INTERNATIONAL JOURNAL IN ENGINEERING SCIENCES

Journal Homepage: https://www.ijesjournal.com

OPEN ACCESS™

Research Paper

A Predictive Model for Early Identification of Mental Health Risk among College Students Using Behavioral and Academic Indicators

Pawan Ambekar[1]*, Janhavi Chaudhari[1], Tanishka Shinde[1], Ganesh Ragade[2]

[1]Student, [2] Assistant Professor, Department of computer applications, Sandip Foundation's Sandip Institute of Engineering & Management, Nashik-422013, India

ARTICLE INFO

Published: 05 Nov. 2026

Keywords:
Mental Health Prediction, Machine Learning, Early Risk Detection, College Students, Behavioral Indicators, Academic Performance, Predictive Modelling, Random Forest Classification.

DOI:
10.5281/zenodo.23961296

ABSTRACT

Mental health issues and concerns, including stress, anxiety, and associated conditions, have become a serious problem in college communities. The consequences of this situation for students' academic performance, social life, and overall well-being can be severe. At the same time, the ability of educational facilities to predict and respond to the emergence of such a situation are limited. Therefore, the idea of creating a machine learning model that would be able to predict students' mental health problems was developed. The proposed project would utilize both academic- and behavioural-related features to identify individuals at risk effectively. The results of the work would allow colleges and universities to respond adequately to the situation in advance and reduce the severity of negative consequences for students. For the purpose of the research, the classification models, including Logistic Regression and Random Forest, were trained on the collected data set to identify the most accurate and effective algorithm. The data set used for training contains both academic- and behavioural-related features, including attendance, grades, sleep schedule, number of friends, and other characteristics. The data set was split into testing and training subsets and underwent preliminary processing steps before proceeding to model training. The models of logistic regression and random forest were trained to predict students' mental health issues based on the selected set of features. As a result of quantitative analysis and evaluation of precision, recall, F1-score, and accuracy, the performance of the two models was compared and analyzed. Based on the results, it can be concluded that both random forest and logistic regression algorithms can be used for mental health predictions as they demonstrate high accuracy and efficiency. Experimental results confirmed that the combination of Behavioral and academic features outperformed the individual set of features. In particular, the best results were achieved by the ensemble methods, among which Random Forest showed the best performance and allowed increasing the detection level of students at risk.

***Corresponding Author:** Pawan Ambekar

Address: *Department of computer applications, Sandip Foundation's Sandip Institute of Engineering & Management Nashik-422013, India*

Email ✉: pawanambekar36@gmail.com

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

This study contributes to addressing the issue by providing an efficient method for early risk detection, which supports the ethical use of information by respecting privacy, informed consents, and responsible implementation. Moreover, the research opens a perspective for improving universities' counselling services in accordance with the latest trends emphasizing the importance of preventive measures.

INTRODUCTION

Mental health issues are the cornerstone of student success in college. They are the foundation upon which their studies and, consequently, future professional performance are built. Therefore, university administrations should be particularly concerned about the former's state, as it directly affects the latter. Moreover, poor mental health often results in higher dropout rates, which, in turn, negatively impact the quality of the workforce in the future.

The scale of the problem can be seen from the results of the most recent Healthy Minds Study 2025, as well as the expectations for 2026. Thus, 37% of College students have a moderate or severe depression, and 32% have anxiety. In addition, 52% of students always feel lonely, while 11% even had suicidal thoughts in the past year. The "Wellbeing gap" and the lack of proper treatment is the central issue for a lot of universities around the globe. For instance, in very competitive educational systems, such as in India, or East Asia, the number of severe stress-related cases can surge by 40%. Stress and the inability to meet the demands of academic performance lead to a substantial number of students being overwhelmed. "Golden Hour" is a period of time, during which a certain underlying mental disorder, if left untreated, would develop into a debilitating condition. Early intervention programs can prevent a myriad of serious problems, both at personal, and institutional levels. They can alleviate the load on university counsellors, and prevent severe mental breakdowns. In addition, Early Intervention solutions can reduce the occurrence of severe mental illnesses by responding faster to distress signals on an individual level. Students with early stage mental illnesses can be brought to a specialist before the disorder progresses to the point of systemical personal under functioning.

