RESEARCH ARTICLE

A Software-Oriented Deep Learning Framework for Multiclass Academic Stress and Mental Health Risk Prediction

Shoaib Ahmad[1], **Ali Imran**[1], **Tahoona Kshif**[1], **Zubdah Tur Rasheed**[1], **Ayesha Khan**[1], **Aamir Ali**[1,*] **and Muhammad Ali**[1]

[1] Department of Computer Science, Quaid-e-Azam College of Engineering and Technology, Sahiwal 57000, Pakistan

Abstract

Student mental health and academic stress are significant concerns in higher education, creating a need for effective approaches to early risk identification. This study proposes a multiclass predictive framework for academic stress and mental health risk classification among students. The framework was developed using a Kaggle dataset containing 25,000 records with demographic, academic, behavioral, psychological, lifestyle, and medical attributes. A unified comparison was conducted across six conventional machine learning algorithms, namely Decision Tree, Random Forest, XGBoost, CatBoost, K-Nearest Neighbors (KNN), and Naïve Bayes, and three deep learning architectures, namely Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM). Class imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE), applied only to the training data, and model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrices. CatBoost and BiLSTM achieved the highest classification accuracy of 97%, with weighted F1-scores of 0.97 and macro F1-scores of 0.96. The results indicate that ensemble and deep learning approaches can effectively support multiclass mental health risk classification. Because the provenance and label construction of the secondary dataset are not documented, the results should be read as within-dataset performance. The findings provide a basis for future real-world validation and decision-support applications.

Keywords: academic stress, mental health risk, deep learning, multiclass classification, BiLSTM, CatBoost, student well-being, SMOTE, predictive analytics.

1 Introduction

Mental health challenges among university students are influenced by academic workload, financial pressures, social conditions, and psychological stress, with potential consequences for students' well-being and academic performance [1–5]. These

Submitted: 04 August 2026
Revised: 12 September 2026
Accepted: 17 September 2026
Published: 22 September 2026

Vol. 2, **No.** 3, 2026.
10.62762/JSE.2026.690220

***Corresponding author:**
✉ Aamir Ali
amirali4436823@gmail.com

Citation
Ahmad, S., Imran, A., Kshif, T., Rasheed, Z. T., Khan, A., Ali, A., & Ali, M. (2026). A Software-Oriented Deep Learning Framework for Multiclass Academic Stress and Mental Health Risk Prediction. *ICCK Journal of Software Engineering*, 2(3), 220–247.

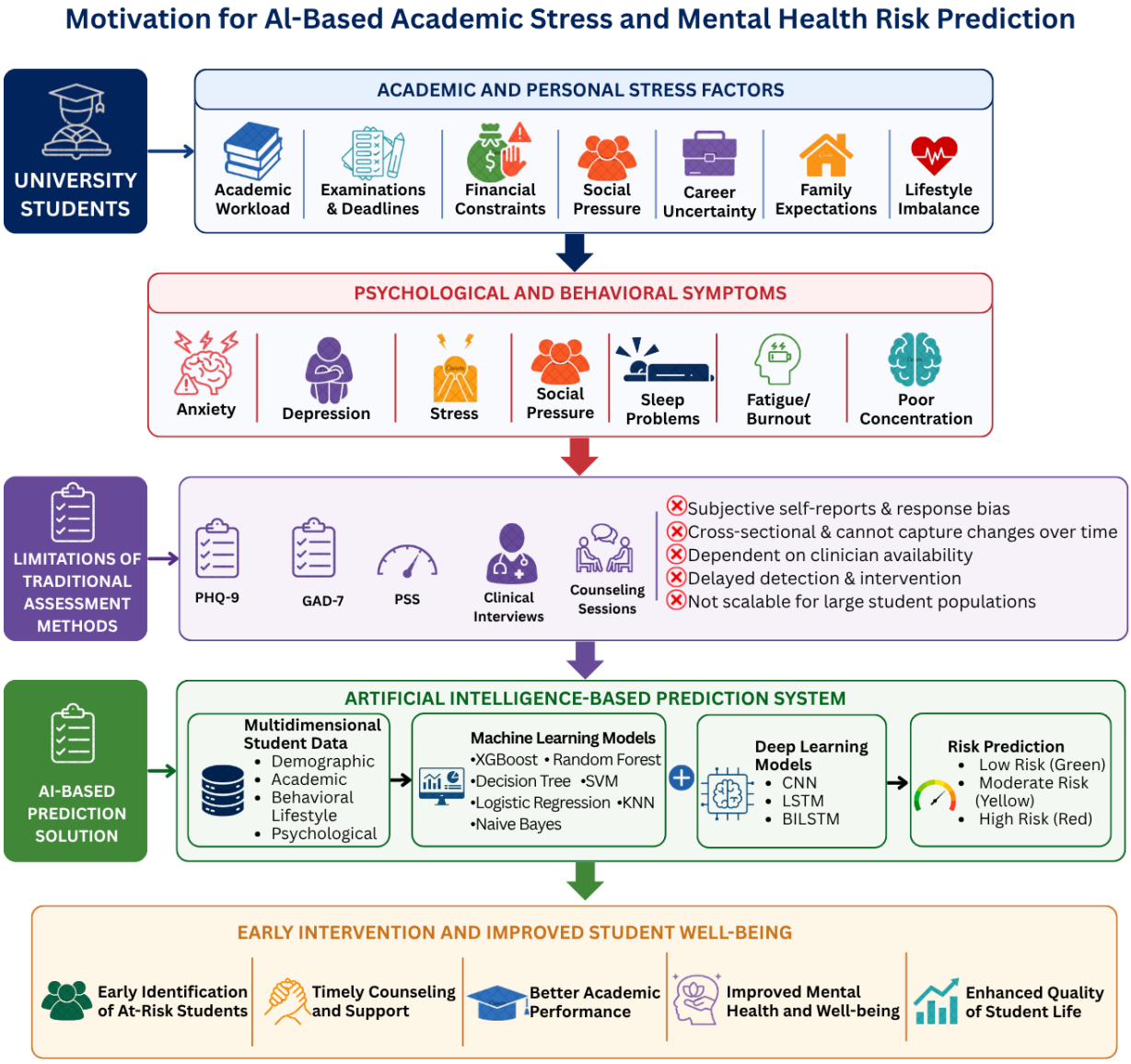

Figure 1. Motivation for multiclass prediction of student mental health risk.

challenges have increased the need for effective approaches that can assist in identifying students at different levels of mental health risk. Conventional mental health screening approaches for students, including self-reported measures, counseling, and clinical interviews, have limited sensitivity and specificity for identifying individuals at risk, and their reach is constrained by professional availability and participation rates. Computational approaches using demographic, lifestyle, and behavioral data have been proposed to support more scalable risk assessment among student populations [6]. However, such screening data may also be affected by response bias and measurement variability, which can reduce the reliability of model inputs derived from self-reported assessments [7]. These limitations motivate the development of computational approaches that can analyze complex student-related data and assist in the identification of individuals who may be at different levels of mental health risk [8, 9]. Artificial Intelligence (AI), particularly Machine Learning (ML) and Deep Learning (DL), has increasingly been applied to the analysis and prediction of mental health-related outcomes [10, 11]. These approaches can support automated analysis of behavioral and psychological factors and have demonstrated potential for predictive risk detection [12]. Figure 1 presents the motivation for the proposed framework by linking student-related stressors and risk factors with multiclass mental health risk prediction.

The framework is motivated by the need to examine

multiple student-related factors and classify mental health risk into low, moderate, and high categories. This motivation directly supports the study's focus on evaluating alternative machine learning and deep learning models for multiclass risk classification. Machine learning algorithms have increasingly been used for predicting mental health-related outcomes. Traditional approaches such as Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) have been applied to the classification of stress, depression, anxiety, and related psychological outcomes [13]. Association rule mining approaches have also been explored for identifying patterns in student mental health risk data [14]. Deep learning approaches, including Long Short-Term Memory (LSTM), Gated Recurrent Units (GRUs), and Convolutional Neural Networks (CNNs), have also demonstrated potential in mental health prediction across diverse data sources, including clinical records, student data, and social media text [15–17]. However, existing approaches vary in their classification settings, model selection, datasets, and evaluation strategies, making direct comparison across conventional machine learning and deep learning approaches difficult.

Despite the increasing application of machine learning and deep learning to mental health prediction, several gaps remain in the existing literature. Existing studies have frequently examined binary or limited-class prediction settings and individual machine learning or deep learning approaches, while direct comparison between conventional machine learning and deep learning models remains limited. In addition, relatively limited attention has been given to multiclass prediction of different mental health risk levels using a combination of demographic, academic, behavioral, psychological, lifestyle, and medical factors. Class imbalance can further affect multiclass prediction performance, particularly for less represented risk categories. These gaps motivate the present study, which develops a three-level multiclass prediction framework and evaluates six conventional machine learning algorithms alongside three deep learning architectures under a common experimental setting. A comparison of the relevant existing studies and their methodological characteristics is provided in Table 1 in Section 2.

This study develops a deep learning-based multiclass prediction framework for assessing student mental health risk across low, moderate, and high risk levels. The study examines demographic, behavioral, psychological, academic, and lifestyle-related factors and evaluates the performance of CNN, LSTM, and BiLSTM models in comparison with six conventional machine learning algorithms: Decision Tree, Random Forest, XGBoost, CatBoost, K-Nearest Neighbors (KNN), and Naïve Bayes. The objective is to identify the strongest-performing models for multiclass mental health risk classification. This study makes the following contributions:

1. It proposes a multiclass predictive framework for identifying students across low, moderate, and high mental health risk levels, extending the prediction task beyond binary classification and providing a basis for future early risk assessment and decision-support applications.

2. It evaluates three deep learning architectures, namely CNN, LSTM, and BiLSTM, for multiclass mental health risk classification and examines their performance alongside conventional machine learning approaches.

3. It provides a unified comparative evaluation of six machine learning algorithms, namely Decision Tree, Random Forest, XGBoost, CatBoost, K-Nearest Neighbors, and Naïve Bayes, against the selected deep learning architectures using common evaluation criteria.

4. It incorporates demographic, academic, behavioral, psychological, lifestyle, and medical attributes within a multidimensional feature space and addresses class imbalance through the Synthetic Minority Over-sampling Technique (SMOTE) [18] applied to the training data.

To guide the investigation, this study addresses the following research questions:

RQ1: Which demographic, academic, behavioral, psychological, lifestyle, and medical factors contribute to the classification of students into low, moderate, and high mental health risk levels?

RQ2: How do conventional machine learning and deep learning models differ in their performance for multiclass student mental health risk classification?

RQ3: Which of the evaluated machine learning and deep learning models achieves the strongest overall performance across the selected classification metrics?

RQ4: How does the performance of the best-performing models compare with previously

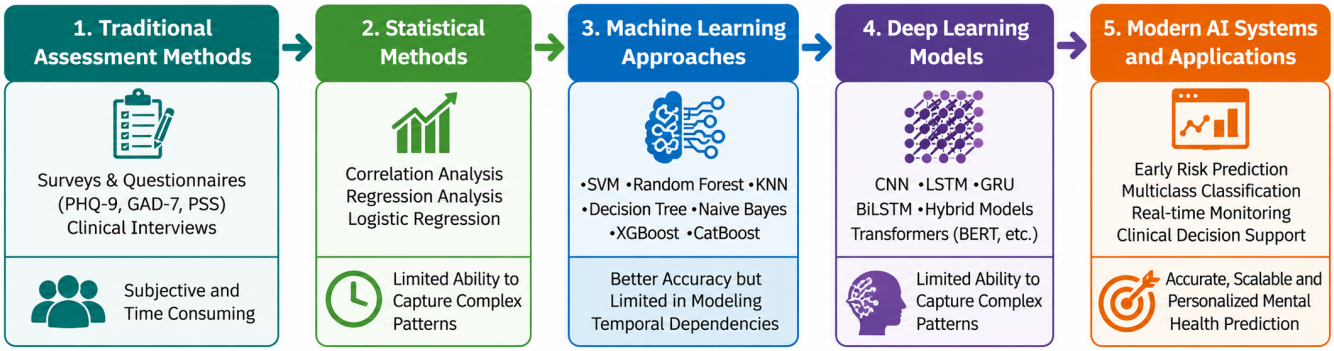

Figure 2. Evolution of AI techniques for mental health prediction.

reported machine learning and deep learning approaches for student mental health risk prediction?

This paper is organized as follows. Section 2 reviews related work on machine learning and deep learning approaches for mental health prediction. Section 3 presents the research methodology, including dataset preparation, preprocessing, data balancing, model development, and evaluation. Section 4 presents and discusses the experimental results and comparative performance of the evaluated machine learning and deep learning models. Section 5 discusses the study limitations and future research directions. Section 6 concludes the paper and summarizes the main findings and potential directions for future work.

2 Literature Review

Artificial Intelligence (AI) based approaches for mental health prediction have evolved from conventional machine learning models toward deep learning and hybrid architectures. Earlier studies primarily employed structured survey, behavioral, academic, and clinical variables with conventional classifiers, while more recent research has incorporated deep learning, transformer-based architectures, and heterogeneous data sources such as social media and electronic health records. Figure 2 summarizes this methodological evolution and provides context for the progression from conventional machine learning toward more complex learning architectures in mental health prediction.

This evolution also reveals an important methodological issue. Although increasingly sophisticated models have been introduced, their reported performance cannot always be compared directly because studies differ in target definitions, datasets, feature spaces, class structures, and evaluation protocols. The present study therefore focuses on a common experimental setting in which conventional machine learning and deep learning models are evaluated for the same multiclass student mental health risk classification task.

2.1 Machine Learning Approaches for Mental Health Prediction

Conventional machine learning approaches have been widely applied to mental health prediction using clinical and behavioral data, including data from electronic health records [19], as well as student-related survey and academic data. Garriga et al. [20] developed an XGBoost-based model for mental health crisis prediction using longitudinal electronic health records from 17,122 patients. The model achieved an AUROC of 0.797, while the lack of data from multiple centers limited its generalizability to other populations. Rahman et al. [3] investigated mental well-being among 15,366 university students across multiple ASEAN universities using health behavioral indicators. Random Forest achieved the strongest reported performance, with 90.1% accuracy and an AUC of 0.966. These studies illustrate the applicability of tree-based machine learning approaches to mental health prediction across different data sources and populations, although their research contexts, datasets, and prediction targets differ substantially.

Tutun et al. [21] developed an AI-based decision support approach for predicting mental health disorders using more than 6,000 responses to SCL-90-R psychological assessments from a general clinical population. Logistic Regression and Random Forest were evaluated, and the reported

accuracy was 89%. The study illustrates the use of structured psychological assessment data for automated prediction, but its research context differs from student-focused settings, and its applicability to broader student populations and operational educational environments remains limited.

