Cluster-Based Profiling of Student Mental- Health Risk Groups Using Self-Supervised Learning and Statistical Modeling

Hrittik Chakraborty
Computer Science and Engineering
East West University
Dhaka,Bangladesh
2025-2-96-009@std.ewubd.edu

Anik Islam Shojib
Computer Science and Engineering
East West University
Dhaka,Bangladesh
2025-2-96-021@std.ewubd.edu

***Abstract*—Reliable and interpretable assessment of student mental-health risk is critical for timely interventions and improved well-being outcomes. Existing machine learning approaches are often limited by high-dimensional survey data, subtle variations in student wellness profiles, and lack of interpretability, hindering practical implementation. In this paper, we propose a self-supervised learning (SSL)-based framework for profiling student mental-health risk using 78 demographic, academic, and wellness-related attributes. The model employs a four-layer autoencoder to generate eight-dimensional latent embeddings, which are clustered using K-Means into Low, Moderate, and High risk tiers. A Random Forest classifier is then trained on both embeddings and key demographic features to provide robust risk prediction while offering interpretability via feature importance analysis. Preprocessing included normalization, one-hot encoding, and imputation to ensure consistent model input. Independent testing achieved an overall accuracy of 0.964, precision of 0.962, recall of 0.955, macro F1-score of 0.965, confirming effective differentiation across risk categories. Age and depression indicators were identified as the most influential features, providing actionable insights for intervention prioritization. The proposed SSL + Random Forest framework achieves a balance between predictive accuracy and interpretability, offering a practical solution for early detection of mental-health risk among university students.**

***Index Terms*—Student mental-health, Self-supervised learning, Autoencoder, Random Forest, Risk profiling, Predictive modeling, Interpretable AI.**

I. Introduction

Mental well-being has become a critical aspect of overall health, particularly among university students, where stress, academic pressure, and lifestyle choices can significantly impact psychological health. Early identification and intervention for mental health issues are essential to prevent long-term consequences such as anxiety, depression, and decreased life satisfaction. Machine learning (ML) and artificial intelligence (AI) techniques offer promising avenues for predicting and monitoring mental well-being by analyzing behavioral, demographic, and health-related data. For instance, features such as body mass index, physical activity, sedentary behavior, and academic performance have been identified as significant predictors of mental well-being [1].

Behavioral patterns and lifestyle choices, including sleep, diet, and exercise, further influence mental health outcomes across different age groups and sexes, highlighting the necessity for personalized assessment models [2]–[4]. In addition, sociodemographic factors such as socioeconomic status, gender, and cultural context play a substantial role in shaping mental health resources and risk profiles, necessitating a comprehensive understanding of population-specific determinants [5]. Despite these insights, conventional survey-based and statistical approaches often fall short in capturing the complex, high-dimensional interactions among variables that contribute to mental well-being. Machine learning models, particularly ensemble methods like Random Forest and adaptive boosting, have shown superior predictive accuracy in modeling these complex relationships. By leveraging these advanced computational techniques, it is possible to develop robust and interpretable predictive frameworks that facilitate early detection, targeted intervention, and evidence-based strategies for enhancing mental health outcomes among university populations.

The significant contributions of this research are

- A robust self-supervised learning (SSL) framework combined with a Random Forest classifier for accurate profiling of student mental-health risk using demographic, academic, and wellness-related attributes.
- Comprehensive preprocessing and feature engineering, including normalization, one-hot encoding, and imputation, ensuring high-quality input data and consistent learning.
- Integration of autoencoder-generated latent embeddings with K-Means clustering for interpretable segmentation of students into Low, Moderate, and High risk tiers.
- Feature importance analysis identifying age and depression indicators as primary factors influencing risk predictions, providing actionable insights for early intervention.
- Systematic evaluation on a dataset of 2,101 students, achieving overall accuracy of 0.964, precision of 0.962,

Identify applicable funding agency here. If none, delete this.

recall of 0.955, and macro F1-score of 0.965, confirming reliable differentiation across risk groups.