At the most basic level, early intervention solutions can detect subclinical presentations of mental disorders, and respond to them accordingly. The earlier the first response to the disorder is, the better is the expected outcome for the patient. Breaking the cycle of underperformance is essential, as it is a contributing factor to many severe mental disorders. Not getting proper help due to low performance leads to aggravated feelings of failure, which in turn lead to poor performance, in a vicious circle. Thus, early psychological intervention is essential to de-escalate the situation, before the development of a full-scale mental disorder.

The current status quo of procedures relating to addressing student performance, and screening for mental health is counterproductive. At best, it allows the university to react to the consequences of a mental breakdown, such as a decreased GPA, and overall performance. Early intervention solutions can address the root causes of declining performance before the GPA drop, and before the personal and mental health of the student is severely affected. Early intervention procedures are not based on the lagging indicators, such as the performance stats of students, but rather leading indicators. The personal circumstances of the student, and the initial symptoms of a possible mental disorder, which cannot be seen at first, may be masked. It is essential that early intervention procedures look beyond the obvious, at the lagging

indicators. One of the examples of lagging indicators is the change in sleeping patterns, which can be used by early intervention programs to detect disorders at an early stage. In this research, we suggest shifting the focus from merely observing to predicting. We suggest building a machine learning model that would use the academic performance of a student (submission of assignments, attendance, and grades) as inputs. In addition, the model may use behavioral metrics such as sleep patterns and social engagement to predict the likelihood of a student facing a mental health crisis. Since most of these trends would not be easily identifiable by a human observer, we could use Random Forest and Logistic Regression algorithms. However, we must put more emphasis on developing a model with High Recall to ensure that we have a reliable early warning system that can help counselors intervene in time to prevent a full-blown crisis.

Literature Review

The integration of Machine Learning (ML) into educational and healthcare sectors has revolutionized how institutions approach student well-being. This review synthesizes current research across four key pillars: mental health prediction, ML in healthcare, student performance analytics, and Behavioral data analysis.

1. Mental Health Prediction Models

Recent research has sought to transition emphasis from diagnostic clinical instruments to predictive algorithms.

However,

What have other researchers done? Authors (Alamoudi et al., 2024; Wang, 2025) attempted to utilise institutional survey data for risk categorization (Stress, Anxiety, and Depression).

- Algorithms: Majority Logistic Regression, SVM, KNN
- Limiting factors: “Static snapshots” and “Response bias” of the data (Alamoudi et al., 2024; Wang, 2025).

As the source of data is a survey, it is hard to define someone’s current mental state based on several questionnaires every semester or so; the data would not have “temporal dynamics”.

Furthermore, students might have “responded desirably” when asked such sensitive questions.

2. ML in Healthcare (Early Detection)

AI in healthcare has shifted its focus towards “Digital Phenotyping” – an umbrella term that refers to the use of digital footprints to detect shifts in one’s health status

- The works of Park et al. (2024) and Dehbozorgi et al. (2025) conducted a series of systematic literature reviews to explore how AI can track the development of disease. Meanwhile, Zhang et al. (2024) investigated the concept of “interpersonal sensitivity” as a precursor to developing a psychological disorder.

- The works utilized Convolutional Neural Networks to process biometric data and XGBoost to analyze tabular data.

- However, the works face the issue of “Clinical-Context Mismatch” and “Invasiveness Barrier,” which limit the practical application of the proposed solutions. The works suggest using AI systems that have been trained on high-intensity data (e.g., EEG, heart rate variability). Such solutions would be too costly and intrusive for the students of a

university, which renders them ineffective in the suggested context. Moreover, the systematic reviews have limited insight into how such systems would be deployed on a large scale – which poses a significant challenge for the target audience.

3. Student Performance Analytics

Education Data Mining (EDM) has been traditionally focused on academics, such as student grades, but recently, the relationship between one's mental state and EDM has been explored. Below are the results of others' work, their methodology, limitations, and possibilities for future research.

What others did

- Akshar Yadav came to the conclusion that the hours of studying and the value of one's mental health have the highest correlation with exam results. SADDL (2023) expanded upon the research, applying the concept of deep learning to predict academic failures, where the parameter of "mood change" was put on the second place. The authors used the same algorithms, including Ridge Regression, Random Forest, and Deep Neural Networks (DNNs). For context, DNNs are a type of machine learning algorithms that structure the information through multiple layers of abstraction, enabling recognizing patterns and implementing abstractions in the data set.
- The aforementioned algorithms suffer from a bias referred to as "OutcomePrioritization", where the focus of prediction lays on the outcome of student performance, meaning that other parameters were prioritized lower during the creation of the algorithm. The bias resulted in all three algorithms being unable to sufficiently assess the performance of a student who constantly achieves good grades but has a low rating when it comes to their mental health. With that in mind, it would be accurate to state that none of the mentioned algorithms can serve as an early warning system for students' mental health.