Student-focused studies have also applied conventional machine learning to stress, depression, anxiety, and related outcomes. Mutalib et al. [22] evaluated Decision Tree, Neural Network, SVM, Naïve Bayes, and Logistic Regression using survey data from 629 higher education students. Decision Tree achieved 84.44% accuracy for stress prediction, while SVM achieved 88.15% for depression prediction. The Neural Network produced the strongest result for anxiety prediction. The relatively small sample restricted prediction robustness, illustrating the importance of adequate and diverse training data for student mental health prediction [22].

Meda et al. [23] examined severe depressive symptoms and suicidal ideation among university students using baseline and six-month follow-up data. Random Forest models achieved a balanced accuracy of 0.85 for predicting sustained well-being. However, the model showed limited ability to predict worsening symptoms, which restricts its usefulness for early identification of deteriorating mental health.

Luo et al. [24] used a Random Forest (ExtraTrees) model with 33 variables to predict depression among 9,478 Chinese undergraduate students (from 10,043 initial respondents). The model achieved 87.5% accuracy and an AUC of 0.927. However, its cross-sectional design, single-region sample, and reliance on self-reported data limit the generalizability and interpretation of the findings [24]. Hasan et al. [25] applied seven machine learning techniques to Depression Anxiety Stress Scales (DASS-21) data from 1,697 Bangladeshi university students. SVM performed best for depression prediction, Logistic Regression for anxiety, and CatBoost for stress prediction. The cross-sectional design and narrow student population limit the generalizability of the reported findings [25]. Macalli et al. [26] used Random Forest with 70 baseline predictors to estimate suicidal behaviors and thoughts among 5,066 students in the French i-Share cohort. Although the model showed good predictive performance (AUC of about 0.8), reliance on self-reported data and low follow-up participation remained important limitations.

Sharma et al. [27] evaluated Naïve Bayes, Logistic Regression, Multilayer Perceptron, Bayes Net, J48, and Random Forest using data from more than 200 university students. Random Forest achieved the highest accuracy of 94.73%. However, the university-based sample was relatively small, limiting the generalizability of the findings [27]. Chung and Teo [28] similarly compared multiple single and ensemble machine learning classifiers using the Open Sourcing Mental Illness (OSMI) mental health survey dataset. Gradient Boosting achieved 88.80% accuracy followed by Neural Networks at 88.00%. However, the study focused on binary prediction using survey-based data.

2.2 Deep Learning and Hybrid Approaches

Deep learning and hybrid approaches have increasingly been applied to mental health-related prediction, particularly when datasets contain linguistic, social media, or complex clinical information [29]. Machine learning methods have also been applied to structured survey data from university student populations to predict depression, anxiety, and stress outcomes, demonstrating that conventional classifiers can effectively support mental health risk assessment in educational contexts [30]. Venkatachalam and Sivanraju proposed the SADDL model, combining LSTM, a deep convolutional neural network (DCNN), and a multilayer perceptron (MLP) with Latent Dirichlet Allocation (LDA)-based feature extraction to predict student academic performance using academic, demographic, mental health, mood-related, and linguistic attributes. The model achieved 91.71% accuracy with an 80:20 split and 89.3% with a 70:30 split [17]. However, the limited dataset size and time-intensive data collection constrain the scalability and generalizability of the approach [17].

Shrestha and Srinivasan compared deep learning and conventional machine learning approaches using 150,085 psychiatric clinical notes from hospital records. The evaluated models included CNN, LSTM, attention-based models, transformers, SVM, and Logistic Regression. The authors' proposed CB-MH architecture achieved an F1-score of 0.62, while an attention-based model achieved an F2-score of 0.71 [31]. Despite the large clinical dataset, uncertainty in complex clinical contexts contributed to false-negative predictions [31]. The clinical and text-based nature of this study differs substantially from structured tabular student data, limiting direct comparability with student-focused prediction settings. Social

Table 1. Summary of related studies on machine learning and deep learning-based mental health risk prediction.

Author Name and Reference	Objective	Dataset Used	Method / Approach	Key Findings / Accuracy	Limitations
Venkatachalam and Sivanraju [17]	To predict student academic performance using academic, demographic, mental health, mood-related, and linguistic attributes.	Data from engineering students and social media posts from Twitter, Facebook, and Instagram.	LDA for feature extraction; LSTM, DCNN, and MLP in the SADDL model.	Achieved 91.71% accuracy with 80:20 split and 89.3% accuracy with 70:30 split.	Dataset size was limited, and data collection was time-consuming.
Rahman et al. [3]	To predict mental well-being among university students using health behavior indicators.	15,366 students from multiple ASEAN universities.	Random Forest, KNN, Neural Networks, and other ML models.	Random Forest achieved the highest accuracy of 90.1% and AUC of 0.966.	Relied on self-reported survey data, which may introduce response bias.
Mutalib et al. [22]	To predict stress, depression, and anxiety among higher education students.	Survey data from 629 students in Kuala Terengganu using DASS-21 and WHOQOL-based factors.	Decision Tree, Neural Network, SVM, Naïve Bayes, and Logistic Regression.	Decision Tree achieved 84.44% accuracy for stress; SVM achieved 88.15% accuracy for depression.	Small sample size limited prediction robustness.
Shrestha and Srinivasan [31]	To compare DL and conventional ML models for mental illness prediction.	150,085 psychiatry clinical notes from history of present illness records.	CNN, LSTM, attention models, transformers, SVM, and Logistic Regression.	CB-MH achieved best F1-score of 0.62; attention model achieved best F2-score of 0.71.	False negatives were difficult due to unclear clinical contexts.
Karamat et al. [32]	To perform multiclass mental illness prediction using social media text.	Social media text related to depression, anxiety, bipolar disorder, and PTSD.	MentalBERT, MelBERT, and CNN-based hybrid transformer architecture.	Achieved 92% accuracy and 92% F1-score.	Computationally complex and dependent on social media text quality.
Luo et al. [24]	To identify depression predictors among Chinese college students.	10,043 undergraduate students surveyed (9,478 analysed); depression assessed using the CES-D scale.	Random Forest Algorithm with 33 variables.	Achieved 87.5% accuracy and AUC of 0.927.	Cross-sectional design, single-region sample, and self-reported data limited causal interpretation.
Sharma et al. [27]	To predict student stress using machine learning techniques.	More than 200 university students assessed using PSS scale.	Naïve Bayes, Logistic Regression, MLP, Bayes Net, J48, and Random Forest.	Random Forest achieved the highest accuracy of 94.73%.	Small university-based dataset limited generalizability.
Chung and Teo [28]	To compare single and ensemble ML classifiers for mental health prediction.	OSMI mental health survey dataset.	Logistic Regression, Gradient Boosting, Neural Networks, KNN, SVM, XGBoost, DNN, and ensemble model.	Gradient Boosting achieved the highest accuracy of 88.80%, followed by Neural Networks with 88.00%.	Focused on binary prediction and survey-based data.

media-based deep learning approaches have also been investigated: Gkotsis et al. [16] applied convolutional neural networks to Reddit posts, reaching 91.08% accuracy in separating mental health-related posts from other posts and a weighted accuracy of 71.37% across 11 mental-illness themes. However, such text-based approaches depend on social media data, which limits their applicability to structured multiclass student risk assessment.

Karamat et al. [32] developed a hybrid transformer-based architecture for multiclass mental illness prediction using social media text. The approach combined MentalBERT, MelBERT, and CNN components to distinguish depression, anxiety, bipolar disorder, and PTSD. It achieved 92% accuracy and a 92% F1-score. However, the computational complexity of the approach and its dependence on social media text quality and pretrained language models represent important practical limitations [32].

2.3 Multiclass Mental Health Prediction

Multiclass prediction has been investigated in previous mental health research, and therefore the research gap should not be framed as an absence of multiclass studies. Karamat et al. [32], for example, performed multiclass mental illness prediction across depression, anxiety, bipolar disorder, and PTSD using social media text and a hybrid transformer architecture. Other studies have examined multiple psychological outcomes, including stress, depression, and anxiety, using several machine learning classifiers [22, 25]. However, the prediction setting addressed in these studies differs from the specific task considered

in the present research. Existing multiclass or multi-outcome studies primarily distinguish different mental disorders or predict individual psychological outcomes, whereas comparatively limited attention has been given to three-level student mental health risk classification, specifically the classification of students into low, moderate, and high risk categories using a multidimensional combination of demographic, academic, behavioral, psychological, lifestyle, and medical factors. In addition, direct comparison of conventional machine learning and deep learning models within the same experimental setting remains limited.

Therefore, the contribution of the present study is not the introduction of multiclass prediction itself. Instead, it lies in evaluating a specific three-level student mental health risk classification task using a common experimental setting and comparing six conventional machine learning algorithms with three deep learning architectures. Table 1 provides a consolidated comparison of representative studies in terms of their objectives, datasets, methodological approaches, reported performance, and limitations. The table highlights the diversity of prediction targets, data sources, model families, and methodological constraints across the existing literature.

Figure 3 summarizes the major research gaps identified from the reviewed literature and maps them to the corresponding scope of the present study. The first gap concerns the prediction setting. Although multiclass prediction has been investigated, comparatively limited work addresses three-level student mental health risk classification into low, moderate, and high categories. The second gap concerns methodological comparison, as relatively limited research directly evaluates conventional machine learning and deep learning approaches within the same experimental setting. The third gap concerns feature diversity, as previous studies frequently rely on specific data sources such as psychological assessments, clinical records, or social media, whereas the present study considers demographic, academic, behavioral, psychological, lifestyle, and medical attributes. The fourth gap concerns class imbalance, which can affect the representation and prediction of less frequent risk categories. The present study addresses this issue by applying SMOTE only to the training data.

The final gap concerns the comparability of reported predictive performance. Differences in datasets, prediction targets, class definitions, and evaluation protocols make direct comparison across previous studies difficult. The present study addresses this limitation by evaluating six conventional machine learning algorithms and three deep learning architectures using a common dataset, preprocessing procedure, class-balancing strategy, and evaluation framework.

2.4 Critical Synthesis and Research Gaps

The comparison of previous studies reveals several recurring methodological limitations. First, differences in sample size and population characteristics influence the generalizability of predictive findings. Studies based on relatively small or institution-specific student populations may not adequately represent the diversity of broader student populations [22, 27]. Second, reliance on self-reported survey data can introduce response bias and measurement variability, potentially affecting the reliability of model inputs [3, 24–26]. Third, cross-sectional designs provide limited evidence regarding changes in mental health risk over time and constrain interpretation of predictive factors [24, 25]. Fourth, studies based on social media or clinical text depend on domain-specific data quality and representations, which can limit their applicability to other prediction settings [16, 31, 32].

A further limitation is the lack of methodological comparability across studies. Existing research differs in prediction targets, class definitions, datasets, feature spaces, data splitting strategies, and evaluation measures. Consequently, reported accuracy and F1-scores cannot always be interpreted as direct evidence that one model family is superior to another. For example, the 92% accuracy reported by Karamat et al.[32] was obtained for multiclass mental illness prediction using social media text, whereas other studies focus on specific outcomes such as stress, depression, anxiety, or mental well-being [3, 22, 24, 25]. A common experimental setting is therefore needed to provide a more controlled comparison between conventional machine learning and deep learning approaches.

3 Materials and Methods

3.1 Research Design

This study employed a supervised multiclass classification design to develop and evaluate machine learning and deep learning models for student mental health risk classification. The primary objective was to classify the individuals represented in the dataset into three mental health risk

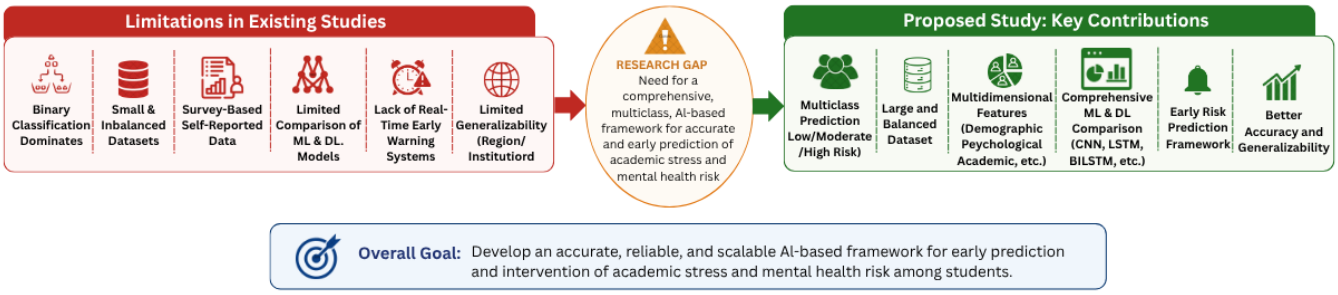

Figure 3. Research Gap Analysis and Proposed Contribution.

categories, namely low, moderate, and high risk, using demographic, academic, employment, behavioral, lifestyle, psychological, and medical or family-history related attributes.

The overall research workflow consisted of dataset acquisition and inspection, data preprocessing, stratified division into training, validation, and testing subsets, class balancing using SMOTE on the training data only, model development, hyperparameter optimization, model training, and independent evaluation. Six conventional machine learning models, namely Decision Tree, Random Forest, XGBoost, CatBoost, K-Nearest Neighbors, and Naïve Bayes, were compared with three deep learning architectures, namely Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM).

The models were evaluated under the same experimental setting using the same training, validation, and independent test partitions. This design enabled a controlled comparison of conventional machine learning and deep learning approaches and facilitated the identification of the strongest-performing model for the three-class mental health risk prediction task.

In this study, the term software-oriented refers to the implementation of the proposed prediction approach as a computational pipeline in which data preparation, feature transformation, class balancing, model training, prediction, and performance evaluation are integrated into a reproducible machine learning and deep learning workflow. The term emphasizes the computational implementation and comparative evaluation of the framework rather than implying clinical deployment or diagnosis. The framework is intended for automated risk classification and research-oriented decision support rather than clinical diagnosis. Figure 4 presents the overall workflow of the proposed software-oriented prediction framework.

