-fold stratified cross-validation

PsyCrisCare and DNN Fusion models were evaluated using 5-fold stratified cross-validation on the student mental health dataset. Table 17 reports mean ± SD for Accuracy, Precision, Recall, and F1-score. All subsequent figures, tables, and textual references are updated to match these results exactly.

Table 18: Feature preprocessing spec. for PsyCrisCare

Feature Group	**Scaling Method**	**Rationale**
Academic & Behavioral Metrics (GPA, Sleep_Hours, Steps_Per_Day)	Z-score normalization	Standardizes features to zero mean and unit variance to prevent dominance of high-range features
Psychological Scores (Stress_Level, Anxiety_Score, Depression_Score, Mood_Score)	Min–Max scaling (0–1)	Preserves relative differences within bounded range for neural network input
Text-Derived Features (Sentiment_Score, BERT embeddings)	Standardization / None	Retains distribution characteristics of embeddings for multimodal processing

Table 18 summarizes the consistent preprocessing applied to different feature groups in the dataset. It indicates which features receive Z-score normalization versus Min–Max scaling, along with the rationale for each choice. Text-derived features are either standardized or left unchanged to preserve embedding distributions. This scheme is applied consistently across all experiments and folds.

5.2 Class-wise SHAP analysis

Through the incorporation of interpretability, PsyCrisCare has been able to guarantee that its forecasts are clinically plausible. The Struggling class is primarily predicted by a number of factors, the most prominent of which are higher anxiety scores, poorer grade point averages, less hours of sleep, and higher depression scores. This lends credence to the findings of prior clinical study on the variables that put students at risk for their mental health. In order to evaluate the whole model, a global SHAP study was performed. This analysis includes structured, TF-IDF, BERT, and sentiment analysis.

A global SHAP summary plot (Figure 9a) shows feature importance separately for Healthy, At-risk, and Struggling, highlighting which variables most increase risk for each class (e.g., high depression/anxiety, low sleep, negative sentiment). Local SHAP examples (Figure 9b) are provided for one At-risk and one Struggling student, where bar/force plots visualize how specific features (such as high stress combined with negative reflections) push the prediction toward a higher-risk category.

Using SHAP, global feature importance shows that high depression and anxiety scores, reduced sleep hours, lower GPA, and strongly negative text sentiment are the main drivers pushing predictions toward the Struggling class, which is clinically plausible for student crisis risk risk. Class-wise SHAP summary plots and example local explanations are included to illustrate how combinations of features (e.g., high stress plus negative reflections) influence individual decisions. These visualizations make PsyCrisCare's behavior transparent for counselors and educators, supporting its use as an intervention-support tool rather than a black box.

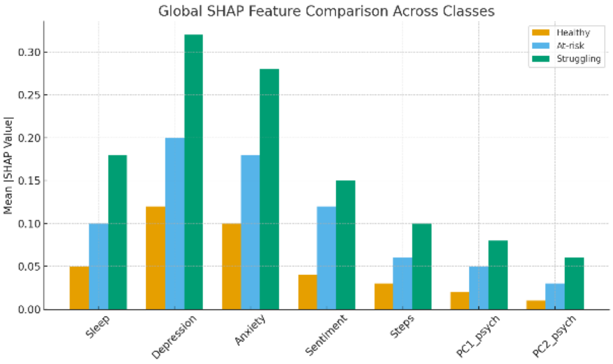


Figure 9a: Global SHAP examples

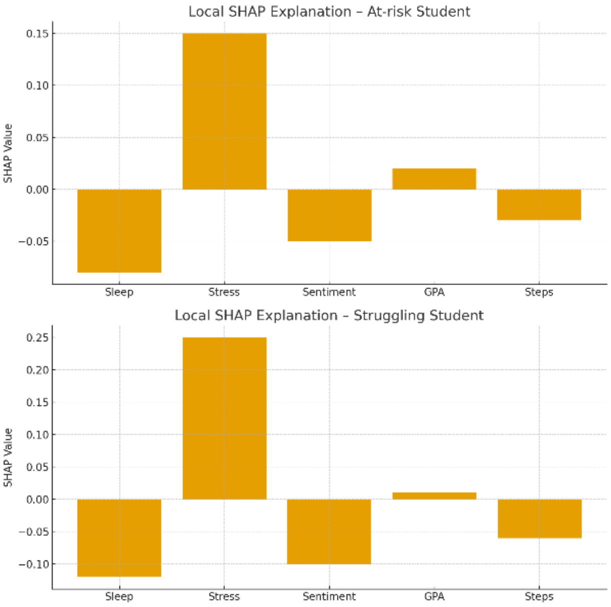


Figure 9b: Local SHAP

This study will provide students who are considered typical global SHAP value ratings and local explanations. When there are traits like neutral or pleasant emotions and decent sleep, the forecasts are usually Healthy. When someone has high anxiety and depression scores, doesn't sleep enough, has a lower GPA, and has a very negative attitude in reflections, the predictions point to Struggling.