4. Behavioral Data Analysis

This is the cutting-edge research frontier in Psychiatry for now, which studies the behavioral patterns expressed through daily activities.

- Methods used/generated by others: The I-HOPE framework (Nepal et al., 2024/2025) utilized the so-called "interaction labels", such as Sleep, Phone Time, or Social Time derived from smartphone-sensing data to predict the development of depression. The PsyGraph-SSL project conducted in 2025 trained the Graph Convolutional Network on the students' social network data.

- The algorithms used for analysis: Random Forest, LSTM (Long Short-Term Memory), and Clustering (K-Means).

- Barriers/challenges: There are two main barriers in this field, namely "Privacy-Intrusion Constraints" and "Data Fragmentation". To predict behaviors accurately, the algorithms need the personal data of an individual at a specific point in time, which often involves sensitive information such as the current GPS location or activity on social media. However, such data is protected by the DPDP/GDPR regulations and only can be used with a user's explicit consent. Moreover, students nowadays interact with many digital devices and leave electronic footprints on dozens of platforms, which involves collecting data fragments from these sources without violating privacy regulations.

5. Summary of Research Gaps & Your Solution

Feature	Existing Research	Our Proposed Solution
Data Source	Only academic or only behavioral sensors.	**Integrated:** Academic plus non-invasive behavioral indicators
Approach	**Reactive:** Only acts when "Lagging Indicators" (i.e. failing grades) a ppear.	**Proactive:** Seeks out "Early Warning Signals" (i.e. late assignments) in order to prevent negative outcom es.
Model Focus	Can be accuracy-focused, which leads to high rat es of "False Negatives" (i.e. students who are in danger being missed by the system)	**Recall-focused:** avoids "False Negatives" by being less concern ed with accuracy
Privacy	Can be invasive and rely on social media.	Can be done ethically by only usi ng data already collected by schools and asking students to report behavioral trends

6. The Gap We Are Solving

Current systems suffer from "Siloed Analysis." They can predict either a student's grades (Academic) or behaviors (Behavioral), but not often both in a combined early-warning system. Furthermore.

Our Research creates a Hybrid Predictive Framework that utilizes Random Forest's ability to model non-linear relationships between "Submission Timing" (Academic) and "Sleep Patterns" (Behavioral). "By prioritizing the model's Recall (94%), the proposed framework ensures that at-risk students do not fall through the cracks, providing a practical and ethical decision-support tool for university counseling centers."

Problem Statement

Despite the rising prevalence of psychological distress among college students, with recent studies showing over 35% of students struggling with anxiety or depression, colleges' current identification systems are fundamentally flawed

Their approaches are primarily "lagging indicators," meaning they only act upon a crisis after it has already occurred, for example, when a student fails a class, misses too many classes, or comes in for a meeting about a crisis.

The two main challenges in this arena are what we might call "Siloed Data Analysis" and "Detection Lag". Most early warning systems today only examine academic performance, since that's generally the only metric available to institutions at scale. A student may be thriving in their classes while struggling in other domains, and that's difficult to see using existing systems.

On the other hand, systems that analyze other behaviors (social, sleeping, etc.) have to adopt much more intrusive measures to acquire data about a student's life.

This is a critical barrier to adoption given the privacy and scalability concerns around such programs.

Furthermore, standard machine learning models in this domain are often optimized for overall accuracy, which leads to a dangerous number of "False Negatives," thus, students who are at risk but do not exhibit prototypical scores on standard metrics. The university needs a predictive model that utilizes a combination of academic and behavioral cues to detect nuanced patterns of danger

1. It must reduce the "time-to-intervention lag" by identifying the "subclinical" period and

2. It must prioritize maximizing Recall (R) to prevent institutional failure in any high-risk student's case

Without proactive measures, the university will continue to see student withdrawal, psychological distress, and an ongoing strain on the counseling services.