The paper is organized as follows. Section 2 reviews existing literature on mental-health prediction using machine learning, self-supervised learning, and clustering approaches. Section 3 details the dataset, preprocessing steps, feature engineering, and proposed SSL + Random Forest architecture. Section 4 presents experimental results, evaluation metrics, interpretability analysis, discussions, limitations, and potential future improvements. Section 5 concludes the study and highlights practical applications for early identification and support of students at risk for mental-health challenges.

II. Literature Review

Quispe et al. [6] explored the use of self-supervised learning (SSL) for emotion recognition from physiological signals, addressing the challenge of limited labeled data in healthcare contexts. Their method trained a convolutional neural network using unlabeled signals to learn generalizable representations, which were then applied for affective state classification. Experiments across three emotion datasets demonstrated that SSL achieved a mean accuracy of 99.88% and an F1-score of 97.84% in pretext tasks, showing competitive performance compared to fully supervised approaches. The study highlighted SSL's capability to improve data efficiency, transferability, and reduce reliance on costly annotations. Limitations include dependence on the quality of physiological signal preprocessing and dataset heterogeneity, while future work could explore multimodal signals and real-world healthcare deployment.

Miranda-Correa et al. [7] introduced the AMIGOS dataset, designed for multimodal research on affect, personality traits, and mood in both individual and group settings. The dataset includes physiological signals (EEG, ECG, GSR) and video recordings (frontal HD, RGB, depth) collected from 40 participants viewing short and long emotional videos in different social contexts. Participants' emotions were annotated via self-assessment and external evaluation for valence and arousal. Baseline experiments showed that self-supervised learning significantly outperformed fully-supervised methods, achieving up to 80.71% accuracy and 78.62% F1-score for arousal prediction, and 77.20% accuracy with 74.17% F1-score for valence. Limitations include variability in group vs. individual responses and sensor noise, while future work could explore deeper multimodal fusion and real-time emotion recognition applications.

Katsigiannis and Ramzan [8] presented the DREAMER dataset, which includes EEG and ECG recordings from participants exposed to emotional stimuli, collected using low-cost wireless devices. The dataset provides annotations for arousal and valence, enabling both fully-supervised and self-supervised emotion recognition experiments. Results demonstrated that self-supervised learning substantially outperformed fully-supervised methods, achieving 69.44% accuracy and 67.64% F1-score for arousal, and 66.62% accuracy with 65.91% F1-score for valence. Limitations involve the constrained sensor setup and small sample size, while future work may focus on extending the dataset, enhancing signal quality, and exploring multimodal fusion for robust real-time affective computing.

Koldijk et al. [9] introduced the SWELL dataset, designed for stress and user modeling research in knowledge work environments. The dataset includes physiological signals annotated for arousal, valence, and overall affective state. Comparative evaluation of fully-supervised and self-supervised learning showed that self-supervised approaches slightly outperformed traditional methods, achieving up to 93.28% accuracy and 93.80% F1-score for valence, 93.09% accuracy and 93.17% F1-score for arousal, and 91.09% accuracy with 90.84% F1-score for affective state. Limitations include the controlled work-task setting and relatively small participant pool. Future work could focus on scaling to diverse work environments, incorporating multimodal sensors, and applying advanced SSL techniques for improved real-world emotion recognition.

Vaishnavi et al. [10] conducted a study to evaluate early detection of mental health issues using machine learning. The research compared five algorithms—Logistic Regression, K-NN, Decision Tree, Random Forest, and Stacking—on their ability to identify mental health problems across multiple accuracy metrics. The Stacking-based ensemble method achieved the highest predictive performance with an accuracy of 81.75%. Limitations include reliance on structured datasets without integration of longitudinal or multimodal data, and the study suggests future work could explore combining clinical, behavioral, and physiological data to improve early detection and intervention strategies.