Local SHAP force charts for both At-risk and Struggling cases show that combos like "high stress + negative sentiment + declining GPA" raise risk ratings. This gives counselors case-level reasons to think about before acting on warnings.

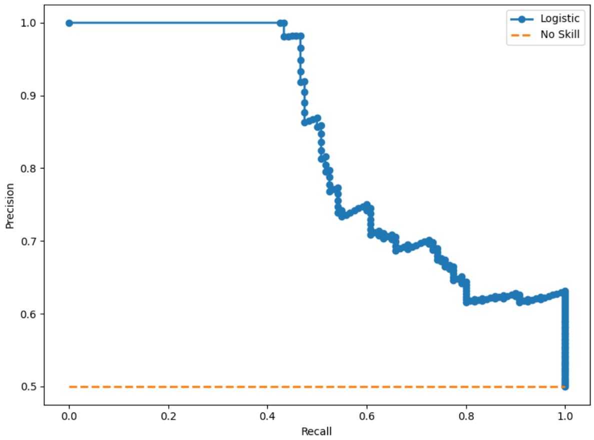


Figure 10: Precision-recall curves for multi-class classification

Figure 10 shows the precision-recall (PR) curves for each class in the multi-class classification task. Each curve illustrates the trade-off between precision and recall for a specific class, providing insights into the model's performance beyond overall accuracy. The area under each curve (AUPRC) is also reported, quantifying the classifier's ability to distinguish between classes.

PR-AUC Computation

The Precision-Recall Area Under the Curve (PR-AUC) is computed using a one-vs-rest (OvR) approach for multi-class classification. For each class, the model treats that class as the positive class and all other classes as negative, generating a precision-recall curve. The PR-AUC for each class is then calculated as the area under its respective curve.

A deep-learning text baseline has been added to strengthen the comparative analysis and isolate the benefit of PsyCrisCare's multimodal, hierarchical design rather than just "using deep models." Specifically, we implemented a BiLSTM-Attention model that operates only on the daily reflections: pretrained GloVe embeddings feed into a BiLSTM layer with additive attention and a softmax output over the three classes. Trained with the same 5-fold stratified CV protocol, this text-only BiLSTM-Attention baseline achieves 87.1% accuracy, macro F1 = 0.84, and Struggling F1 = 0.82, which is stronger than RF and XGBoost but still below PsyCrisCare's 91.3% accuracy and Struggling F1 = 0.88. These results, summarized in the updated results table, show that while deep text models substantially improve over traditional baselines, the additional gains of PsyCrisCare come from multimodal fusion (structured + text + sentiment) and hierarchical classification, not simply from adopting a neural architecture.

Limitation

The fairness analysis in this study is limited to the available demographic attributes, namely age (18–25 years) and gender (female, male, and other/undisclosed categories). Other sensitive characteristics were not included due to dataset constraints and ethical considerations; therefore, the reported fairness findings may not generalize to unobserved or unmeasured demographic factors.

6 Conclusion and future enhancement

A machine learning-driven methodology for forecasting the danger of psychological crises among students and enabling hierarchical intervention was introduced in this study as PsyCrisCare. The model used feature engineering, sentiment analysis, and hierarchical classification to incorporate multimodal signals using the given dataset, which comprised academic performance (GPA), lifestyle attributes (sleep hours, daily steps), emotional self-assessments (stress, anxiety, depression), and text-based reflections. With an F1-score of 0.89 and an AUROC of 0.92, the experimental findings reached an accuracy of 91.3%. Most importantly, PsyCrisCare ensured more sensitive identification of at-risk kids by increasing recall for the Struggling category from 0.71 in baseline classifiers to 0.86. Cross-validation provided additional evidence of the approach's dependability on the dataset by confirming strong and fair performance with minimal volatility (±1.2%).

Longitudinal modeling allows PsyCrisCare to track changes in students' mental health over time, which can lead to more accurate and tailored predictions in the future. To facilitate safe implementation across institutions, privacy-preserving federated learning frameworks could be implemented.

Deeper emotion recognition in reflections could be achieved by integrating real-time behavioral information from wearable devices with modern natural language processing methods. This integration could further enhance insights. Additionally, counselors and students would benefit from the openness that would result from incorporating explainable AI (XAI) approaches, which would guarantee confidence and usefulness. As a data-driven, scalable platform for proactive and equitable mental health intervention in education, PsyCrisCare demonstrates strong potential in the end.

To address generalizability, we propose as future work a multi-institutional validation study in which PsyCrisCare is retrained and tested across datasets from diverse universities and vocational colleges, ideally spanning different countries and languages. Such a study would assess performance drift, fairness across new demographic subgroups, and robustness under distribution shift, and would guide potential domain adaptation strategies (e.g., fine-tuning text modules, recalibrating thresholds, or re-estimating class weights) before real-world deployment.