3.2 Dataset Description

The study used a publicly available secondary dataset obtained from Kaggle [33]. The dataset contains 25,000 records and 24 predictor variables covering demographic, academic, behavioral, psychological, lifestyle, and medical characteristics. Because the dataset was obtained as a secondary dataset, the present study did not participate in participant recruitment, primary data collection, questionnaire administration, or assignment of the original risk labels. The available dataset documentation does not provide sufficient information to independently establish the original sampling frame, recruitment procedure, participating institutions, geographic distribution, response rate, or inclusion and exclusion criteria of the participants. Therefore, the original population and sampling characteristics cannot be reliably reconstructed from the available source. These limitations are explicitly acknowledged, and the dataset is treated as a secondary computational resource rather than as a representative sample of the wider student population. Accordingly, the reported results should be interpreted within the characteristics and distribution of this dataset and should not be generalized to students from other institutions or populations without external validation. Table 2 summarizes the source, dataset characteristics, population information, and sampling details available from the original Kaggle source. Where the original documentation does not provide sufficient information, this is explicitly reported as not documented rather than inferred by the authors.

The available dataset documentation does not specify whether the low, moderate, and high risk labels were clinically assessed, survey-derived, expert-assigned, or algorithmically generated. Therefore, the present study treats mental_health_risk as a pre-existing target variable in the secondary dataset and does not interpret these labels as clinically validated diagnoses

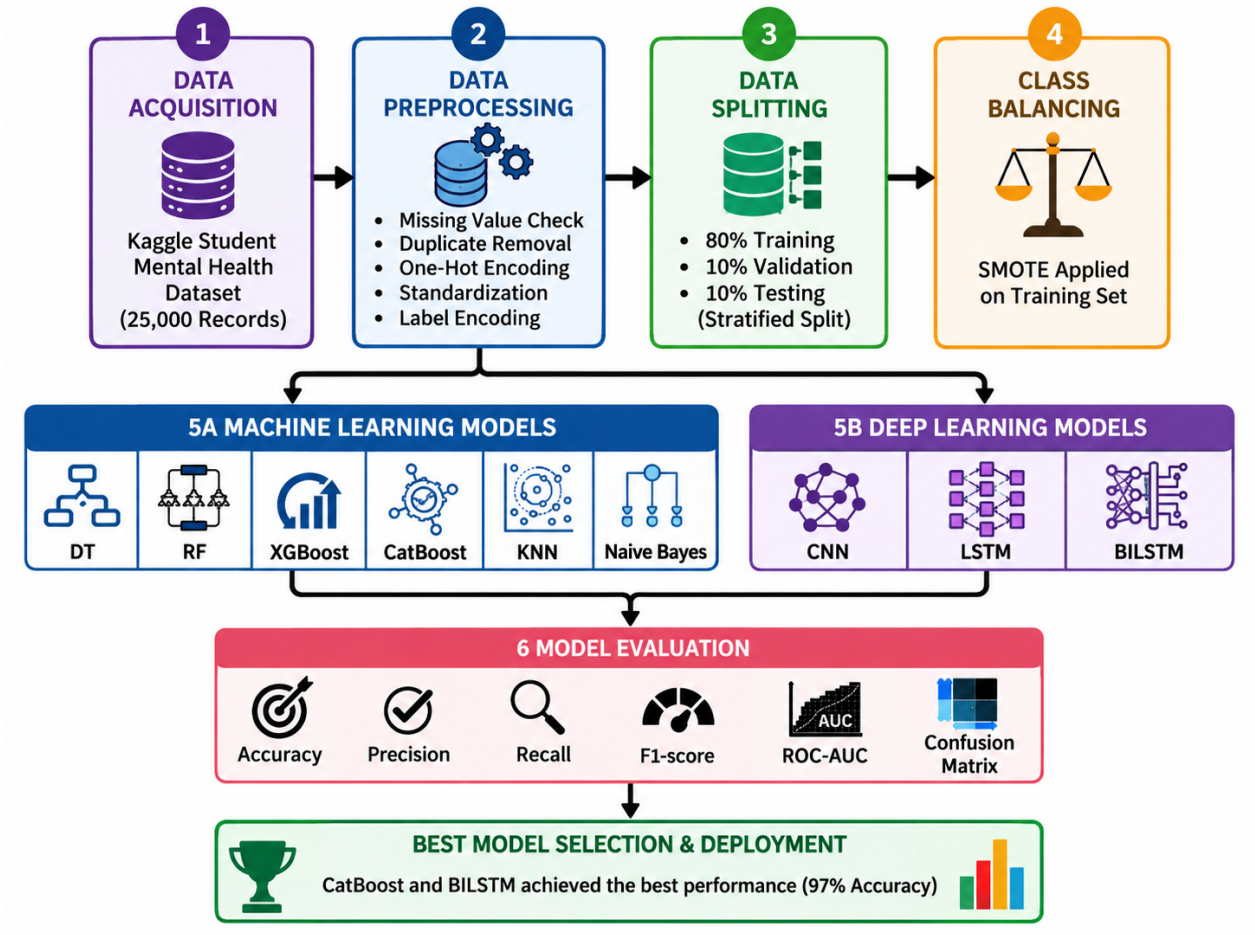

Figure 4. Overall workflow of the proposed multiclass mental health risk prediction framework.

or assessments. To assess potential target leakage, the predictor variables were treated as independent input features and mental_health_risk was retained exclusively as the target variable. No direct copy, transformation, or derived version of the target variable was included among the predictors. However, because variables such as anxiety, depression, stress, panic attack history, previous mental health diagnosis, and therapy history are conceptually related to mental health risk, their use may introduce conceptual overlap with the target. The available dataset documentation does not provide sufficient information to determine whether these variables were used to construct the original mental_health_risk labels. Therefore, the possibility of label relatedness cannot be completely excluded and is acknowledged as a limitation of the secondary dataset.

The dataset features are numeric, binary, or categorical, including age, sleep hours, physical activity, screen time, social support score, work and academic pressure levels, job satisfaction score, financial stress, working hours per week, anxiety, depression, stress, mood swings, difficulty in concentration, panic attacks, family history of mental illness, previous mental health diagnosis, therapy history, and substance use. All feature categories in the dataset are shown in Table 3.

Ethical and Privacy Considerations: This study used a publicly available secondary dataset accessed through Kaggle and did not involve direct recruitment, interaction with participants, or primary collection of personal data by the authors. Accordingly, the study did not involve direct access to identifiable participant records. Nevertheless, mental health-related information is sensitive. Real-world use should apply restricted access, secure storage, and data minimization to protect student privacy.

The available dataset documentation does not provide sufficient information to verify whether informed consent was obtained from the original participants or covered secondary research use. This limitation is therefore acknowledged. The authors did not control the original data collection, sampling, labeling, or governance procedures. Future deployment should include controlled access, secure storage, documented provenance, and appropriate data stewardship. The

Table 2. Dataset provenance and available sampling information.

Item	Information
Dataset source	Kaggle
Dataset title	Mental Health Risk Prediction Dataset (Kaggle URL slug: mental-health-disorder)
Kaggle uploader	guriya79
Dataset size	25,000 records
Variables	24 predictors + 1 target
Population	Student population as represented by the dataset; original sampling frame not documented in the available source
Data collection	Secondary dataset; original collection procedure not documented in the available source
Recruitment procedure	Not reported in the available dataset documentation
Geographic coverage	Not reported in the available dataset documentation
Participating institutions	Not reported in the available dataset documentation
Sampling method	Not reported in the available dataset documentation
Inclusion/exclusion criteria	Not reported in the available dataset documentation
Response rate	Not reported in the available dataset documentation
Target variable	mental_health_risk
Risk classes	0 = Low, 1 = Moderate, 2 = High

ethical status of the original data collection remains dependent on the procedures followed by the original dataset creators. The present study therefore focuses on computational analysis of the publicly available dataset and does not claim primary human-subject data collection or clinical validation.

Figure 5 categorizes the features used in the proposed prediction framework.

Table 4 presents the characteristics of the Mental Health Risk Prediction Dataset. The dataset was examined for missing values, duplicate records, variable types, feature ranges, and target-class distribution. No missing values or duplicate records were identified. The target variable showed class imbalance, with 9,357 records in class 0, 11,823 records in class 1, and 3,820 records in class 2. Because class 2 contained substantially fewer observations than the other two classes, class balancing was applied to reduce majority-class dominance during model training and improve representation of the high-risk

Table 3. Feature categories in the dataset.

Category	Features
Demographic Information	Age, Gender, Marital Status, Education Level, Employment Status
Lifestyle Factors	Sleep Duration, Physical Activity, Screen Time, Social Support Score
Academic and Stress Indicators	Academic Pressure, Work Stress Level, Job Satisfaction, Financial Stress, Weekly Working Hours
Psychological Indicators	Anxiety Score, Depression Score, Overall Stress Level, Mood Swings Frequency, Concentration Difficulty, Panic Attack History
Medical and Family History	Family History of Mental Illness, Previous Diagnosis, Therapy History, Substance Use
Target Class	Mental Health Risk (Low, Moderate, High)

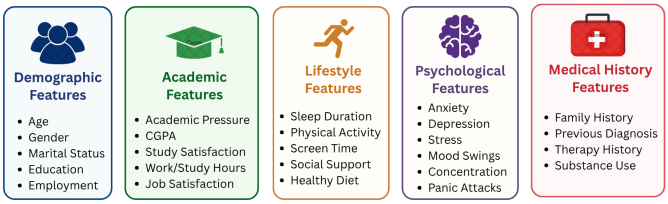

Figure 5. Categories of input features used for mental health prediction.

category.

The preprocessing pipeline adopted in this study is presented in Figure 6.

3.3 Data Preprocessing

Data preprocessing was performed to ensure consistency and compatibility of the dataset with the machine learning and deep learning models. First, categorical variables were converted into numerical representations using one-hot encoding. This transformation allowed categorical attributes to be processed without imposing an artificial ordinal relationship between categories. Numerical features were standardized using the StandardScaler transformation, which centers each feature around zero and scales it according to its standard deviation:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{1}$$

where x_{ij} is the value of numerical feature j for sample i, μ_j and σ_j are the mean and standard deviation of feature j computed on the training subset only, and z_{ij} is the standardized value. Standardization was

Table 4. Description of the mental health Risk prediction dataset.

Category	Features
Dataset Source	Kaggle Mental Health Risk Prediction Dataset
Dataset Size	25,000 Records
Problem Type	Multiclass Classification
Target Variable	Mental Health Risk
Class Labels	0 = Low Risk (9,357), 1 = Moderate Risk (11,823), 2 = High Risk (3,820)
Number of Features	24 predictor variables
Data Type	Structured Tabular Data
Application Domain	Student Mental Health Risk Prediction

applied to place numerical variables on comparable scales and to prevent features with larger numerical ranges from disproportionately influencing models that are sensitive to feature magnitude, particularly K-Nearest Neighbors and neural network models. The parameters required for standardization were estimated from the training subset only and then used to transform the validation and test subsets. The target variable was encoded into three class labels corresponding to low, moderate, and high mental health risk. For the deep learning models, the target labels were subsequently represented using one-hot encoding to support the three-unit softmax output layer. The dataset was first divided into training, validation, and testing subsets using stratified sampling. The preprocessing transformations were fitted using the training subset only and subsequently applied to the validation and test subsets. SMOTE was applied only to the training subset after categorical variables had been transformed into numerical representations. This allowed SMOTE to operate on the numerical feature space required by the implemented pipeline while ensuring that synthetic samples were generated exclusively from training data. The validation and test subsets were not oversampled. However, because one-hot encoded categorical variables are included in the feature space, SMOTE interpolation may produce synthetic values between encoded categories. This limitation is acknowledged, and the resulting performance should therefore be interpreted with caution.

3.4 Data Splitting and Class Balancing

To preserve the original class proportions across the training, validation, and testing subsets, the dataset was divided using stratified sampling. An

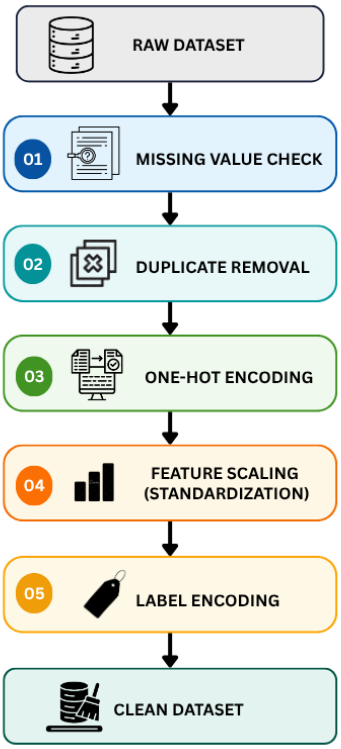

Figure 6. Data preprocessing pipeline.

80:10:10 split was used; 80% of the data was used for training the model, 10% for validating it, and 10% for final model evaluation. With the original dataset comprising 25,000 records, the training subset contained 20,000 records, the validation subset contained 2,500 records, and the testing subset contained 2,500 records. The validation subset was used for hyperparameter tuning, model selection, and implementation of early stopping during the training of deep learning models. As a preparation for the model-building process, the training subset was subjected to SMOTE to address potential imbalances in the target variable. To accomplish that, SMOTE added minority-class samples by interpolating between existing minority-class samples in the feature space [18]. Each synthetic sample was generated as:

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda\,(\mathbf{x}_{nn} - \mathbf{x}_i), \qquad \lambda \sim U(0,1), \tag{2}$$

where $\mathbf{x}_i$ is a minority-class training sample, $\mathbf{x}_{nn}$ is one of its k nearest minority-class neighbours, and λ is a random number drawn uniformly from $[0,1]$. This procedure was intended to reduce majority-class dominance during training and provide more balanced decision boundaries across the three risk categories. The order of applying SMOTE after one-hot encoding was retained as part of the implemented pipeline; its limitation is described in Section 3.3.

Alternative class balancing approaches, including SMOTENC and class weighting, were not evaluated in the present study. Therefore, the effectiveness of the selected SMOTE strategy relative to categorical-aware or cost-sensitive balancing methods cannot be established. This is acknowledged as a methodological limitation and represents a direction for future comparative evaluation.

3.5 Model Development

The model set was selected to represent complementary learning paradigms and provide a controlled comparison between conventional machine learning and deep learning approaches for structured multiclass data. The selected conventional models include tree-based, ensemble, distance-based, and probabilistic methods, while the deep learning models represent convolutional and recurrent neural architectures. Six machine learning models were designed and built for this purpose. Decision Tree was used as an interpretable baseline model, as it can capture nonlinear decision rules and feature interactions. Random Forest was considered as an ensemble model that can reduce overfitting and provide greater generalization. XGBoost [34] was included as a gradient boosting method that can effectively model complex relationships in structured, tabular data. CatBoost [35] was included as a gradient boosting approach known for its strong performance on structured tabular classification tasks and its ability to model complex nonlinear relationships while incorporating regularization mechanisms that can reduce overfitting. K-Nearest Neighbors was included as a distance-based classifier, while Naïve Bayes provided a probabilistic baseline based on conditional feature distributions. Optimization of hyperparameters was done using the validation dataset. The tuning of parameters was done for tree-based models by varying the maximum tree depth, the number of estimators, the learning rate, and the minimum number of samples per split. For KNN, we tuned the number of neighbors and the distance metric. For Naïve Bayes, we tuned the smoothing parameters. The macro F1-score on the validation set was used as the primary criterion for selecting the final model configuration, with other validation metrics considered as supporting performance measures.