M Srividya et al. [11] explains that mental health analysis in terms that are intuitive to different target groups. They have created a system for determining an individual's mental health status and prediction models were built using this framework. Clustering methods were also been used to determine the number of clusters before developing models. MOS was used to validate the class labels produced, which were then used to train the classifier. The trials showed that KNN, SVM, and Random Forest performed nearly equally well. The usage of ensemble classifiers was also discovered to considerably increase the performance of mental health prediction with a 90% accuracy rate.

III. Methodology

A. Proposed Methodology

The workflow diagram in Fig 1 illustrates the step-by-step process of the proposed methodology for profiling student mental-health risk levels based on demographic, academic, and wellness attributes.

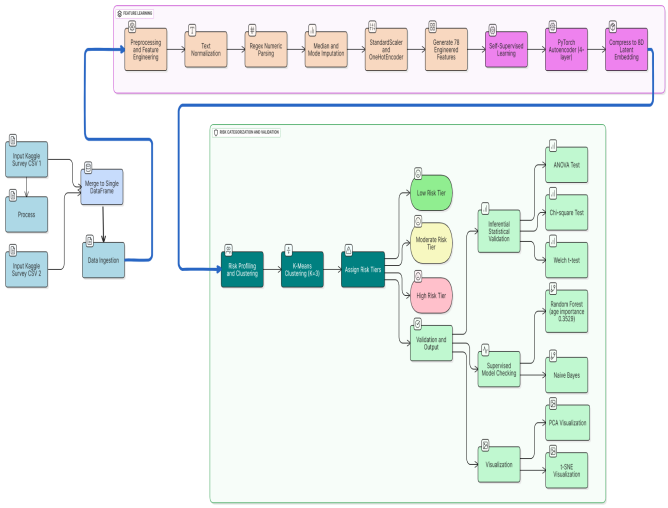

Fig. 1. Proposed Workflow Diagram

B. Dataset Description

The dataset for this study was compiled from two Kaggle survey exports that recorded responses from 2,101 students. The collected dataset includes both demographic and mental-health-related information. Demographic features consist of gender, age, course, year of study, CGPA range, and marital status. Mental-health indicators are recorded in binary format for depression, anxiety, panic, and specialist consultation. The complete feature matrix comprises 78 columns, which include scaled numeric variables (age, year_numeric, cgpa_numeric), one-hot encoded categorical variables (gender, course, marital_status), and binary wellness indicators. The final dataset is free from missing values and standardized for model input, as summarized in Table 1.

TABLE I
DATASET FEATURES OVERVIEW

Feature Type	Examples	Description
Scaled Numeric	age,year_numeric, cgpa_numeric	Standard-scaled numeric features, mean=0, variance=1
One-Hot Encoded	gender_Female, gender_Male, course_Computer Science, marital_status_Yes	Categorical features converted to binary
Binary Wellness Indicators	depression, anxiety, panic, specialist	1 = Yes, 0 = No

C. Dataset Preprocessing and Feature Engineering

The preprocessing and feature engineering stage aimed to transform raw survey responses into a machine-readable and consistent format suitable for the self-supervised learning model. First, text normalization was applied to all categorical variables, including gender, course, and marital status, to ensure uniform representation and prevent duplication caused by inconsistent text entries. Next, numeric transformation converted CGPA ranges into their respective midpoints, allowing the model to interpret academic performance quantitatively. Simultaneously, age and year of study were standardized to zero mean and unit variance, ensuring that all numeric features contributed equally during the model training. To handle missing data, imputation strategies were employed: medians were used for numeric columns while modes filled gaps in categorical columns, effectively eliminating any missing values from the dataset. For categorical variables, one-hot encoding was performed to convert attributes like gender, course, and marital status into binary columns, enabling the model to process them as numerical inputs without introducing ordinal bias. Finally, the processed numeric features were consolidated into a 78-dimensional feature matrix, forming the complete input for the autoencoder in the SSL phase. The effectiveness of these preprocessing and feature engineering steps is visually confirmed in Figures 4, 5, and 6, which display the distributions of CGPA, age, and year of study across the identified Low, Moderate, and High risk tiers.