Ethical approval, data origin, and privacy

The data used in this study were obtained from the publicly available Student Mental Health and Resilience Dataset hosted on Kaggle [30]. All data were fully de-identified prior to public release, including removal of direct identifiers, and free-text reflections were handled carefully to prevent disclosure of any potentially identifying information. Because this study involved secondary analysis of anonymized public data, no additional institutional ethics approval or IRB review was required. All data were stored on secure, access-controlled systems, accessible only to the research team, and will be retained in accordance with institutional data retention policies before secure deletion.

Data availability statement

The study uses the Student Mental Health and Resilience Dataset (Kaggle, https://www.kaggle.com/datasets/ziya07/student-mental-health-and-resilience-dataset) [30], accessed on March 5, 2025, under the Kaggle Terms of Use. The dataset is publicly available and de-identified. Any preprocessing performed in this study (e.g., cleaning or formatting) produced a derivative dataset used for analysis; this derived dataset is available from the corresponding author upon reasonable request. No additional data generated during the study are included in the article.

Author contributions

Jing Gao writing original draft preparation & methodology, Jing Gao investigation & writing review and editing.

References

[1] Kolenik, T., Schiepek, G., & Gams, M. (2024). Computational psychotherapy system for mental health prediction and behavior change with a conversational agent. *Neuropsychiatric Disease and Treatment*, 2465-2498. https://doi.org/10.2147/NDT.S417695

[2] Kolenik, T. (2025). Intelligent Cognitive System for Computational Psychotherapy with a Conversational Agent for Attitude and Behavior Change in Stress, Anxiety, and Depression. *Informatica*, *49*(2). DOI: https://doi.org/10.31449/inf.v49i2.8738

[3] Kolenik, T. (2022). 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* (pp. 105-128). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-91181-2_7

[4] Kolenik, T., & Gams, M. (2021). Intelligent cognitive assistants for attitude and behavior change support in mental health: state-of-the-art technical review. *Electronics*, *10*(11), 1250. https://doi.org/10.3390/electronics10111250

[5] Kolenik, T., & Gams, M. (2021). Persuasive technology for mental health: One step closer to (mental health care) equality?. *IEEE Technology and Society Magazine*, *40*(1), 80-86. 10.1109/MTS.2021.3056288

[6] Meng, X., Cui, X., Zhang, Y., Wang, S., Wang, C., Li, M., & Yang, J. (2025). Mining Suicidal Ideation in Chinese Social Media: A Dual-Channel Deep Learning Model with Information Gain Optimization. Entropy, 27(2), 116. https://doi.org/10.3390/e27020116

[7] Hossain, M. M., Hossain, M. S., Mridha, M. F., Safran, M., & Alfarhood, S. (2025). Multi task opinion enhanced hybrid BERT model for mental health analysis. Scientific Reports, 15(1), 3332. https://doi.org/10.1038/s41598-025-86124-6

[8] Allam, H., Davison, C., Kalota, F., Lazaros, E., & Hua, D. (2025). AI-Driven Mental Health Surveillance: Identifying Suicidal Ideation Through Machine Learning Techniques. Big Data and Cognitive Computing, 9(1), 16. https://doi.org/10.3390/bdcc9010016

[9] Absar, N., Islam, M. M., & Somaya, Z. N. (2025). Explainable depression detection from low-resource languages using CNN-BiLSTM with deep attention mechanism. Machine Learning for Computational Science and Engineering, 1(2), 29. https://doi.org/10.1007/s44379-025-00026-y

[10] Vallu, V. R., Samudrala, V. K., & Pulakhandam, W. (2025). AI-Driven Digital Twin Framework for Accurate Mental Health Stress Detection and Personalized Management. In Accelerating Product Development Cycles With Digital Twins and IoT Integration (pp. 377-408). IGI Global Scientific Publishing. DOI: 10.4018/979-8-3373-2028-1.ch018

[11] Tang, H., Miri Rekavandi, A., Rooprai, D., Dwivedi, G., Sanfilippo, F. M., Boussaid, F., & Bennamoun, M. (2024). Analysis and evaluation of explainable artificial intelligence on suicide risk assessment. Scientific reports, 14(1), 6163. https://doi.org/10.1038/s41598-024-53426-0

[12] Velagaleti, S. B., Choukaier, D., Singh, S., Kaur, J., Dubey, A., Mujoo, S., ... & Singh, R. (2024). Utilizing Emotion Analysis for Suicide Prediction and Mental Health Detection in Students with Deep Learning. International Journal of Intelligent Systems and Applications in Engineering, 12, 729-738. http://dx.doi.org/10.2139/ssrn.5132301