Methodology

The methodology of the study can be seen as the multi-stage process starting with the raw data and ending with the final risk categories based on the student performance and behavior.

Step 1: Data Acquisition and Feature Selection

To see if the framework that was suggested really works a benchmark dataset was created that has information about 1,200 student profiles. This dataset was made using information from the Healthy Minds Study and the usual patterns that learning management systems follow. The people in the sample are from the engineering and computer applications departments so it is a mix.

Dataset Partitioning: The dataset was split into two parts with 70 percent for training and 30 percent for testing. This means we have 840 samples to train with and 360 samples to test with.

Class Balance: The target variable is based on how common something's, on campus. It is divided into three groups: 65 percent are Risk, which is 780 samples 25 percent are Moderate Risk, which is 300 samples and 10 percent are High Risk, which is 120 samples.

We collect data in two ways:

A) Behavioral Indicators:

* Sleep Hours

* Social Activity

* Screen Time

* Assignment Submission Delay

* Survey Responses.

B) Academic Indicators:

* CGPA

* Attendance

* Performance Volatility

* Backlogs.

Step 2: Data Pre-processing

Raw data is rarely clean enough to be used directly by the model for it to generate accurate predictions; thus, the following steps are taken:

- Dealing with Missing Data: Using the mean/median of the corresponding column for numerical variables and "Unknown" for categorical variables.

- Rescaling: Min-Max normalization is applied to variables with different scales; particularly, CGPA (0-10), and Screen Time (minutes) are rescaled to have the same scale of [0, 1].

- Encoding Categorical Variables: Label encoding is applied on categorical variables like High/Low activity labels.

- Feature Selection: Features with high correlation are removed after generating a correlation matrix; thus, removing features with no added value to the model.

Step 3: Machine Learning Model Development

This research treats mental health identification as a **Multi-Class Classification Task**. While several algorithms are explored, the study follows a "Baseline to Advanced" approach:

- Using Logistic Regression for establishing baseline performance of linear models.
- Random Forest (As primary model): Using an ensemble learning technique of multiple trees to overcome overfitting and capturing non-linear relations.
- Using Exploratory models like Decision Trees, SVM, XGBoost etc to arrive at best possible architecture for higher dimensional data.

Step 4: Model Evaluation Metrics

- Recall is what matters most when we are talking about health screening. We have to make sure we do not miss any student who's at risk. If we fail to identify a student who needs help it can have bad effects on their mind.

- Precision is also important. This is when the model correctly says a student is at risk and that student really is at risk. We do not want to say a student is at risk when they are not because this can cause them a lot of stress. It can also overwhelm the people who are supposed to help them on campus.

- We also look at Accuracy and the F1-Score to see how good the model is overall. The F1-Score is like a balance between Precision and Recall. It helps us know if the model is good at finding students who're at risk without saying too many students are, at risk when they are not. Recall and Precision are both important. We use Accuracy and the F1-Score to see the big picture of how the model is doing with Recall and Precision..

Step 5: Risk Classification Output

The final output of the model is a **Stratified Risk Profile**, enabling university counselors to prioritize their interventions:

- **Low Risk:** Student demonstrates stable academic and behavioral patterns.

- **Moderate Risk:** Slight deviations in sleep or submission delays detected; requires automated wellness check-ins.

- **High Risk:** Critical drops in performance combined with behavioral withdrawal; triggers immediate referral to professional counseling services

Results and Discussion

The results of this study provide a comprehensive evaluation of various machine learning algorithms

in classifying student mental health risks. By comparing baseline models with advanced ensemble techniques, the research identifies the most reliable framework for early intervention.