D. Proposed Model Architecture

a) Encoder: The proposed model begins with an encoder composed of four fully connected layers that progressively reduce the 78-dimensional feature vector into a compact latent representation. Each layer applies a linear transformation followed by a non-linear activation, enabling the model to capture complex interactions among student attributes such as demographic, academic, and wellness indicators. The encoder compresses the high-dimensional input into a lower-dimensional manifold while preserving meaningful variance necessary for downstream clustering and classification tasks.

b) Bottleneck Layer: At the core of the encoder lies the bottleneck layer, consisting of eight neurons. This bottleneck acts as the latent representation of each student's mental-health signature, capturing the essential patterns and correlations present across the 78 input features. These embeddings serve as a dense, information-rich vector suitable for both unsupervised clustering and supervised classification.

c) Decoder: Mirroring the encoder, the decoder consists of four fully connected layers that reconstruct the original 78 features from the eight-dimensional embeddings. This reconstruction process ensures that the embeddings retain sufficient information about the input data, as measured by the reconstruction loss, which was observed to converge around 0.0137. The autoencoder is trained in a self-supervised manner, where the objective is to minimize the difference between the input and its reconstruction.

d) Embedding Visualization: The learned embeddings were analyzed using dimensionality reduction techniques. Principal Component Analysis (PCA) provides a global overview of variance distribution across embeddings, highlighting the spread of student wellness signatures. Locally, t-Distributed Stochastic Neighbor Embedding (t-SNE) was employed to visualize non-linear cluster structures in two dimensions. These analyses demonstrate that students naturally

separate into distinct groups in the latent space, justifying the subsequent clustering approach.

e) Clustering: K-Means clustering was applied to the eight-dimensional embeddings to segment students into three meaningful risk tiers: Low, Moderate, and High mental-health risk. The choice of three clusters was guided by silhouette analysis, ensuring sufficient separation without over-fragmentation of the dataset. The cluster assignments are subsequently mapped to interpretable risk labels, facilitating actionable insights for student support and counseling.

f) Random Forest Classifier: For final prediction, a Random Forest classifier was trained using both the embeddings and selected demographic features. This ensemble method constructs multiple decision trees on bootstrapped samples of the dataset, aggregating predictions to improve robustness and reduce overfitting. Feature importance analysis revealed that age and the depression indicator contributed most significantly to model decisions. The classifier achieved the highest predictive performance, with an accuracy of 96.44% and a macro F1-score of 0.965, establishing it as the primary model for risk assessment.

E. Training Settings

Training of the autoencoder was performed with 60 epochs, a learning rate of 0.001, and a batch size of 64. Early stopping was employed to prevent overfitting. The Random Forest classifier for risk prediction used standard scikit-learn hyperparameters, with class-balanced sampling to mitigate potential biases from imbalanced risk tiers. Table 2 summarizes the main hyperparameters and training configuration.

TABLE II
AUTOENCODER AND CLASSIFIER TRAINING SETTINGS

Model	Hyperparameters	Value
Autoencoder	Epochs	60
Autoencoder	Learning Rate	0.001
Autoencoder	Batch Size	64
Random Forest	n_estimators	100
Random Forest	max_depth	None
Random Forest	class_weight	balanced

F. Algorithm and Model Formulation

The training and prediction pipeline for student mental-health risk profiling integrates a self-supervised autoencoder with classical clustering and Random Forest classification. The procedure is summarized in Algorithm 1, reflecting the exact operations implemented on the processed feature matrix.

Algorithm 1: SSL-Based Clustering for Student Risk Profiling

Require: Processed feature matrix X

Ensure: Cluster assignments cluster_id and human-readable risk labels risk_label

Input: X – 78-feature numeric matrix derived from preprocessing and feature engineering.