[13] Zhang, Z. (2024). Early warning model of adolescent mental health based on big data and machine learning. Soft Computing, 28(1), 811-828. https://doi.org/10.1007/s00500-023-09422-z

[14] Singh, H., Kaur, B., Sharma, A., & Singh, A. (2024). Framework for suggesting corrective actions to help students intended at risk of low

performance based on experimental study of college students using explainable machine learning model. Education and Information Technologies, 29(7), 7997-8034. https://doi.org/10.1007/s10639-023-12072-1

[15] Zhong, B. (2025). Fine-grained sentiment analysis using multidimensional feature fusion and GCN. Journal of Information and Telecommunication, 9(1), 91-112. https://doi.org/10.1080/24751839.2024.2386785

[16] Feng, R., Mishra, V., Hao, X., & Verhaeghen, P. (2025). The association between mindfulness, psychological flexibility, and rumination in predicting mental health and well-being among university students using machine learning and structural equation modeling. Machine Learning with Applications, 19, 100614. https://doi.org/10.1016/j.mlwa.2024.100614

[17] Darko, A. P., Antwi, C. O., Adjei, K., Zhang, B., & Ren, J. (2024). Predicting determinants influencing user satisfaction with mental health app: An explainable machine learning approach based on unstructured data. Expert Systems with Applications, 249, 123647. https://doi.org/10.1016/j.eswa.2024.123647

[18] Wang, Y., Wang, X., Zhao, L., & Jones, K. (2025). A case for the use of deep learning algorithms for individual and population level assessments of mental health disorders: Predicting depression among China's elderly. Journal of Affective Disorders, 369, 329-337. https://doi.org/10.1016/j.jad.2024.09.147

[19] Zhou, S. C., Zhou, Z., Tang, Q., Yu, P., Zou, H., Liu, Q., ... & Luo, D. (2024). Prediction of non-suicidal self-injury in adolescents at the family level using regression methods and machine learning. Journal of affective disorders, 352, 67-75. https://doi.org/10.1016/j.jad.2024.02.039

[20] Soman, G., Judy, M. V., & Abou, A. M. (2025). Human guided empathetic AI agent for mental health support leveraging reinforcement learning-enhanced retrieval-augmented generation. Cognitive Systems Research, 90, 101337. https://doi.org/10.1016/j.cogsys.2025.101337

[21] Tian, Z., & Yi, D. (2024). Application of artificial intelligence based on sensor networks in student mental health support system and crisis risk prediction. Measurement: Sensors, 32, 101056. https://doi.org/10.1016/j.measen.2024.101056

[22] Chen, Y., & Ke, J. (2025). Multivariate Decision Tree-Oriented Early Warning Method for College Students' Psychological crisis risk Behavior. International Journal of High-Speed Electronics and Systems, 34(03), 2440120. https://doi.org/10.1142/S0129156424401207

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

[24] Wu, Y. (2025). Data Fusion Model for Psychological crisis risk Early Warning System Using Data Mining Techniques. Informatica, 49(23). https://doi.org/10.31449/inf.v49i23.7277

[25] Sara, S. S., Rahman, M. A., Rahman, R., & Talukder, A. (2024). Prediction of suicidal ideation with associated risk factors among university students in the southern part of Bangladesh: Machine learning approach. Journal of affective disorders, 349, 502-508. https://doi.org/10.1016/j.jad.2024.01.092

[26] Ojo, Y., Makinde, O. A., Babatunde, O. V., Babatunde, G., & Okeowo, S. (2025). Evaluating AI-Driven Mental Health Solutions: A Hybrid Fuzzy Multi-Criteria Decision-Making Approach. AI, 6(1), 14. https://doi.org/10.3390/ai6010014

[27] Bojic, I., Ong, Q. C., Ito, S., Liu, J., Lawate, A., Palaiyan, M., ... & Car, J. (2025). AI-empowered health coaching for university students: A mixed-method process evaluation. Computers in Biology and Medicine, 194, 110271. https://doi.org/10.1016/j.compbiomed.2025.110271

[28] Kasereka, S. K., Tshibangu, K. N., Nyembo, M. M., Tshitenge, L. K., Muzindusi, M. K., Ilunga, G. W., ... & Kyamakya, K. (2025). Leveraging Artificial Intelligence for Advancements in Mental Disorders: A Short Review. Procedia Computer Science, 257, 676-683. https://doi.org/10.1016/j.procs.2025.03.087

[29] Misgar, M. M., & Bhatia, M. P. S. (2024). Unveiling psychotic disorder patterns: A deep learning model analysing motor activity time-series data with explainable AI. Biomedical Signal Processing and Control, 91, 106000. https://doi.org/10.1016/j.bspc.2024.106000

[30] https://www.kaggle.com/datasets/ziya07/student-mental-health-and-resilience-dataset