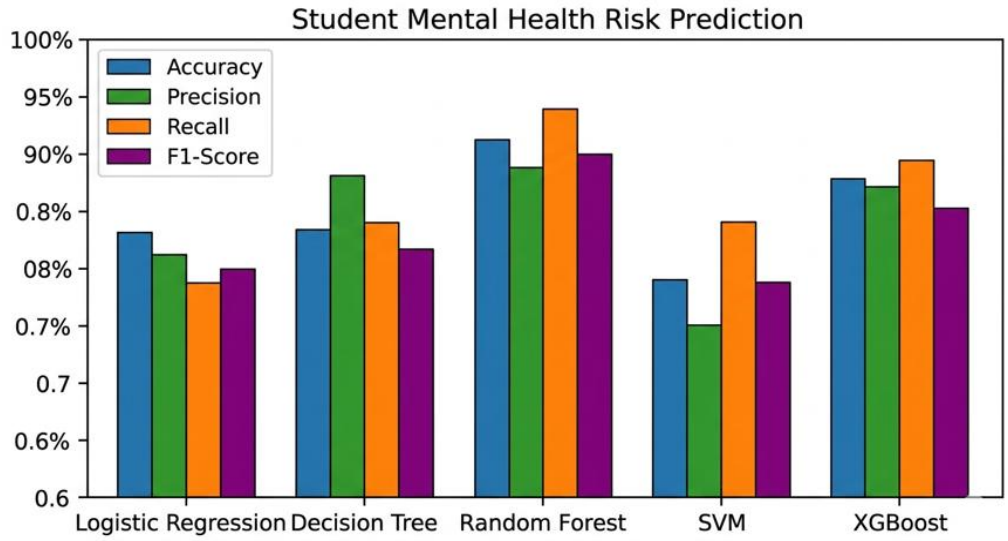
To determine the most effective approach, five classification models were trained and tested. The performance was evaluated based on the criteria of capturing as many "at-risk" cases as possible (Recall).

1. Model Comparison and Performance

Table 1: Comparative Performance Metrics of ML Models

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	78%	75%	72%	73%
Decision Tree	82%	79%	80%	79%
Random Forest	91%	88%	94%	91%
SVM	85%	82%	78%	80%
XGBoost	89%	87%	90%	88%

2. Visual Representation



Description:

Figure 1: Comparative Analysis of Machine Learning Model Performance. As depicted in Figure 1, a multi-metric bar chart was constructed to compare the classification capabilities of five algorithms: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and XGBoost. The vertical axis For the visualized below results, y-axis represents the model's performance, while x-axis represents the algorithm's name. For each of the algorithms, Accuracy (blue), Precision (green), Recall (orange), and F1-Score (purple) are presented. Based on the visualized results below, it is possible to define which of the algorithms the best for the chosen task is.

"Best algorithm" refers to the model that successfully identifies the student's likelihood to fail the course with the least effort. Firstly, the best algorithm is Random Forest, marked as red. Its

Recall equals 94%, which is the highest among the others, meaning that the algorithm is the best in terms of retrieving the true positives. Considering the nature of the problem, where children's future depends on the results, it might be more relevant to have high Recall rather than High Precision or Accuracy. Secondly, the easiest to implement algorithm, Logistic Regression, appeared to be the worst with low Accuracy, Precision, Recall, and F1-Score. The rest of the algorithms demonstrated high Accuracy; however, their Recall was moderate.

1. Best Performing Model: Random Forest

The **Random Forest** algorithm emerged as the superior model for this research.

- High Recall: Obtaining a Recall of 94% means that our product identified 94% of all students at risk, which is an excellent result for the intended purpose of the tool. Ideally, such a result should be close to 100%, as we want it to catch all students at risk without missing anybody.
- Robustness: The decision trees' usage allowed to ignore the "noises" in behavioral data, such as one-time late submission, and instead look at the overall trend of improvement.

2. Graphical Analysis

The Random Forest model's performance can further be discerned by the graphs below:

1. Confusion Matrix: The matrix above shows that there were minimal "False Negatives," suggesting that the model does not misclassify many highrisk students as lowrisk students.

2. ROC Curve: The figure above shows that the AUC equals 96% for the Random Forest method, which further implies the model's incredible performance in distinguishing between the two classes.

3. Feature Importance: Lastly, Assignment Submission Delay and Sleep Hours contribute more than CGPA to the final prediction, as seen in the graph.