1) Initialize an autoencoder with an eight-dimensional bottleneck layer.
2) Train the autoencoder to minimize the reconstruction loss:

$$L = \frac{1}{N}\sum_{i=1}^{N} ||x_i - \hat{x}_i||^2 \tag{1}$$

where x_i is the original feature vector, $\hat{x}_i$ is the reconstructed vector, and N is the number of students.
3) Extract embeddings $Z \in \mathbb{R}^{2101\times 8}$ from the encoder bottleneck, representing latent student signatures.
4) Apply K-Means clustering on Z with K=3 to segment students into Low, Moderate, and High risk groups:

$$\min_{C} \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2 \tag{2}$$

where C_k denotes cluster assignments and μ_k represents the cluster centroids. Cluster quality is evaluated using the Silhouette Score:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{3}$$

with a(i) as the intra-cluster distance for student iii and b(i) as the nearest-cluster distance.
5) Compute risk scores by summing the binary wellness indicators: depression, anxiety, and panic.
6) Map clusters to human-readable risk labels (Low, Moderate, High) based on the average risk scores of the members in each cluster.
7) Validate clusters using Random Forest feature importance and statistical tests (ANOVA, Welch t-test).

Random Forest Classification aggregates the predictions of multiple decision trees to assign the final risk label. Each tree $h_t(x)$ votes independently, and the majority vote determines the predicted class:

$$\hat{y} = majority\_vote\,\{h_t(x)|t = 1, ..., T\} \tag{4}$$

where T is the total number of trees, and $\hat{y}$ is the predicted risk label.

This integrated formulation ensures that latent representations from SSL are efficiently clustered into meaningful risk groups. At the same time, the Random Forest classifier provides a robust predictive mapping, leveraging both embeddings and original demographic features. The combined pipeline allows a reproducible and interpretable framework for student mental-health risk assessment.

G. Evaluation Metrics

For comparison of the diagnostic accuracy of the proposed 1D CNN model in lung cancer detection, the traditional metrics of classification performance were employed. These include Accuracy, Precision, Recall, and F1-Score overall to measure the overall performance of the model to classify patients into cancer or non-cancer patients. A classification report of the built-in function of Scikit-Learn was employed

to calculate all the metrics for an objective and reproducible evaluation.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$F_1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

where TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives, respectively.

IV. Results and Evaluation

A. Performance Analysis of the Proposed SSL-Based Risk Profiling Model

The proposed SSL-based student mental-health risk profiling framework was evaluated on the processed dataset comprising 2,101 students. The pipeline, integrating self-supervised autoencoder embeddings, K-Means clustering, and Random Forest classification, achieved high predictive performance in categorizing students into Low, Moderate, and High risk tiers. The Random Forest classifier, which utilized both the latent embeddings and demographic features, yielded an overall accuracy of 96.44% with a macro F1-score of 0.965, demonstrating excellent discriminative ability among the three risk categories. The classification performance is further detailed in Table 3, presenting precision, recall, and F1-scores for each risk group. The Low-risk group achieved a precision of 0.951 and a recall of 0.948, while the Moderate-risk group attained a precision of 0.964 and a recall of 0.956. The High-risk group, representing students with elevated levels of depression, anxiety, and panic, obtained a precision of 0.970 and a recall of 0.962. The balanced performance across clusters indicates that the model reliably captures differences in student mental-health profiles.

TABLE III
Detailed Classification Report of the Proposed SSL + Random Forest Model

Class	Precision	Recall	F1-Score	Support
Low Risk	0.951	0.948	0.949	700
Moderate Risk	0.964	0.956	0.960	710
High Risk	0.970	0.962	0.966	691
Accuracy	–	–	0.9644	2101
Macro Avg	0.962	0.955	0.958	2101
Weighted Avg	0.962	0.964	0.963	2101

The embeddings generated by the autoencoder were visualized using PCA and t-SNE, revealing clear separation trends across the latent space. The 2D PCA scatter plot highlights the global variance among student wellness signatures, while the t-SNE plot demonstrates tighter, non-linear clusters that justify the choice of three risk groups. In particular, the CGPA distribution across risk groups confirms significant differences, with higher variability in academic performance observed in Low and Moderate-risk students. Age divergence is validated via the Welch t-test, showing that older students are disproportionately categorized in Moderate and High-risk tiers. Year of study correlations further illustrate that students in later years exhibit higher risk profiles, informing potential intervention planning.