The candidate configurations were evaluated systematically on the validation subset, and the configuration achieving the highest validation macro F1-score was selected for each model. The search space included parameters controlling model complexity, ensemble size, learning rate, neighborhood structure, and probabilistic smoothing. The hyperparameter configurations considered for each model are summarized in Table 5.

The selected configurations were subsequently evaluated on the independent test subset, which remained unseen during model selection.

Three deep learning models were trained to compare neural-network-based learning with the conventional machine learning models. One of the models employs a one-dimensional Convolutional Neural Network (CNN) for learning local patterns in transformed feature vectors. The CNN architecture consists of an input layer, one-dimensional convolutional layers, activation functions, pooling layers, dropout regularization, fully connected layers, and a final output layer containing softmax for three-class classification. The LSTM model examines classification performance with sequential feature representation. As the dataset was tabular and not temporal, the transformed feature vector is reshaped to a feature sequence and then fed to the LSTM layer. The LSTM model also has an input layer, LSTM units, dropout layers, fully connected layers, and a softmax output layer. The Bidirectional LSTM model adds to this previous architecture and processes a feature sequence in both the forward and backward directions. This configuration allowed the model to process the transformed feature sequence in both directions and provided a bidirectional representation of the input feature sequence. All deep learning models were trained using categorical cross-entropy loss and the Adam optimizer. Figure 7 provides a visual representation of the deep learning architecture used in the proposed prediction framework.

To provide a more precise description of the implemented deep learning models, Table 6 summarizes their principal layer configurations, activation functions, dropout settings, and output structure.

The CNN, LSTM, and BiLSTM models used the same optimization settings to ensure a consistent comparison. The CNN employed two one-dimensional convolutional layers followed by max pooling and a dense layer, whereas the LSTM and BiLSTM models used recurrent layers followed by dense layers. All models produced three class probabilities through a softmax output layer.

Table 5. Hyperparameter search space for conventional machine learning models.

Model	Hyperparameters Tuned	Search Values / Range
Decision Tree	Maximum depth, minimum samples split	max_depth = [5, 10, 15, 20, None]; min_samples_split = [2, 5, 10]
Random Forest	Number of estimators, maximum depth, minimum samples split	n_estimators = [100, 200, 300]; max_depth = [10, 20, 30, None]; min_samples_split = [2, 5, 10]
XGBoost	Number of estimators, learning rate, maximum depth	n_estimators = [100, 200, 300]; learning_rate = [0.01, 0.05, 0.1]; max_depth = [3, 5, 7]
CatBoost	Number of iterations, learning rate, depth	iterations = [100, 200, 300]; learning_rate = [0.01, 0.05, 0.1]; depth = [4, 6, 8]
KNN	Number of neighbors, distance metric	n_neighbors = [3, 5, 7, 9, 11]; metric = [euclidean, manhattan]
Naïve Bayes	Smoothing parameter	var_smoothing = [1e-11, 1e-10, 1e-9, 1e-8, 1e-7]

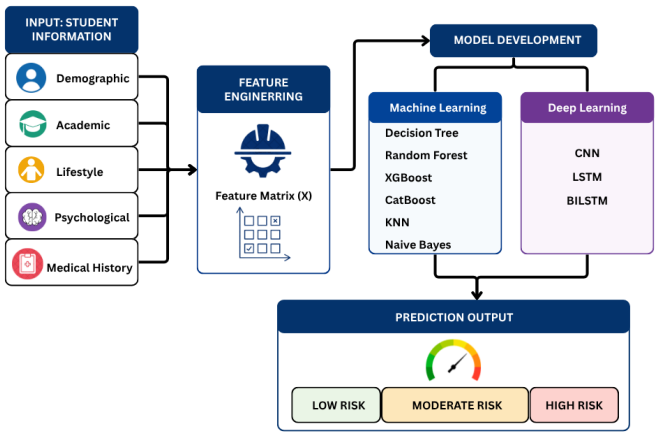

Figure 7. Architecture of the proposed multiclass prediction framework.

No systematic hyperparameter search was performed for the deep learning models. The architectures and training parameters were selected prior to final test evaluation. The selected configuration used a learning rate of 0.001, a batch size of 32, a maximum of 50 epochs, a dropout rate of 0.30, and an early stopping patience of 5 epochs. Confidence intervals were also not calculated because the experiments were evaluated using a single stratified train, validation, and independent test split rather than repeated stratified runs. Calibration measures were also not evaluated in the present study; therefore, the reported results assess classification discrimination rather than the reliability of predicted probabilities. Moreover, balanced accuracy was not reported because the evaluation focused on accuracy, class-wise precision, recall, F1-score, and macro F1-score.

Software environment: The experiments were implemented using Python 3.10, TensorFlow/Keras 2.15, Scikit-learn 1.4, XGBoost 2.0, CatBoost 1.2, and imbalanced-learn 0.12. The experiments were executed in Google Colab using a GPU-enabled runtime.

Code availability: The complete source code and processed data are not publicly available at present. The preprocessing procedure, model architectures, training configuration, and evaluation protocol are reported in sufficient detail to document the experimental procedure. The code may be made available upon reasonable request, subject to the terms governing redistribution of the secondary dataset.

3.6 Model Evaluation Metrics

Multiple metrics were used to compare the classifiers systematically and to gauge their overall performance. For each class c, let TP_c, FP_c and FN_c denote the numbers of true positives, false positives and false negatives when class c is treated as the positive class (one-vs-rest), let N be the number of test samples, n_c the number of test samples of class c, and $C = 3$ the number of classes. Accuracy is the proportion of correctly classified samples:

$$\text{Accuracy} = \frac{1}{N}\sum_{c=1}^{C} TP_c. \tag{3}$$

Precision is the proportion of samples predicted as class c that truly belong to class c:

$$P_c = \frac{TP_c}{TP_c + FP_c}. \tag{4}$$

Recall is the proportion of samples of class c that are correctly identified:

$$R_c = \frac{TP_c}{TP_c + FN_c}. \tag{5}$$

The F1-score is the harmonic mean of precision and recall:

$$F1_c = \frac{2\,P_c R_c}{P_c + R_c}. \tag{6}$$

Table 6. Deep learning model architectures and training configuration.

Parameter	CNN	LSTM	BiLSTM
Layer configuration	Input → Conv1D (64 filters, kernel size 3) → MaxPooling1D → Conv1D (128 filters, kernel size 3) → Flatten → Dense (128) → Dropout	Input → LSTM (128 units) → Dropout → Dense (64) → Dropout	Input → Bidirectional LSTM (128 units) → Dropout → Dense (64) → Dropout
Activation	ReLU (hidden layers)	tanh and sigmoid within the LSTM cell; ReLU in the dense layer	tanh and sigmoid within the BiLSTM cell; ReLU in the dense layer
Dropout rate	0.3	0.3	0.3
Output layer	Dense(3), Softmax	Dense(3), Softmax	Dense(3), Softmax
Optimizer	Adam	Adam	Adam
Loss function	Categorical cross entropy	Categorical cross entropy	Categorical cross entropy
Learning rate	0.001	0.001	0.001
Batch size	32	32	32
Maximum epochs	50	50	50
Early stopping monitor	Validation loss	Validation loss	Validation loss
Early stopping patience	5 epochs	5 epochs	5 epochs
Restore best weights	Yes	Yes	Yes
Random seed	42	42	42

The macro-averaged F1-score, used as the primary comparative metric because it assigns equal importance to each class and is therefore less dominated by the majority class, is

$$F1_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C} F1_c, \tag{7}$$

and the weighted F1-score, which weights each class by its support, is

$$F1_{\text{weighted}} = \sum_{c=1}^{C} \frac{n_c}{N} F1_c. \tag{8}$$

Confusion matrices were generated to analyze class-wise prediction behavior and to identify which risk categories were most frequently misclassified. This helped to analyze the discrimination power of each model for the different mental health risk categories. Figure 8 summarizes the conventional machine learning and deep learning approaches evaluated in this study.

Since the experimental evaluation was conducted using a fixed independent test set, statistical significance testing based on repeated independent model evaluations was not performed. Instead, model differences were assessed using multiple complementary performance metrics, including accuracy, precision, recall, macro F1-score, weighted F1-score, and confusion matrices. The absence of repeated cross-validation or repeated independent test evaluations is acknowledged as a limitation and represents an opportunity for further validation.

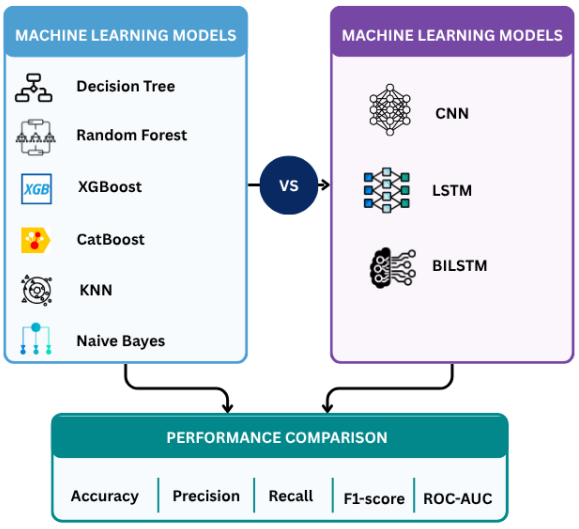

Figure 8. Comparative framework of machine learning and deep learning models.

4 Results and Discussion

This section presents and critically discusses the experimental results of the multiclass mental health risk prediction framework. The evaluated models were compared using accuracy, precision, recall, macro F1-score, weighted F1-score, and confusion matrices. Particular emphasis is placed on class-wise performance because the original dataset contained unequal class distributions. The analysis examines predictive performance, feature importance, differences between conventional machine learning and deep learning models, and the factors contributing to classification errors across the low, moderate, and high risk categories.

4.1 Experimental Setup

Table 7 provides the experimental configuration for the machine learning and deep learning models used for mental health risk prediction.

Table 7. Experimental setup summary for proposed model.

Component	Configuration
Problem Type	Multiclass classification (Mental Health Risk: Low, Moderate, High)
Dataset	Kaggle Mental Health Risk Prediction Dataset
Samples	25,000 records
Features	24 predictors + 1 target variable
Feature Groups	Demographic, Academic, Behavioral, Psychological, Lifestyle, Medical
Preprocessing	One-hot encoding, standardization
Class Balancing	SMOTE applied only on training set
Data Split	80% Train / 10% Validation / 10% Test (Stratified)
Machine Learning Models	Decision Tree, Random Forest, XGBoost, CatBoost, KNN, Naïve Bayes
Deep Learning Models	CNN (1D Conv), LSTM, BiLSTM
DL Architecture	Conv1D / LSTM layers + Dense layers + Dropout + Softmax
Loss Function	Categorical Cross-Entropy
Optimizer	Adam
Hyperparameter Tuning	Validation-based tuning (conventional machine learning models only; no systematic search for the deep learning models)
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, Confusion Matrix
Tools & Frameworks	Python, TensorFlow/Keras, Scikit-learn, XGBoost, Imbalanced-learn
Hardware	HP Laptop, Intel Core i5 8th Generation, 16 GB RAM, 512 GB SSD, (Google Colab) (GPU supported)

4.2 Performance of the Conventional Machine Learning Models

The Decision Tree feature importance analysis identified sleep hours as the most influential predictor, followed by panic attack history, anxiety score, and depression score. Financial stress, work stress, social support, physical activity, and family history of mental illness also contributed to the prediction. These findings indicate that mental health risk classification was influenced by a combination of psychological, lifestyle, social, and stress-related factors rather than by a single predictor. Figure 9 presents the feature importance and confusion matrix of the Decision Tree model.

The Decision Tree showed stronger recognition of the low and high risk classes than of the moderate risk class. Its Class 1 recall was 0.86, compared with 1.00 for Class 0 and 0.96 for Class 2. The lower recall for Class 1 indicates that some moderate risk cases were assigned to the high risk category, reflecting overlap between the two risk groups.

Table 8 shows that the Decision Tree achieved an accuracy of 0.93 and a macro F1-score of 0.91. Although performance for Class 0 was very strong, the lower Class 1 recall and Class 2 precision indicate that the model produced more uncertainty around the boundary between moderate and high risk. The macro F1-score therefore provides a more informative summary than accuracy alone because it gives equal importance to all three classes. Figure 10 presents the Naïve Bayes classification results. In contrast to the stronger ensemble models, Naïve Bayes showed substantially greater class overlap and weaker class-wise discrimination.

Naïve Bayes showed the weakest class-wise discrimination among the principal models considered in the detailed analysis. Class 2 was particularly difficult to identify, with a recall of 0.58 and an F1-score of 0.49. Class 1 also showed limited discrimination, with both precision and recall of 0.67. The greater overlap between the moderate and high risk categories suggests that the probabilistic assumptions of Naïve Bayes were insufficient to represent the complex relationships among the psychological, behavioral, lifestyle, and stress-related predictors.

Table 9 confirms the comparatively weak performance

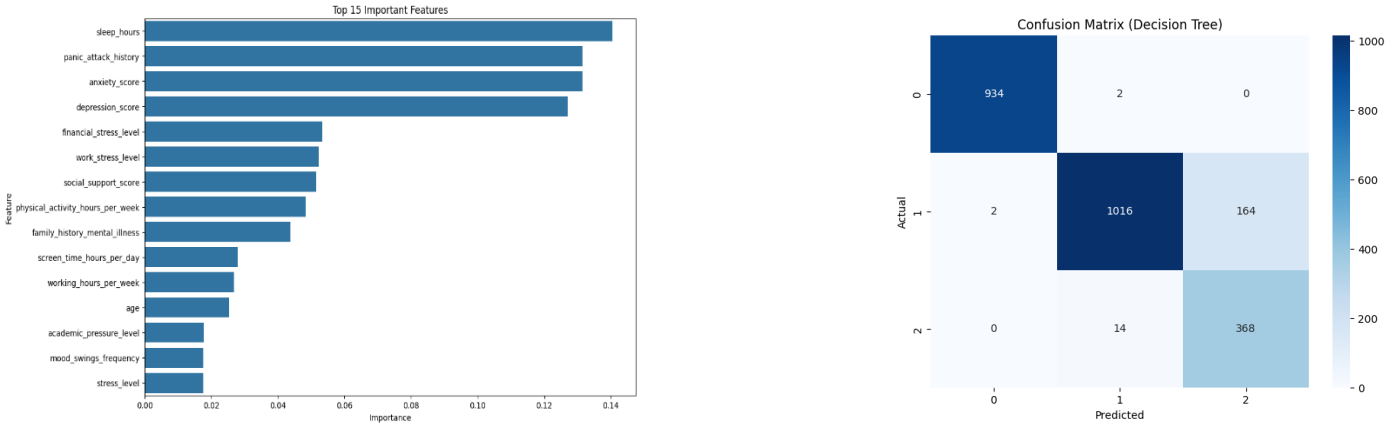

Figure 9. Feature importance and confusion matrix of the Decision Tree model.