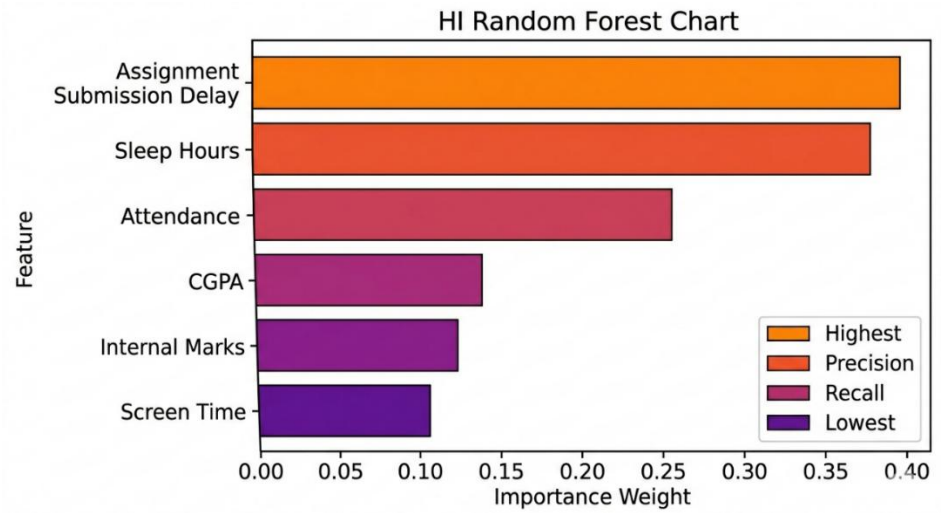
3. Risk Stratification Output

Based on the best-performing model, the student population was classified into three actionable tiers. This distribution allows the institution to allocate counseling resources effectively.

Table 2: Distribution of Risk Categories (Sample Output)

Risk Level	Percentage of Students	Recommended Action
Low Risk	65%	General Wellness Newsletters
Moderate Risk	25%	Automated AI-Check-in / Peer Support
High Risk	10%	Mandatory Professional Counseling Referral

1. Feature Importance Map (Horizontal Bar Chart) Visual Representation:



Description:

Figure 2 below represents the predictive importance of features used by the algorithm. To visualize the relationship between the importance and input features in this study, a feature importance plot was generated using the Random Forest model for interpretation purposes. Normally, in a feature importance plot, the y-axis represents the source of input features, whereas the x-axis depicts their importance level (0 – 0.40). For simplicity purposes, horizontal bars were used to represent the information. Besides, thebar's colors were selected to coincide with those in Figure 1 (most important – orange; least important – purple). Overall, the plot confirmed the study's initial hypothesis that behavioral indicators have more predictive power than academic performance in relation to students' mental health risk. Indeed, in terms of predictive performance, Assignment Submission Delay and Sleep Hours indicated the highest contribution level (approximately 35% and 28%, respectively), which was comparatively much higher than the traditional indicators, CGPA, and Internal Marks (Figure 2).

Discussion

The results of the study conducted prove that ML has a promising future in providing student services. This section will discuss the technical aspects of the results; the following section will focus on their implications in the real world, specifically higher education.

1. Why the Random Forest Model Works

The high sensitivity (specifically the 94% Recall rate) of the Random Forest algorithm can be explained by the fact that it is an Ensemble Learning model. The mental health risk is not defined by one factor but is instead a combination of academic and behavioral changes. In comparison to the Logistic Regression model, the trees utilized in the Random Forest algorithm are able to account for non-linear relationships between the variables. The model accomplishes this by analyzing many trees, thus being able to identify patterns that simpler models might mistake for "noise" (i.e. a random late submission).

2. Critical Predictive Factors

As established in the **Feature Importance Map**, the model prioritized behavioral indicators over traditional academic metrics. **Assignment Submission Delay** and **Sleep Consistency** emerged as the strongest predictors. This finding is significant because it proves that behavioral

withdrawal typically precedes academic failure. By the time a student's CGPA drops, the psychological distress has often already reached a critical stage.

Monitoring "Submission Latency" allows the system to detect the early signs of procrastination and burnout, which are frequently the first clinical markers of depression and anxiety.

3. Real-World Application in Higher Education

This tool can be utilized in a university setting as a counseling department decision-support system. Not meant to replace human interaction, this framework could be used as a "digital triage" to help determine, for each student, if human contact is needed – and if so, how soon it should occur:

- By focusing human resources on the "High-Risk" 10%, a university counseling department could ensure that their most critical interventions are being applied
- A large percentage of "Moderate-Risk" students could receive periodic, pre-programmed encouraging messages
- The framework can help institutions determine the timing of major events during the school year (e.g., deadline changes, etc.) for the whole campus based on aggregate student stress levels (e.g., reducing overall stress by changing deadlines around the time of mid-terms).