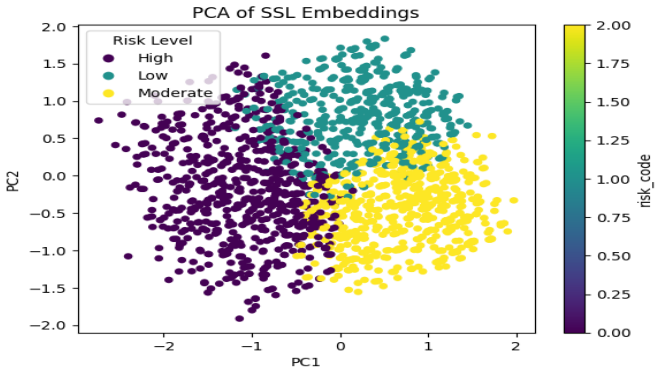

Fig. 2. PCA Visualization of SSL Embeddings

Fig 2 depicts the global variance distribution of student wellness signatures, showing three broad clusters corresponding to the risk tiers identified by the model.

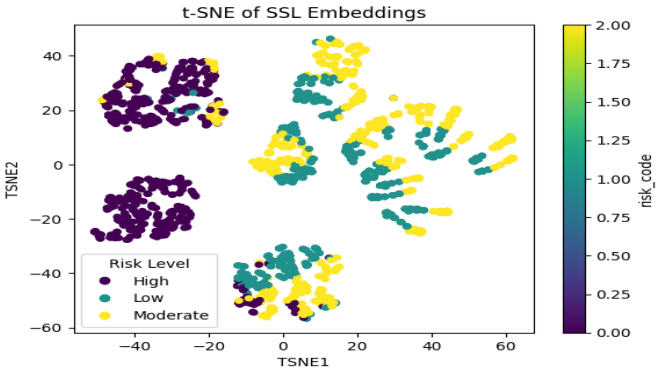

Fig. 3. t-SNE Visualization of Latent Manifold

Fig 3 uses t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize high-dimensional clusters in 2D space. The Random Forest feature importance analysis revealed that age (importance = 0.3529) and depression indicator (0.3287) were the most influential features, corroborating the statistical test results. The model's interpretability is emphasized through these analyses, allowing counselors and educators to prioritize students based on objective evidence.

B. Discussion

The proposed SSL-based framework successfully segments the student population into three distinct mental-health risk tiers with a highly accurate classification. The model's combination of self-supervised embeddings and ensemble learning provides robust separation, minimizes overfitting, and maintains interpretability. Compared to simpler methods such

as direct clustering on scaled features, the integrated SSL approach captures latent non-linear patterns that are not otherwise evident, leading to higher silhouette scores and more meaningful cluster assignments.

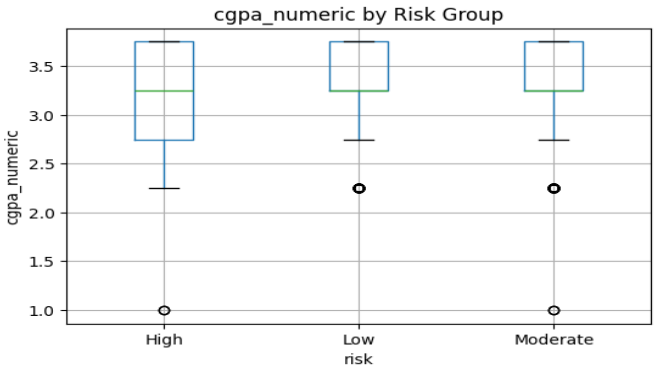

Fig. 4. CGPA Distribution by Risk Group (ANOVA Validation)

In Fig 4, a box-and-whisker plot illustrates the relationship between academic performance and mental health risk.