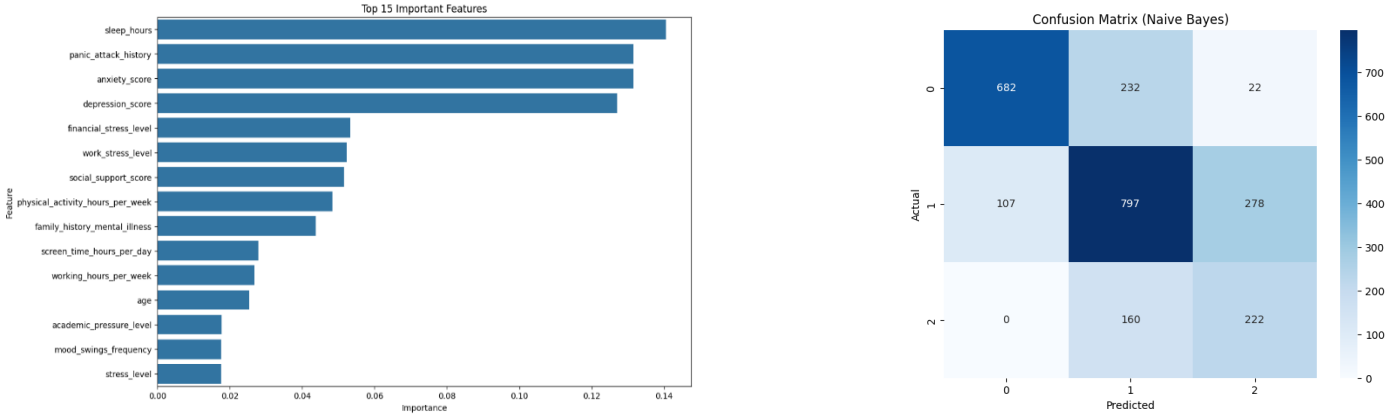

Figure 10. Feature importance and confusion matrix of the Naïve Bayes model.

Table 8. Decision Tree model performance for multiclass mental health risk classification.

Class / Metric	Precision	Recall	F1-score	Support
Class 0	1.00	1.00	1.00	936
Class 1	0.98	0.86	0.92	1182
Class 2	0.69	0.96	0.81	382
Accuracy	—	—	0.93	2500
Macro Average	0.89	0.94	0.91	2500
Weighted Average	0.94	0.93	0.93	2500

Table 9. Classification report of the Naïve Bayes model.

Class / Metric	Precision	Recall	F1-score	Support
Class 0	0.86	0.73	0.79	936
Class 1	0.67	0.67	0.67	1182
Class 2	0.43	0.58	0.49	382
Accuracy	—	—	0.68	2500
Macro Average	0.65	0.66	0.65	2500
Weighted Average	0.71	0.68	0.69	2500

of Naïve Bayes, with an accuracy of 0.68 and macro F1-score of 0.65. The Class 2 F1-score of 0.49 was substantially lower than the Class 0 F1-score of 0.79, demonstrating that accuracy alone would obscure the difficulty of identifying the high risk category. The results suggest that the conditional independence assumptions underlying the model were not sufficient to capture the interactions among the predictors in this multiclass setting. CatBoost identified depression score, anxiety score, panic attack history, and sleep hours as the most influential predictors of mental health risk classification. Social support, work stress, financial stress, family history of mental illness, and physical activity also contributed to prediction. Figure 11 presents the feature importance and confusion matrix of the CatBoost model.

CatBoost showed strong class-wise discrimination, particularly for the high risk category. It correctly classified 373 of 382 Class 2 cases, while the remaining 9 cases were classified as Class 1. For Class 1, 70 of 1182 cases were misclassified, with most of these errors assigned to Class 2. The concentration of errors between the moderate and high risk categories indicates that these two classes have greater overlap than the low risk category. Nevertheless, the total number of classification errors remained low.

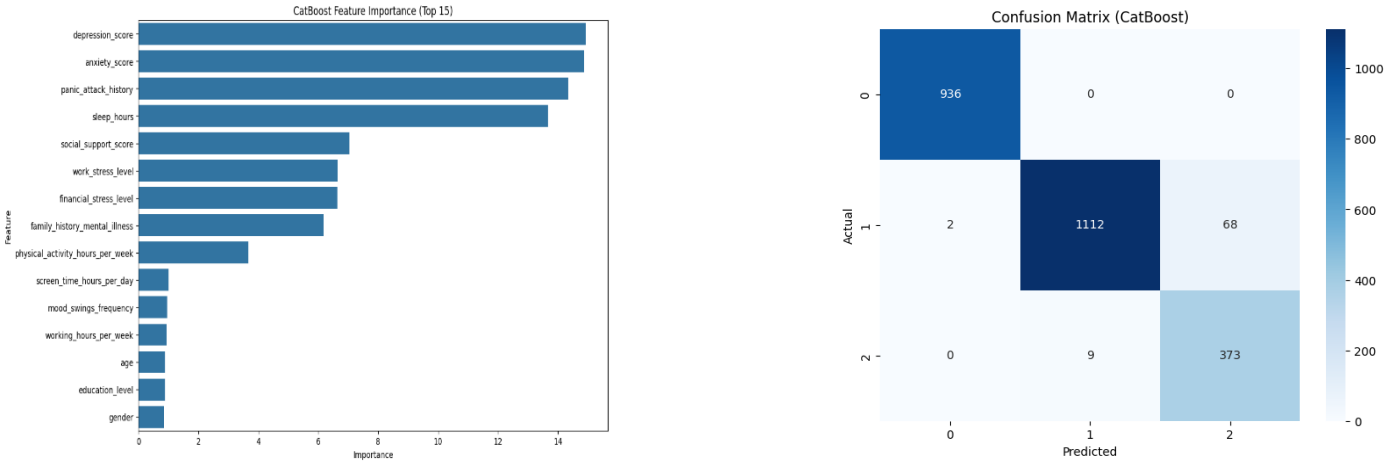

Figure 11. Feature importance and confusion matrix of the CatBoost model.

Table 10. Classification report of the CatBoost model.

Class / Metric	Precision	Recall	F1-score	Support
Class 0	1.00	1.00	1.00	936
Class 1	0.99	0.94	0.97	1182
Class 2	0.85	0.98	0.91	382
Accuracy	—	—	0.97	2500
Macro Average	0.95	0.97	0.96	2500
Weighted Average	0.97	0.97	0.97	2500

Table 11. Classification report of the Random Forest model.

Class / Metric	Precision	Recall	F1-score	Support
Class 0	0.99	1.00	0.99	936
Class 1	0.96	0.92	0.94	1182
Class 2	0.81	0.90	0.85	382
Accuracy	—	—	0.95	2500
Macro Average	0.92	0.94	0.93	2500
Weighted Average	0.95	0.95	0.95	2500

Table 10 confirms the strong and relatively balanced performance of CatBoost, with an accuracy of 0.97 and macro F1-score of 0.96. Class 2 achieved a recall of 0.98, although its precision was lower at 0.85, indicating that some moderate risk cases were classified as high risk. The macro F1-score of 0.96 demonstrates that the strong overall performance was not driven solely by the majority class.

Figure 12 presents the Random Forest feature importance and confusion matrix. The feature importance pattern was broadly consistent with the other tree-based models, with sleep, psychological symptoms, stress indicators, social support, physical activity, and family history contributing to classification.

Random Forest showed strong overall class-wise performance, although errors remained concentrated around the moderate and high risk categories. Its Class 2 precision of 0.81 was lower than its Class 2 recall of 0.90, indicating that some moderate risk cases were assigned to the high risk category.

Table 11 shows that Random Forest achieved an accuracy of 0.95 and a macro F1-score of 0.93. The relatively small difference between weighted F1-score (0.95) and macro F1-score (0.93) indicates that performance remained comparatively consistent across the three classes.

Figure 13 indicates that sleep hours were among the most influential predictors in the KNN classification analysis. Panic attacks, anxiety, and depression scores were also among the more influential predictors. Financial stress, work stress, social support, physical activity, and family history of mental illness also contributed to the classification, indicating that mental health risk was associated with a combination of psychological, behavioral, lifestyle, and social factors.

KNN showed substantially greater class overlap than the ensemble models. Class 1 was particularly difficult to classify, with only 503 of 1182 cases correctly identified. The model also produced substantial confusion between Class 0 and Class 1 and between Class 1 and Class 2. This pattern indicates that distance-based classification was less effective for the nonlinear and overlapping class boundaries present in the dataset. Table 12 confirms that KNN was one of the weakest models, with an accuracy of 0.60 and macro F1-score of 0.59. Class 1 had the lowest recall at 0.43, indicating substantial difficulty in identifying moderate risk cases. The results suggest that distance-based classification was less suitable for representing the complex interactions and overlapping

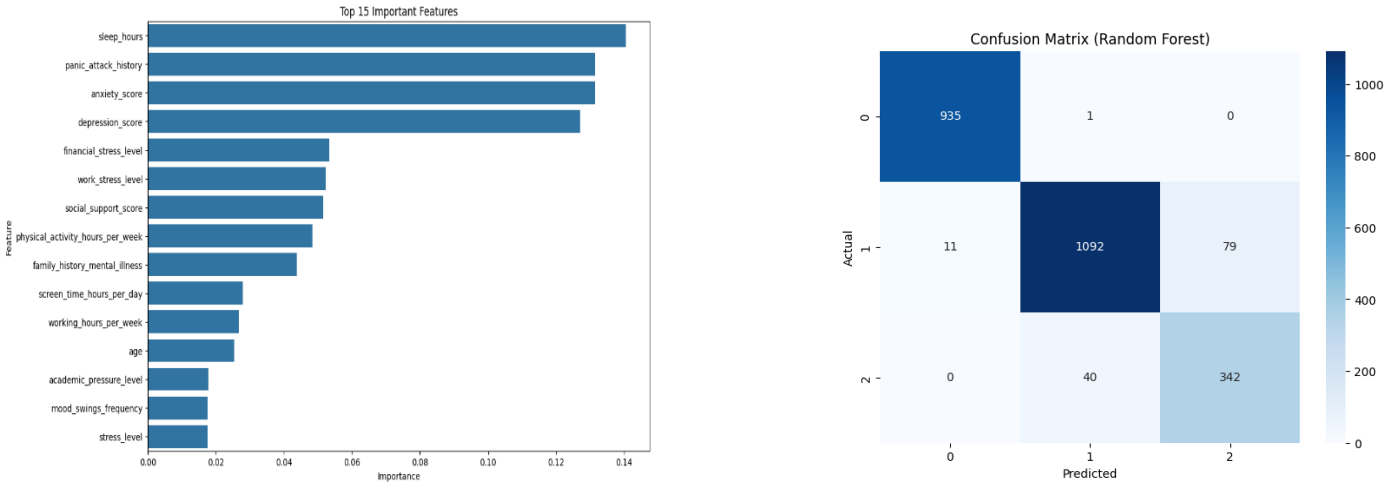

Figure 12. Feature importance and confusion matrix of the Random Forest model.

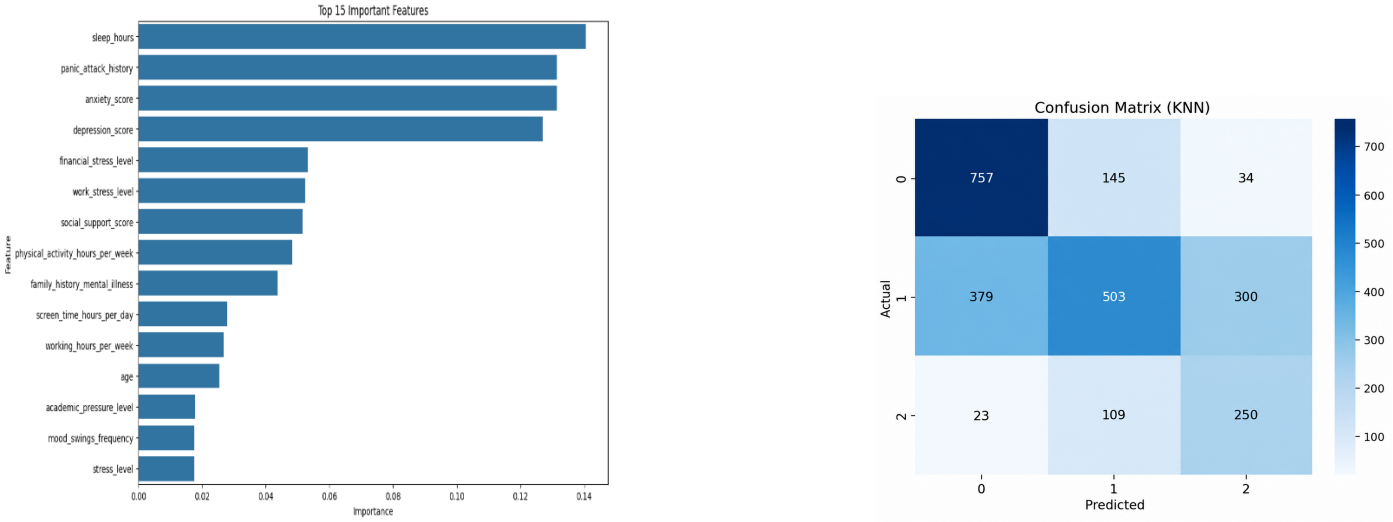

Figure 13. Feature importance and confusion matrix of KNN model.

Table 12. Classification report of the KNN model.

Class / Metric	Precision	Recall	F1-score	Support
Class 0	0.65	0.81	0.72	936
Class 1	0.66	0.43	0.52	1182
Class 2	0.43	0.65	0.52	382
Accuracy	—	—	0.60	2500
Macro Average	0.58	0.63	0.59	2500
Weighted Average	0.62	0.60	0.59	2500

boundaries among the three risk categories.