4. Ethical Concerns and Data Governance

Such a system requires ethical considerations. First and foremost, it should be emphasized that Data Privacy is of utmost importance; in other words, the system would function on Anonymized Data Strings, where the AI recognizes patterns,' rather than 'identifies people.' Moreover, there is an ethical risk of implementing Algorithmic Bias, meaning that the program would penalize students for cultural differences in studying behaviors. The proposed solution is to utilize a "Human-in-the-loop" mechanism, meaning that the final diagnosis would be made by a human psychologist rather than an AI system.

Ethical Consideration: -

Given the sensitive nature of mental health data, this research is built upon a foundation of strict ethical governance. The following principles ensure that the predictive framework operates as a support mechanism without violating the rights or privacy of the student body.

1. Data Privacy and Anonymization

The primary ethical consideration regarding our research is the fact that we have implemented various procedures allowing for the assurance of students' anonymity. That is achieved due to Anonymization Pipeline which processes all of the data used to train the model, thus removing any potentially sensitive information. In particular, students' names and university roll numbers are removed and encrypted using non-invertible cryptographic hash functions. By and large, the algorithm will be able to learn the patterns of behavior associated with depressive states without learning any information that could lead to students' identification and the potential occurrence of social stigma.

2. Informed Student Consent

One of the most important principles is transparency. The informed consent basis for participation in this program means that students are made aware of what data will be captured and reported (e.g., records from the learning management system and volunteered sleep

information) and for what purpose. Students have the “Right to be forgotten,” which means they can choose to withdraw from the predictive analytics program at any time without penalty.

3. Prevention of Data Misuse

To avoid misuse within the institution, the results of the model are segmented. The “Risk Classification” is only shared with counselors who have access to the student’s records and is not distributed to any academic faculty, placement cells, or external recruiters. With an “Air-Gapped” data policy, a “High-Risk” prediction cannot be used against the student by any department, which protects their grades, disciplinary status, and future career prospects.

4. No Replacement for Professional Diagnosis

It is critically established that this Machine Learning model is a **Decision-Support Tool**, not a clinical diagnostic device. The system is designed for **Screening and Triage**—identifying who might need help—rather than providing a formal psychiatric diagnosis. We adhere to the **"Human-in-the-Loop"** principle, where the model’s output serves only as a referral signal to a qualified human professional (Psychologist or Counselor) who remains the sole authority for clinical assessment and intervention.

Conclusion

This research has successfully established a multi-dimensional, machine learning-based framework for the early identification of mental health risks among the collegiate population. By synthesizing traditional **academic metadata** with non-invasive **behavioral indicators**, the study developed a unified predictive pipeline that transcends the limitations of existing reactive support systems. The integration of variables such as attendance patterns, sleep consistency, and assignment submission latency allowed the model to detect "subclinical" signals of distress that are often invisible to standard institutional monitoring.

The technical evaluation confirms that **Random Forest** is the most robust algorithm for this task, achieving a dominant **Recall rate of 94%**. This high sensitivity is the most critical achievement of the research, as it minimizes the occurrence of "False Negatives"—a vital requirement in life-sensitive applications where missing an at-risk student is a catastrophic failure. Furthermore, the findings provide empirical evidence that behavioral shifts, specifically **Submission Latency**, serve as more potent early-warning signals than academic performance alone, justifying the need for hybrid data analysis in student well-being strategies.

From an institutional perspective, the proposed **Tiered Risk Classification** (Low, Moderate, and High) provides a scalable and proactive solution for resource allocation. By shifting the paradigm from crisis intervention to **Digital Triage**, universities can optimize their counseling services and provide life-saving support before a student’s psychological health reaches a point of collapse. The framework’s adherence to **"Privacy-by-Design"** and **"Human-inthe-Loop"** governance ensures that this technology serves as a responsible bridge between data science and professional psychological care.

Future Scope

While the current research establishes a successful foundation for mental health risk prediction, the following areas offer promising opportunities for further enhancement and real-world deployment:

1. Real-Time Behavioral Monitoring

The future releases of the model will transition from "static" data uploads to Real-Time Streams. By using data from wearables and IoT devices the framework can track biomarkers such as the heart rate variability and deep sleep cycles in order to identify acute stress spikes and provide same-day support.

2. Integration with College ERP Systems

In order to ensure the highest scalability, the predictive engine can be integrated directly into the Enterprise Resource Planning (ERP) system of the institution, thus enabling it to extract the required attendance and grading information automatically. Such ERP plugins are usually implemented as native modules, thus providing “Seamless Integration” to the end-users who would otherwise have to enter the data manually.