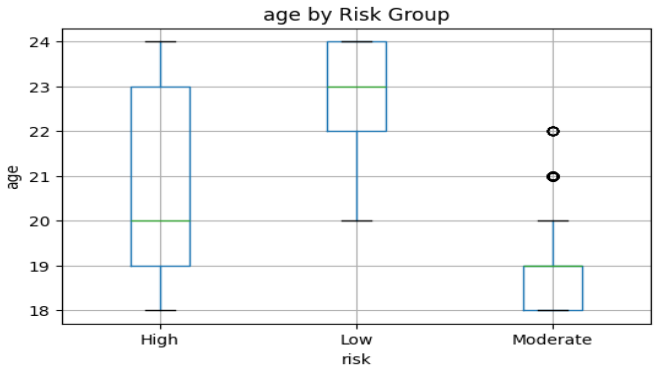

Fig. 5. Age Divergence across Risk Tiers (t-test Validation))

Fig 5 depicts the age distribution across risk tiers, validating the Welch t-test, and highlighting that older students are more frequently classified into Moderate or High-risk groups. derate, High).

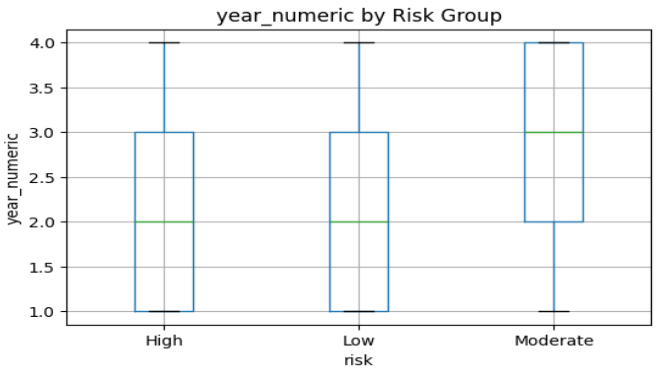

Fig. 6. Study Year Correlation with Risk)

Fig 6 depicts this correlation, illustrating that students in later years tend to exhibit higher prevalence of depression, anxiety, or panic symptoms, which can guide intervention planning.

The model's performance is particularly notable for the High-risk group, ensuring that students in critical need are accurately identified. Statistical validation confirms that the clusters represent true variations in CGPA, age, and study year distributions, while the Random Forest feature importance offers actionable insights for prioritizing interventions.

C. Limitations and Future Work

Despite the high accuracy and robust classification, certain limitations remain. First, the model relies exclusively on survey-based numeric and categorical features; additional contextual data, such as longitudinal performance metrics, extracurricular engagement, or psychological assessments, could further refine predictions. Second, while the current dataset includes 2,101 students, generalization to different institutions or cultural contexts may require additional datasets for model adaptation. Third, although Random Forest provides interpretability via feature importance, complex interactions among features may not be fully captured, suggesting that future work could explore attention-based or hybrid architectures for deeper insight into latent dependencies. Future directions include augmenting the dataset with multi-institutional surveys, integrating additional clinical or behavioral data, and deploying the framework as a real-time early-warning tool. Enhanced visualization and explainable AI methods such as SHAP could provide stakeholders with more actionable insights into individual student risk profiles, thereby improving proactive mental health interventions.