Figure 14 presents the XGBoost feature importance and confusion matrix. Sleep hours, panic attack history, anxiety score, and depression score were among the most influential predictors, while financial stress, work stress, social support, physical activity, and family history of mental illness also contributed to classification. The similarity of this feature pattern with the other strong tree-based models supports the consistency of these predictors across the analysis.

XGBoost showed strong class-wise discrimination, with most errors occurring between the moderate and high risk categories. Class 0 was particularly well separated, while the lower precision observed for Class 2 indicates that some moderate risk cases were classified as high risk.

Table 13 shows that XGBoost achieved an accuracy of 0.96 and macro F1-score of 0.94. The Class 2 F1-score of 0.86 was lower than the corresponding scores for Classes 0 and 1, again indicating greater difficulty at the moderate and high risk boundary. Nevertheless, the relatively high macro F1-score demonstrates strong performance across all three classes.

4.3 Performance of the Deep Learning Models

Three deep learning architectures were evaluated using the same training, validation, and independent

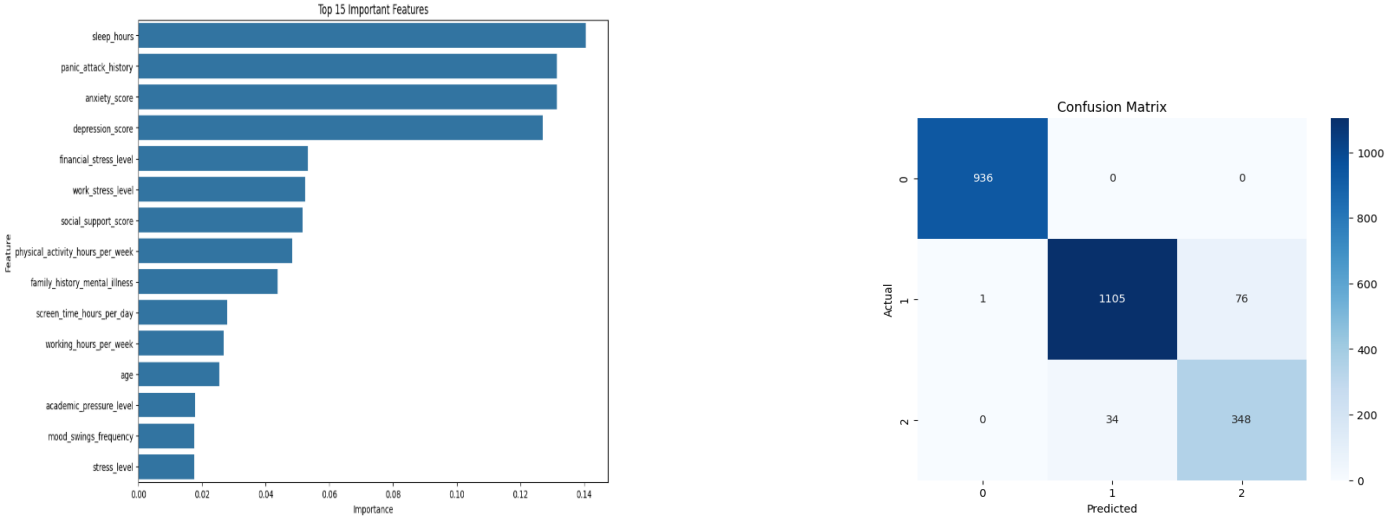

Figure 14. Feature importance and confusion matrix of the XGBoost model.

Table 13. Classification report of the XGBoost model.

Class / Metric	Precision	Recall	F1-score	Support
Class 0	1.00	1.00	1.00	936
Class 1	0.97	0.93	0.95	1182
Class 2	0.82	0.91	0.86	382
Accuracy	—	—	0.96	2500
Macro Average	0.93	0.95	0.94	2500
Weighted Average	0.96	0.96	0.96	2500

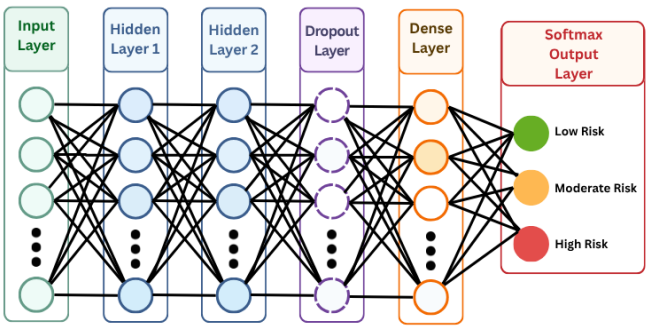

Figure 15. Generalized architecture of the deep learning models.

test partitions. Their results are considered alongside the conventional machine learning models to determine whether increased architectural complexity provides a measurable advantage for the structured tabular prediction task. The architecture of the deep learning models used in this study is illustrated in Figure 15.

The CNN achieved stable training and validation performance, with the accuracy curves converging during training. The corresponding loss curves also showed stabilization, indicating no major divergence between training and validation performance. The training and validation accuracy along with the model loss performance of the final CNN model is presented in Figure 16.

The confusion matrix of the final CNN model, showing classification performance across three classes, is presented in Figure 17. The model correctly classified most samples, with 893 Class 0, 1,040 Class 1, and 343 Class 2 samples correctly identified. Most misclassifications occurred between Class 1 and Class 2, indicating some overlap between these risk categories.

The CNN achieved an accuracy of 0.91 and correctly classified 2276 of 2500 test samples. Class 0 showed the strongest performance, while Class 2 had lower precision, indicating greater overlap between the high risk category and the other classes. The macro F1-score of 0.90 indicates reasonably balanced performance across the three classes.

Table 14 shows that the CNN achieved a weighted precision, recall, and F1-score of 0.91, with a macro F1-score of 0.90. The lower Class 2 precision of 0.80 indicates greater difficulty in distinguishing high risk cases from the other categories. The macro F1-score suggests reasonably balanced performance across the three classes, although the CNN remained below the stronger tree-based ensemble models.

The LSTM training and validation curves showed strong learning performance with a moderate difference between training and validation accuracy. The loss curves remained relatively stable, indicating acceptable generalization to the validation data. The training and validation accuracy of the model along with the loss performance of the final LSTM model is

Table 14. Performance evaluation of the final CNN.

Metric	Class 0	Class 1	Class 2	Overall / Average
Precision	0.94	0.93	0.80	0.91 weighted avg
Recall	0.95	0.88	0.90	0.91 weighted avg
F1-score	0.95	0.90	0.85	0.91 weighted avg
Support	936	1182	382	2500
Correct Predictions	893	1040	343	2276
Class-wise Recall	95.41%	87.99%	89.79%	91.04%

Table 15. Performance evaluation of the LSTM model.

Metric	Class 0	Class 1	Class 2	Overall / Average
Precision	0.95	0.93	0.86	0.93 weighted avg
Recall	0.96	0.92	0.90	0.93 weighted avg
F1-score	0.96	0.92	0.88	0.93 weighted avg
Support	936	1182	382	2500
Correct Predictions	897	1082	344	2323
Class-wise Accuracy / Recall	95.83%	91.54%	90.05%	92.92%

presented in Figure 18.

On the independent test set, the LSTM achieved an accuracy of 0.93 (92.92%) and a weighted F1-score of 0.93. Table 15 summarizes its performance across the three risk classes. Class 0 achieved the highest F1-score of 0.96, followed by Class 1 at 0.92 and Class 2 at 0.88. The lower Class 2 precision (0.86) indicates that some moderate risk cases were classified as high risk.

The BiLSTM model achieved the strongest deep learning performance, as shown in Figure 19. The training and validation accuracy curves remained relatively close during training, while the corresponding loss curves showed stable behavior. This pattern indicates stable learning with no substantial divergence between training and validation performance.

Table 16 summarizes the classification performance of the BiLSTM model across the three risk classes. The model achieved an accuracy of 0.97 and a macro F1-score of 0.96. Class 0 achieved the highest F1-score at 0.99, followed by Class 1 at 0.96 and Class 2 at 0.92.

Although Class 2 was the smallest class, its recall of 0.93 (Table 16) indicates that the model identified most high risk cases. The relatively lower precision of 0.91 indicates that some cases from the other categories were assigned to Class 2. Because the dataset is structured tabular data rather than temporal data, the bidirectional architecture should not be interpreted as learning temporal dependencies. Instead, the transformed feature vector was represented as a feature sequence, allowing the BiLSTM to process feature positions in both forward and backward directions. This representation may have enabled the model to capture complementary relationships within the transformed feature space, contributing to its strong classification performance. Figure 20 illustrates the prediction process of the proposed multiclass framework.

4.4 Feature Importance and Practical Implications

Table 17 summarizes the recurring predictors identified across the feature importance analyses of the evaluated models. Sleep hours, panic attack history, anxiety score, and depression score were repeatedly identified among the influential predictors. Stress related, social, lifestyle, and family history variables also appeared across the models. The consistency of these predictors suggests that the classification task reflects a multidimensional pattern of student mental health risk rather than dependence on a single variable. However, feature importance represents predictive contribution within the fitted models and should not be interpreted as evidence of causal relationships.

The consistent importance of sleep hours also has a practical implication for student support systems. Sleep-related information could be incorporated as one component of a broader screening profile used by educators or counselors to identify students who may benefit from further assessment or supportive intervention. However, feature importance represents predictive association rather than causation. Sleep hours should therefore not be used independently to label a student as being at mental health risk. Instead,

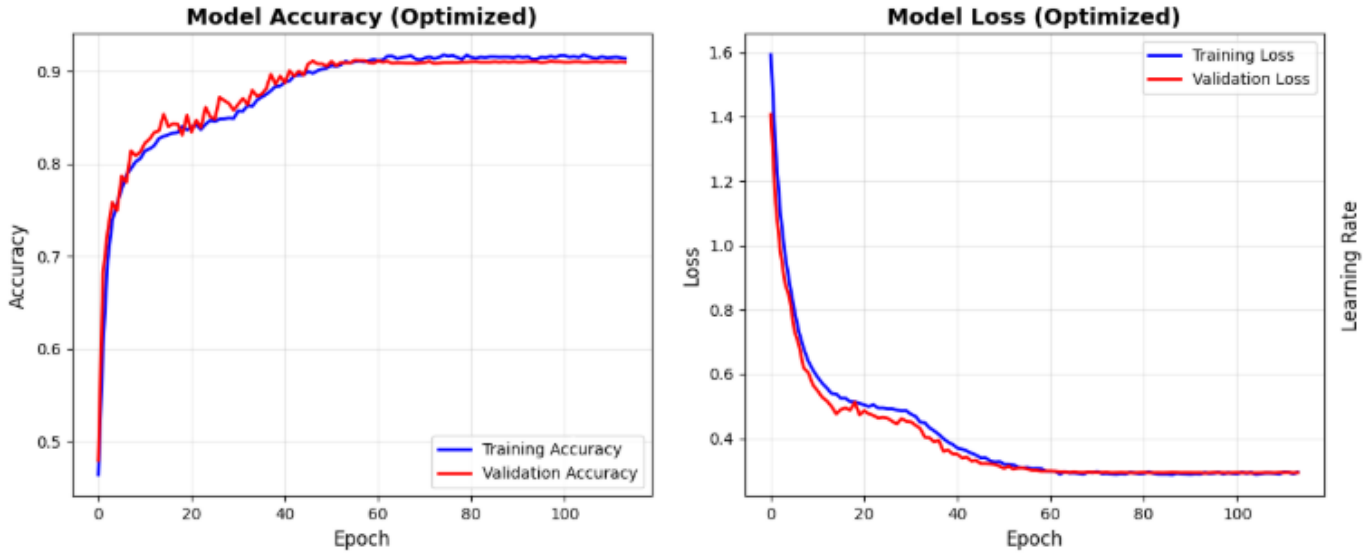

Figure 16. Training and validation accuracy and loss curves of the final CNN model.

Table 16. Classification metrics of the BiLSTM model.

Metric	Class 0	Class 1	Class 2	Overall / Average
Precision	0.98	0.97	0.91	0.97 weighted avg
Recall	0.99	0.96	0.93	0.97 weighted avg
F1-score	0.99	0.96	0.92	0.97 weighted avg
Support	936	1182	382	2500
Accuracy	—	—	—	0.97
Macro average (P / R / F1)	—	—	—	0.95 / 0.96 / 0.96

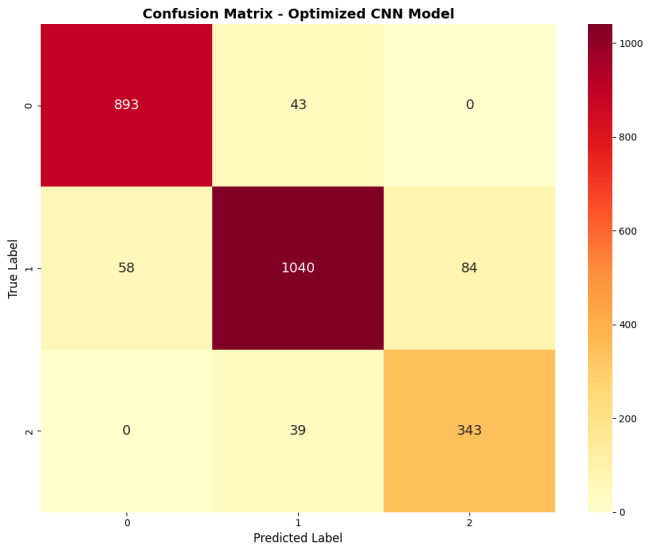

Figure 17. Confusion matrix of the final CNN model, showing classification performance across three classes.

sleep-related information should be considered together with psychological, social, academic, and lifestyle indicators within a broader decision support process.

4.5 Comparative Analysis and Contextual Comparison

Table 18 provides the consolidated performance comparison across all nine evaluated models. Macro F1-score is emphasized alongside accuracy because it gives equal importance to the three risk categories and is therefore less influenced by the majority class. CatBoost and BiLSTM achieved the highest accuracy of 0.97 and macro F1-score of 0.96, followed by XGBoost and Random Forest with macro F1-scores of 0.94 and 0.93, respectively. KNN and Naïve Bayes showed substantially lower macro F1-scores of 0.59 and 0.65.

The comparison between conventional machine learning and deep learning indicates that increased architectural complexity did not automatically produce better predictive performance. CatBoost achieved the same highest accuracy as BiLSTM at 0.97, while XGBoost and Random Forest also achieved strong macro F1-scores of 0.94 and 0.93. This pattern is consistent with the structured tabular nature of the dataset, in which tree-based ensemble methods can effectively capture nonlinear relationships and interactions among heterogeneous demographic, psychological, lifestyle, and stress-related variables.