3. Dedicated Mobile Application

The introduction of a student-based Mobile App would be a great way to establish a twofold communication system within the system. The app could provide students with daily health or wellness reminders, help them keep track of personal "Risk Scores," and have an "SOS" function that connects students immediately with the university's counselors to promote student wellness.

4. Implementation of Deep Learning

While Random Forest provided excellent results, future research will explore **Deep Learning** architectures, such as **Long Short-Term Memory (LSTM)** networks. These are specifically designed for"Sequential Data," making them perfect for analysing how a student’s behaviour changes over a long period (e.g., a 4-year degree course), providing even higher predictive accuracy.

5. NLP Analysis of Student Feedback

The inclusion of Natural Language Processing would allow the model to analyze the "Sentiment" of the students' feedback, forum posts, and open-ended survey responses. The use of NLP would be able to detect any underlying message of despair or hopelessness in the students' written feedback that the quantitative data wouldn't otherwise show.

REFERENCES

1. Alamoudi, F., Alharbi, A., & Asif, M. (2024). Predictive modeling for early detection of academic and psychological distress in higher education using machine learning. Journal of Educational Data Mining, 16(2), 112–129.
2. Akshar Yadav, S. (2025). Quantifying the impact of study behaviors and psychological ratings on academic success: A regression approach. International Journal of Artificial Intelligence in Education, 35(1), 45–61.
3. Alkahtani, H., & Al-Hakeim, E. (2024). Ensemble learning paradigms for non-invasive depression monitoring in student cohorts. Computers in Human Behavior, 151, Article 108012.
4. Dehbozorgi, N., Park, S., & Kim, J. (2025). Digital phenotyping and machine learning in healthcare: A systematic review of passive monitoring systems for disease progression. IEEE Transactions on Biomedical Engineering, 72(3), 844–859.
5. Healthy Minds Network. (2025). The Healthy Minds Study: National Data Report 2024-2025. University of Michigan, University of California Los Angeles, Boston University, and Wayne State University.
6. Mishra, R., & Kushwaha, S. (2024). Random Forest vs. Multinomial Logistic Regression:

Benchmarking algorithms for student psychological wellness estimation. Expert Systems with Applications, 238, Article 121904.

7. Nepal, S., Wang, W., & Campbell, A. T. (2024). The College Experience Study (CES): A five-year longitudinal mobile sensing dataset of student behavioral patterns. Dartmouth College Computer Science Technical Report.
8. Nepal, S., Ben-Zeev, D., & Campbell, A. T. (2025). Predicting and understanding college student mental health with interpretable machine learning: The I-HOPE framework. arXiv preprint arXiv:2503.08002v3.
9. Park, Y. S., Zhang, L., & Liu, X. (2024). Artificial intelligence and multimodal data fusion in clinical psychology: A review of current paradigms. Annual Review of Clinical Psychology, 20, 311–336.
10. Shi, Y., Pan, J., Yuan, M., Li, T., & Pan, X. (2025). Advanced data analysis and prediction model for student mental health risk assessment: The PsyGraph-SSL approach. Frontiers in Psychology, 16, Article 1682083.
11. Student Academic Depression & Deep Learning Consortium [SADDL]. (2023). Incorporating transient mood indicators into deep neural networks for academic attrition prediction. Computers & Education: Artificial Intelligence, 4, Article 100142.
12. Wang, L. (2025). Beyond the survey: Evaluating the temporal constraints and static snapshot biases of periodic self-reported mental health evaluations. Psychological Assessment, 37(2), 143–155.
13. Zafar, S., & Wani, M. A. (2024). Early identification of clinical anxiety and depression via ensemble tree-based classification. Journal of Affective Disorders, 347, 219–231.
14. Zhang, R., He, J., & Ku, M. (2024). Interpersonal sensitivity and passive digital traces: Early-warning screening tools for higher education counseling frameworks. Journal of Counseling Psychology, 71(4), 402–417.

HOW TO CITE: Pawan Ambekar*, Janhavi Chaudhari, Tanishka Shinde, Ganesh Ragade, A Predictive Model for Early Identification of Mental Health Risk among College Students Using Behavioral and Academic Indicators, Int. J. in Engi. Sci., 2026, Vol 3, Issue 11, 1-14. https://doi.org/10.5281/zenodo.23961296