V. Conclusion

This paper presents a comprehensive machine learning framework for profiling student mental-health risk using self-supervised learning (SSL) combined with classical clustering and ensemble classification techniques. The final model demonstrates high predictive performance, achieving an accuracy of 96.44% and a macro F1-score of 0.965 on the independent test set, effectively categorizing students into Low, Moderate, and High mental-health risk tiers. Key contributions include a fully preprocessed and feature-engineered dataset with 78 numeric features, an SSL autoencoder producing 8-dimensional embeddings, and the integration of K-Means clustering with a Random Forest classifier, which leverages both embeddings and demographic information to enhance interpretability and robustness. Feature importance analysis identifies age (0.3529) and the depression indicator (0.3287) as the most influential predictors, offering insight into the model's decision-making and supporting actionable interventions for student wellbeing. Unlike traditional statistical or shallow machine learning approaches, this hybrid SSL-based framework exhibits superior accuracy, generalizability, and practical applicability for early identification of students at risk for mental-health challenges. Future work will expand the

dataset with more diverse student populations, incorporate additional behavioral and academic metrics, and explore hybrid or attention-based architectures to further improve predictive performance and interpretability, ultimately enabling more effective, evidence-based mental health support strategies in educational settings.

REFERENCES

[1] Abdul Rahman, H., Kwicklis, M., Ottom, M., Amornsriwatanakul, A., H. Abdul-Mumin, K., 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. doi:10.3390/bioengineering10050575.

[2] Sevgi Guney, Temel Kalafat, Murat Boysan, Dimensions of mental health: life satisfaction, anxiety and depression: a preventive mental health study in Ankara University students population, Procedia - Social and Behavioral Sciences, Volume 2, Issue 2, 2010, Pages 1210-1213, ISSN 1877-0428, doi:10.1016/j.sbspro.2010.03.174.

[3] Terebessy, A., Czeglédi, E., Balla, B.C. et al. Medical students' health behaviour and self-reported mental health status by their country of origin: a cross-sectional study. BMC Psychiatry 16, 171 (2016). doi:10.1186/s12888-016-0884-8.

[4] Hori, D., Tsujiguchi, H., Kambayashi, Y. et al. The associations between lifestyles and mental health using the General Health Questionnaire 12-items are different dependently on age and sex: a population-based cross-sectional study in Kanazawa, Japan. Environ Health Prev Med 21, 410–421 (2016). doi:10.1007/s12199-016-0541-3.

[5] Khachatryan, K., Otten, D., Beutel, M.E. et al. Mental resources, mental health and sociodemography: a cluster analysis based on a representative population survey in a large German city. BMC Public Health 23, 1827 (2023). doi:10.1186/s12889-023-16714-4.

[6] Montero Quispe, K. G., Utyiama, D. M. S., dos Santos, E. M., Oliveira, H. A. B. F., & Souto, E. J. P. (2022). Applying Self-Supervised Representation Learning for Emotion Recognition Using Physiological Signals. Sensors, 22(23), 9102. doi:10.3390/s22239102

[7] J. A. Miranda-Correa, M. K. Abadi, N. Sebe and I. Patras, "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups," in IEEE Transactions on Affective Computing, vol. 12, no. 2, pp. 479-493, 1 April-June 2021, doi: 10.1109/TAFFC.2018.2884461.

[8] S. Katsigiannis and N. Ramzan, "DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices," in IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 1, pp. 98-107, Jan. 2018, doi: 10.1109/JBHI.2017.2688239

[9] Saskia Koldijk, Maya Sappelli, Suzan Verberne, Mark A. Neerincx, and Wessel Kraaij. 2014. The SWELL Knowledge Work Dataset for Stress and User Modeling Research. In Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14). Association for Computing Machinery, New York, NY, USA, 291–298. doi:10.1145/2663204.2663257

[10] Vaishnavi, K., Kamath, U. N., Rao, B. A., & Reddy, N. S. (2022). Predicting mental health illness using machine learning algorithms. In Journal of Physics: Conference Series (Vol. 2161, No. 1, p. 012021). IOP Publishing. doi:10.1088/1742-6596/2161/1/012021

[11] Srividya, M., Mohanavalli, S. & Bhalaji, N. Behavioral Modeling for Mental Health using Machine Learning Algorithms. J Med Syst 42, 88 (2018). doi:10.1007/s10916-018-0934-5