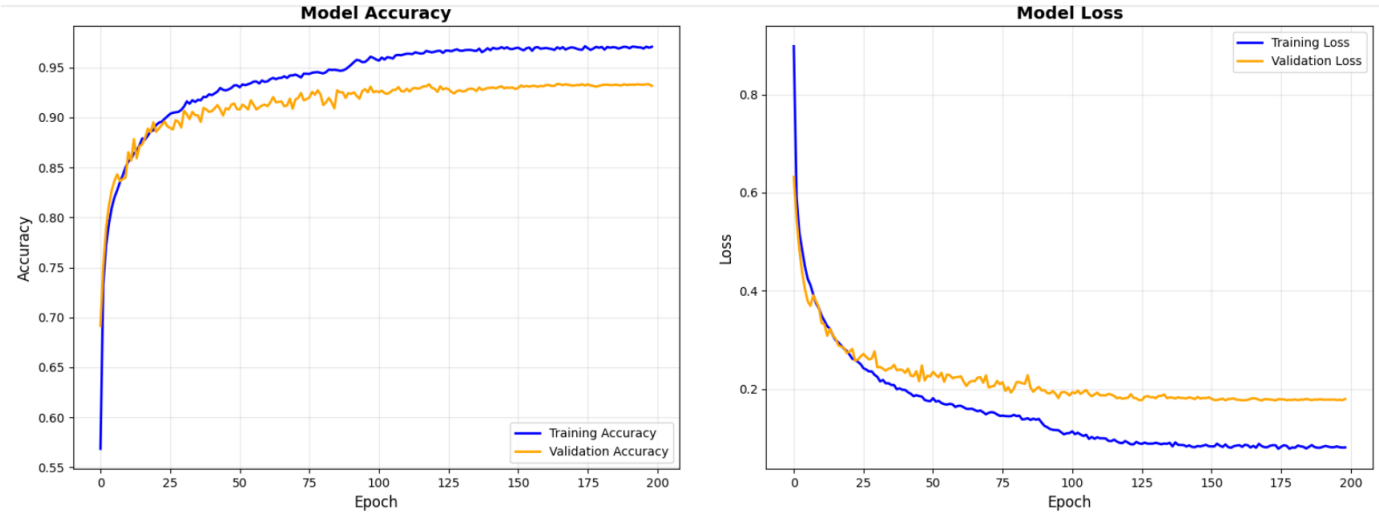

Figure 18. Training and validation accuracy and loss curves of the final LSTM model.

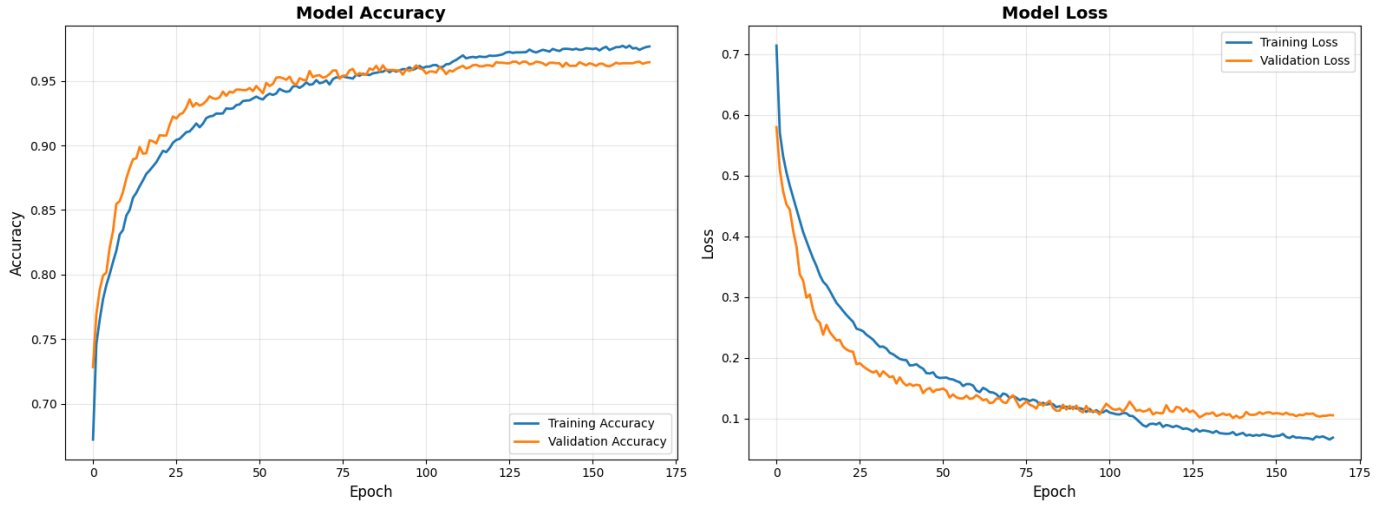

Figure 19. Training and validation accuracy and loss curves for the BiLSTM model.

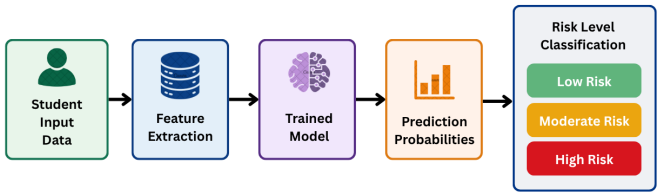

Figure 20. Prediction process of the proposed multiclass framework.

The deep learning architectures were evaluated using transformed feature sequences, but the underlying data were not temporal. Therefore, the additional architectural complexity of LSTM and BiLSTM did not provide a consistent performance advantage over optimized tree-based ensemble models.

The class-wise results show that classification difficulty was not uniform across the three risk categories. The moderate and high risk categories showed greater overlap than the low risk category across several models. This pattern was particularly evident in the weaker models, where Class 1 and Class 2 showed lower precision, recall, or F1 scores. Naïve Bayes and KNN were particularly affected, suggesting that simple probabilistic assumptions and distance-based decision boundaries were insufficient to represent the nonlinear relationships among the predictors. In contrast, CatBoost and XGBoost achieved stronger macro F1-scores, indicating a greater ability to model complex interactions within the multidimensional feature space. These findings demonstrate the importance of reporting class-wise precision, recall, and F1-score rather than relying on overall accuracy alone.

Sensitivity of the minority high risk class was examined using its class-wise recall. The high risk class had a support of 382 samples, compared with

Table 17. Cross model feature importance summary.

Feature	Evidence across model analyses	Overall interpretation
Sleep hours	Frequently identified as highly important	Consistently important
Panic attack history	Frequently identified as highly important	Consistently important
Anxiety score	Frequently identified as highly important	Consistently important
Depression score	Frequently identified as highly important	Consistently important
Financial stress	Repeatedly identified as influential	Frequently important
Work stress	Repeatedly identified as influential	Frequently important
Social support	Repeatedly identified as influential	Frequently important
Physical activity	Repeatedly identified as influential	Frequently important
Family history of mental illness	Repeatedly identified as influential	Frequently important

936 and 1,182 samples for Classes 0 and 1, respectively. Recall for Class 2 varied across the evaluated models, from 0.58 for Naïve Bayes and 0.65 for KNN to 0.98 for CatBoost and 0.93 for BiLSTM. The results indicate that the stronger models provided substantially better sensitivity to high risk cases, while the weaker models showed greater difficulty in identifying this minority class. Therefore, overall accuracy should be interpreted together with Class 2 recall when assessing the suitability of the models for high-risk identification.

Figure 21 compares the accuracy of the nine evaluated models. CatBoost and BiLSTM achieved the highest accuracy at 97%, followed by XGBoost at 96%, Random Forest at 95%, Decision Tree and LSTM at 93%, CNN at 91%, Naïve Bayes at 68%, and KNN at 60%. Because accuracy alone may conceal class specific weaknesses, these results should be interpreted together with macro F1-score and class-wise precision and recall.

Table 19 provides a contextual comparison between the proposed models and selected results reported in previous studies. The reported values are included to position the present findings within the existing literature.

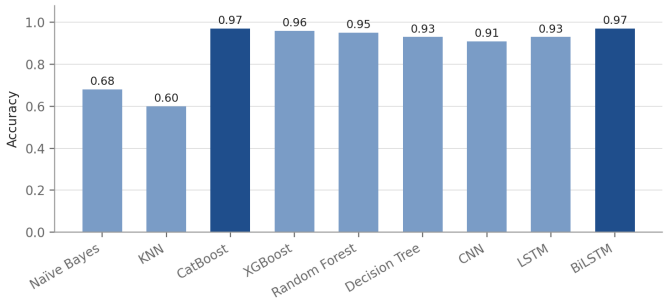

Figure 21. Accuracy comparison of Machine Learning and Deep Learning models.

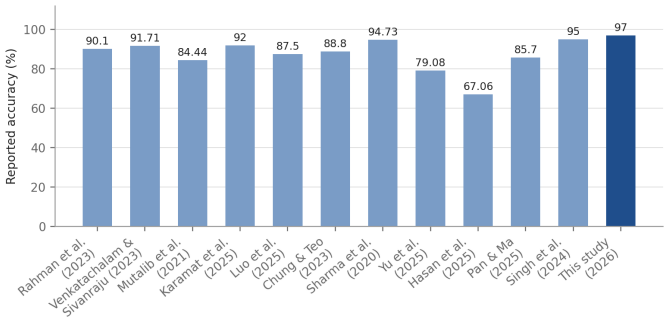

Figure 22. Contextual comparison of the proposed models with selected previously reported approaches.

Figure 22 provides a contextual comparison between the accuracy reported in the present study and selected results from previous studies. These values should not be interpreted as a direct performance benchmark because the studies used different datasets, populations, prediction targets, feature spaces, class definitions, preprocessing procedures, and evaluation protocols. The 97% accuracy obtained by CatBoost and BiLSTM therefore represents performance within the experimental setting of the present study rather than definitive evidence of superiority over all previously reported approaches. The comparison is included to position the observed performance within the broader literature.

The experimental findings provide direct answers to the four research questions. RQ1 is addressed through the feature importance analysis, which identifies sleep hours, panic attack history, anxiety, depression, and several stress, social, and lifestyle factors as recurring influential predictors. RQ2 is addressed through the controlled comparison of six conventional machine learning models and three deep learning architectures under the same experimental setting. RQ3 is addressed by the identification of CatBoost and BiLSTM as the strongest

Table 18. Comparative Performance of Machine Learning and Deep Learning Models for Multiclass Academic Stress and Mental Health Risk Classification.

Model	Accuracy	Weighted precision	Weighted recall	Weighted F1-score	Macro F1-score
Naïve Bayes	0.68	0.71	0.68	0.69	0.65
KNN	0.60	0.62	0.60	0.59	0.59
CatBoost	0.97	0.97	0.97	0.97	0.96
XGBoost	0.96	0.96	0.96	0.96	0.94
Random Forest	0.95	0.95	0.95	0.95	0.93
Decision Tree	0.93	0.94	0.93	0.93	0.91
CNN	0.91	0.91	0.91	0.91	0.90
LSTM	0.93	0.93	0.93	0.93	0.92
BiLSTM	0.97	0.97	0.97	0.97	0.96

Table 19. Contextual comparison of the proposed models with previously reported approaches.

Author name and year	Model	Reported accuracy (%)
Rahman et al. [3]	Random Forest	90.1
Venkatachalam and Sivanraju [17]	Hybrid (LSTM, DCNN, MLP; SADDL)	91.71
Mutalib et al. [22]	Decision Tree	84.44
Karamat et al. [32]	CNN-transformer	92
Luo et al. [24]	Random Forest	87.5
Hasan et al. [25]	CatBoost	67.06
Sharma et al. [27]	Random Forest	94.73
Chung and Teo [28]	Gradient Boosting	88.80
Singh et al. [36]	SVM	95
Pan and Ma [37]	CNN-LSTM	85.7
Yu et al. [38]	Random Forest	79.08
Proposed Research	CatBoost and BiLSTM	97

performing models, with both achieving an accuracy of 0.97 and a macro F1-score of 0.96. RQ4 is addressed through the contextual comparison with previously reported approaches, while recognizing that differences in datasets, populations, prediction targets, and evaluation protocols prevent direct cross-study performance claims.

5 Research Limitations and Future Directions

Although the proposed framework performed well in the multiclass classification of academic stress and mental health risk, several limitations must be acknowledged. First, the study relies on a publicly available secondary Kaggle dataset for which the original sampling frame, recruitment procedure, geographic coverage, participating institutions, and response rate are not documented in the available source. Consequently, the representativeness of the dataset cannot be established, and the findings should not be generalized to the broader student population without external validation using independently collected data from multiple institutions and populations. Second, the dataset primarily covers demographic, academic, lifestyle, psychological, and medical-history variables; other factors, such as family environment, peer pressure, institutional support, and emotional changes, were not included. Third, although SMOTE reduced the class imbalance, synthetic oversampling may produce artificial distributions that do not reflect real-world mental health data, and applying SMOTE after one-hot encoding may generate values between encoded categories (Section 3.3). Fourth, the study focused on classification metrics, and explainability, fairness, privacy, and ethical questions were not addressed; in addition, the mental-health-risk labels may overlap conceptually with several predictors (potential target leakage), and all results derive from a single stratified split without confidence intervals or calibration analysis. Finally, the framework was created for experimental purposes only and has never been used in an actual counseling or academic decision-support system.

Real-world and multi-institutional datasets can now be used to validate the proposed model. Future datasets should also focus on longitudinal studies to include tracking of student stress and student mental health risk over time. Finally, research should focus on developing prediction models that utilize multi-modal data. This could include data such as academic records, attendance, learning management system data, wearable sensor data, and records from other counseling services. Explainable AI methods such

as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) should be incorporated into the model to allow transparency to educational administrators when dealing with student risks. Future work will need to address the ethical implications of utilizing and extending student mental health data, such as the privacy of student data, informed consent policies, and bias mitigation. The framework may also be converted to a real-time intelligent screening dashboard to support early identification and intervention for student mental health risks.

Fairness across demographic groups was not evaluated because the available dataset does not provide sufficient demographic and sampling information for reliable subgroup analysis. Future work should evaluate model performance across relevant demographic groups.

6 Conclusion

This study developed and compared machine learning and deep learning models for the multiclass classification of academic stress and mental health risk into low, moderate, and high categories, which goes beyond the binary case. Using structured data spanning demographic information, academic records, behavioral data, lifestyle, psychology, and medical history, six machine learning models and three deep learning models were built and analyzed. Class imbalance was addressed through data preprocessing, feature scaling, stratified splits, and SMOTE applied to the training data only. The experiments showed that the ensemble models (CatBoost, XGBoost, and Random Forest) and the BiLSTM performed best, whereas the CNN and LSTM did not outperform the tree-based ensembles. CatBoost was the best-performing machine learning model and BiLSTM the best deep learning model; both achieved an accuracy of 0.97 and a macro F1-score of 0.96. The experimental results identified sleep, panic attack history, anxiety, depression, financial and work-related stress, social support, physical activity, and family mental illness history as influential predictors of mental health risk classification. The framework is intended as a research-oriented screening and decision-support approach rather than a diagnostic system. Its practical use requires further external validation and appropriate human assessment. There are some limitations based on the data and models used; therefore, additional validation is necessary before the framework can be applied in practice. Future work should pursue explainable AI, ethical data use, privacy protection, and real-time decision-support systems to build responsible mental health risk prediction systems in educational settings.

Data Availability Statement

The dataset analyzed in this study is publicly available on Kaggle at https://www.kaggle.com/datasets/guriya79/mental-health-disorder/data. The source code and processed data are available from the corresponding author upon reasonable request, subject to the terms governing redistribution of the secondary dataset.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

This study used a publicly available secondary dataset obtained from Kaggle and did not involve the recruitment of participants, interaction with human subjects, or animal experiments. Formal ethical approval was therefore not required. The ethical status of the original data collection, including whether informed consent was obtained from participants, is not documented in the publicly available dataset source and remains the responsibility of the original data creators.

References

[1] Auerbach, R. P., Mortier, P., Bruffaerts, R., Alonso, J., Benjet, C., Cuijpers, P., ... & Kessler, R. C. (2018). WHO World Mental Health Surveys International College Student Project: Prevalence and distribution of mental disorders. *Journal of Abnormal Psychology, 127*(7), 623–638. [CrossRef]

[2] Ibrahim, A. K., Kelly, S. J., Adams, C. E., & Glazebrook, C. (2013). A systematic review of studies of depression prevalence in university students. *Journal of Psychiatric Research, 47*(3), 391–400. [CrossRef]

[3] Rahman, H. A., Kwicklis, M., Ottom, M., Amornsriwatanakul, A., Abdul-Mumin, K. H., Rosenberg, M., & Dinov, I. D. (2023). Machine learning-based prediction of mental well-being

using health behavior data from university students. *Bioengineering, 10*(5), 575. [CrossRef]

[4] Shafiee, N. S. M., & Mutalib, S. (2020). Prediction of mental health problems among higher education student using machine learning. *International Journal of Education and Management Engineering, 10*(6), 1–9. [CrossRef]

[5] Beiter, R., Nash, R., McCrady, M., Rhoades, D., Linscomb, M., Clarahan, M., & Sammut, S. (2015). The prevalence and correlates of depression, anxiety, and stress in a sample of college students. *Journal of Affective Disorders, 173*, 90–96. [CrossRef]

[6] Zhang, L., Zhao, S., Yang, Z., Zheng, H., & Lei, M. (2024). An artificial intelligence tool to assess the risk of severe mental distress among college students in terms of demographics, eating habits, lifestyles, and sport habits: an externally validated study using machine learning. *BMC psychiatry, 24*(1), 581. [CrossRef]

[7] Baba, A., & Bunji, K. (2023). Prediction of mental health problem using annual student health survey: machine learning approach. *JMIR Mental Health, 10*(1), e42420. [CrossRef]

[8] Shatte, A. B., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine, 49*(9), 1426-1448. [CrossRef]

[9] Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports, 21*(11), 116. [CrossRef]

[10] Razavi, M., Ziyadidegan, S., Mahmoudzadeh, A., Kazeminasab, S., Baharlouei, E., Janfaza, V., ... & Sasangohar, F. (2024). Machine learning, deep learning, and data preprocessing techniques for detecting, predicting, and monitoring stress and stress-related mental disorders: scoping review. *JMIR Mental Health, 11*(1), e53714. [CrossRef]

[11] Osman, A. B., Tabassum, F., Patwary, M. J., Imteaj, A., Alam, T., Bhuiyan, M. A. S., & Miraz, M. H. (2022). Examining mental disorder/psychological chaos through various ML and DL techniques: a critical review. *Annals of Emerging Technologies in Computing, 6*(2), 61-71. [CrossRef]

[12] Kannan, K. D., Jagatheesaperumal, S. K., Kandala, R. N., Lotfaliany, M., Alizadehsani, R., & Mohebbi, M. (2026). Advancements in machine learning and deep learning for early detection and management of mental health disorder. *Journal of Affective Disorders Reports, 25*, 101100. [CrossRef]

[13] Zhai, Y., Zhang, Y., Chu, Z., Geng, B., Almaawali, M., Fulmer, R., ... & Du, X. (2025). Machine learning predictive models to guide prevention and intervention allocation for anxiety and depressive disorders among college students. *Journal of Counseling & Development, 103*(1), 110-125. [CrossRef]

[14] Fu, Y., Ren, F., & Lin, J. (2025). Apriori algorithm based prediction of students' mental health risks in the context of artificial intelligence. *Frontiers in Public Health, 13*, 1533934. [CrossRef]

[15] Dhariwal, N., Sengupta, N., Madiajagan, M., Patro, K. K., Kumari, P. L., Abdel Samee, N., ... & Prakash, A. J. (2024). A pilot study on AI-driven approaches for classification of mental health disorders. *Frontiers in Human Neuroscience, 18*, 1376338. [CrossRef]

[16] Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J., Dobson, R. J., & Dutta, R. (2017). Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific reports, 7*(1), 45141. [CrossRef]

[17] Venkatachalam, B., & Sivanraju, K. (2023). Predicting student performance using mental health and linguistic attributes with deep learning. *Revue d'Intelligence Artificielle, 37*(4), 889–899. [CrossRef]

[18] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. [CrossRef]

[19] Nickson, D., Meyer, C., Walasek, L., & Toro, C. (2023). Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review. *BMC medical informatics and decision making, 23*(1), 271. [CrossRef]

[20] Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. *Nature medicine, 28*(6), 1240-1248. [CrossRef]

[21] Tutun, S., Johnson, M. E., Ahmed, A., Albizri, A., Irgil, S., Yesilkaya, I., ... & Harfouche, A. (2023). An AI-based decision support system for predicting mental health disorders. *Information Systems Frontiers, 25*(3), 1261-1276. [CrossRef]

[22] Mutalib, S., Shafiee, N. S. M., & Abdul-Rahman, S. (2021). Mental Health Prediction Models Using Machine Learning in Higher Education. *Turkish Journal of Computer and Mathematics Education Vol, 12*(5), 1782-1792. [CrossRef]

[23] Meda, N., Pardini, S., Rigobello, P., Visioli, F., & Novara, C. (2023). Frequency and machine learning predictors of severe depressive symptoms and suicidal ideation among university students. *Epidemiology and Psychiatric Sciences, 32*, e42. [CrossRef]

[24] Luo, L., Yuan, J., Wu, C., Wang, Y., Zhu, R., Xu, H., ... & Zhang, Z. (2025). Predictors of depression among Chinese college students: A machine learning approach. *BMC Public Health, 25*(1), 470. [CrossRef]

[25] Hasan, M. E., Arif, M., Rakibul Hasan, S. M., Muwanguzi, M., Abaatyo, J., Kaggwa, M. M., ... & Mamun, M. A. (2025). Prevalence, associated factors, and machine learning-based prediction of depression, anxiety, and stress among university students: a

cross-sectional study from Bangladesh. *Journal of Health, Population and Nutrition, 44*(1), 361. [CrossRef]

[26] Macalli, M., Navarro, M., Orri, M., Tournier, M., Thiébaut, R., Côté, S. M., & Tzourio, C. (2021). A machine learning approach for predicting suicidal thoughts and behaviours among college students. *Scientific reports, 11*(1), 11363. [CrossRef]

[27] Sharma, D., Kapoor, N., & Kang, S. S. (2020). Stress prediction of students using machine learning. *International Journal of Mechanical and Production Engineering Research and Development, 10*(3), 5609–5620. [CrossRef]

[28] Chung, J., & Teo, J. (2023). Single classifier vs. ensemble machine learning approaches for mental health prediction. *Brain informatics, 10*(1), 1. [CrossRef]

[29] Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: a scoping review. *Translational psychiatry, 10*(1), 116. [CrossRef]

[30] El Morr, C., Jammal, M., Bou-Hamad, I., Hijazi, S., Ayna, D., Romani, M., & Hoteit, R. (2024). Predictive machine learning models for assessing lebanese university students' depression, anxiety, and stress during COVID-19. *Journal of Primary Care & Community Health, 15*, 21501319241235588. [CrossRef]

[31] Shrestha, I., & Srinivasan, P. (2022). Comparing deep learning and conventional machine learning models for predicting mental illness from history of present illness notations. In *AMIA Annual Symposium Proceedings, 2021* (pp. 1109-1118). https://pubmed.ncbi.nlm.nih.gov/35308915/

[32] Karamat, A., Imran, M., Yaseen, M. U., Bukhsh, R., Aslam, S., & Ashraf, N. (2024). A hybrid transformer architecture for multiclass mental illness prediction using social media text. *IEEE Access, 13*, 12148-12167. [CrossRef]

[33] guriya79. (n.d.). Mental Health Risk Prediction Dataset [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/guriya79/mental-health-disorder/data

[34] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). [CrossRef]

[35] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems, 31*.

[36] Singh, A., Singh, K., Kumar, A., Shrivastava, A., & Kumar, S. (2024, April). Machine learning algorithms for detecting mental stress in college students. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)* (pp. 1-5). IEEE. [CrossRef]

[37] Pan, L., & Ma, H. (2025). A computational CNN-LSTM-based mental health consultation system in a college environment. *Informatica: an international journal of computing and informatics, 49*(10), 147-162. [CrossRef]

[38] Yu, C., Kong, X., Yu, W., Ni, X., Chen, J., & Liao, X. (2025). Machine learning models for predicting the risk of depressive symptoms in Chinese college students. *Frontiers in Psychiatry, 16*, 1648585. [CrossRef]

Shoaib Ahmad is pursuing a Bachelor of Science in Computer Science (BSCS). His research focuses on the design and development of intelligent systems using Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Computer Vision, Natural Language Processing (NLP), Data Mining, and Explainable AI (XAI). He is particularly interested in predictive analytics, healthcare AI, educational data mining, and intelligent decision-support systems. His goal is to contribute to the advancement of AI through innovative research and practical applications that address real-world challenges. (Email: shoaibwatto717@gmail.com)

Ali Imran is pursuing a Bachelor of Science in Computer Science (BSCS). He is passionate about research in Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Computer Vision, Natural Language Processing (NLP), Data Science, and Explainable Artificial Intelligence (XAI). His research focuses on developing intelligent and data-driven solutions for complex real-world problems, with particular interests in predictive analytics, healthcare AI, educational data mining, and intelligent decision-support systems. He is committed to contributing to the advancement of computer science through innovative research, interdisciplinary collaboration, and the practical application of emerging AI technologies. (Email: aliimranworks@gmail.com)

Tahoona Kshif is currently pursuing a Bachelor of Science in Computer Science (BSCS). Her academic interests lie in the fields of Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Computer Vision, Natural Language Processing (NLP), Data Science, and Explainable Artificial Intelligence (XAI). She is actively engaged in research aimed at designing intelligent, data-driven solutions to address challenging real-world problems. Her primary research interests include predictive analytics, AI applications in healthcare, educational data mining, and intelligent decision-support systems. Through innovative research, interdisciplinary collaboration, and the effective use of emerging AI technologies, she aspires to contribute to the continued advancement of computer science and its practical applications. (Email: tahoonakshif@gmail.com)

Zubdah Tur Rasheed is an undergraduate student pursuing a Bachelor of Science in Computer Science (BSCS). Her research interests span a broad range of emerging technologies, including Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Computer Vision, Natural Language Processing (NLP), Data Science, and Explainable Artificial Intelligence (XAI). She is dedicated to exploring innovative AI-driven methodologies for solving complex real-world challenges and advancing intelligent computing systems. Her work particularly emphasizes predictive analytics, healthcare applications of AI, educational data mining, and decision-support technologies. By integrating modern AI techniques with interdisciplinary research, she aims to develop practical, impactful solutions that contribute to the continuous evolution of computer science and its real-world applications. (Email: hafizazubdahtulrasheed@gmail.com)

Ayesha Khan is a Lecturer in the Department of Computer Science at Quaid-e-Azam College of Engineering and Technology. She is dedicated to teaching and research in the field of computer science, with a strong interest in emerging technologies and their practical applications. Her research areas include Artificial Intelligence (AI), Machine Learning (ML), Data Science, Software Engineering, Computer Vision, Natural Language Processing (NLP), and Explainable Artificial Intelligence (XAI). She is actively involved in academic research focused on developing intelligent, data-driven solutions for real-world challenges. Her work emphasizes interdisciplinary collaboration, innovative research methodologies, and the application of advanced computing technologies to address problems in healthcare, education, and intelligent decision-support systems. She is committed to promoting excellence in teaching, fostering research innovation, and contributing to the advancement of computer science through high-quality academic and research activities. (Email: aishakhang09876@gmail.com)

Aamir Ali received the M.S. degree in Computer Science from COMSATS University Islamabad, Sahiwal Campus, Pakistan. He is currently serving as a Lecturer in the Department of Computer Science at Quaid-e-Azam College of Engineering and Technology. His research interests include Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), healthcare analytics, the Internet of Things (IoT), Computer Vision, Natural Language Processing (NLP), and Explainable Artificial Intelligence (XAI). His research focuses on developing intelligent and data-driven solutions for medical diagnosis, speech recognition, facial emotion recognition, predictive healthcare, and clinical decision-support systems. He has also contributed to IoT security research, including malware classification, anomaly detection, and cyberattack prediction. His current research emphasizes the integration of advanced AI techniques into healthcare, cybersecurity, and intelligent systems to address complex real-world challenges through interdisciplinary collaboration and innovative computing approaches. (Email: amirali4436823@gmail.com)

Muhammad Ali is currently pursuing a Bachelor of Science in Computer Science (BSCS). He has a strong academic interest in Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Computer Vision, Natural Language Processing (NLP), Data Science, and Explainable Artificial Intelligence (XAI). His research is centered on designing intelligent, data-driven models to solve complex real-world challenges across diverse application domains. His primary areas of interest include predictive analytics, AI-enabled healthcare, educational data mining, and intelligent decision-support systems. He is dedicated to advancing the field of computer science through innovative research, interdisciplinary collaboration, and the practical implementation of cutting-edge AI technologies to develop impactful and sustainable solutions. (Email: muhammadaliskr1@gmail.com)