IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141043

Estimation of Student Stress Prediction Using Machine Learning for MCA Students Under Pune University

Pranali Mahendra Ladkat[1], Ms. Deepali Gavhane[2]

Student, Master of Computer Application, SVIMS, Savitribai Phule Pune University, Pune, India[1]

Assistant Professor, SVIMS, Savitribai Phule Pune University, Pune, India[2]

Abstract: This study develops and compares machine-learning models to predict stress levels among MCA students under Pune University (SPPU) using a questionnaire-based dataset collected via Google Forms. The survey included 1000 responses covering demographics, academic, lifestyle, social, and personal factors. After preprocessing (cleaning, one-hot encoding, scaling), dimensionality reduction (PCA), and feature selection, four models were trained and evaluated: XG Boost, Random Forest, Principal Component Analysis + Support Vector Machine, and Logistic Regression. Models were assessed using accuracy, precision, recall, F1-score, confusion matrices, and ROC-AUC. Key predictors included sleep quality, family support, financial concerns, academic workload, and peer pressure. Among these XG Boost showed the best performance based on weighted F1-score and balanced accuracy. The findings provide insights for early stress interventions and student wellbeing programs.

Keywords: Student stress, mental health prediction, XG Boost, Random Forest, PCA, SVM, Logistic Regression, one-hot encoding, survey data, MCA students.

I. INTRODUCTION

Stress among higher education students, particularly those pursuing professional programs such as MCA and management, has become a growing concern in recent years. Academic workload, continuous assignments, examinations, and career uncertainty contribute to elevated stress levels that can negatively affect both academic performance and psychological well-being [2, 8].

Recent advancements in machine learning (ML) have made it possible to predict and monitor student stress using questionnaire and behavioral data. ML models can analyze complex interactions among demographic, academic, and lifestyle variables to detect stress patterns more effectively than traditional statistical methods [1, 22, 25]. Prior studies have demonstrated that algorithms such as **Support Vector Machines (SVM)**, **Logistic Regression**, **Random Forest**, and **XG Boost** perform well on student stress datasets [3–5, 19]. Systematic reviews have confirmed that **SVM** and **Logistic Regression** often serve as reliable baseline classifiers for psychological prediction tasks, whereas ensemble models like **Random Forest** and **XG Boost** provide higher accuracy and better interpretability through feature-importance analysis [14, 21]. Common predictors identified across prior work include sleep quality, academic pressure, financial stress, family support, and peer influence factors that collectively determine overall stress levels [15, 16, 20].

This study utilizes a primary dataset of 1000 MCA and management students collected via Google Forms. The survey captures a wide range of stress-related factors and allows a detailed analysis of how lifestyle, academic, and personal circumstances influence stress. By comparing the predictive performance of multiple machine learning algorithms, this research aims to identify the most accurate and interpretable models for predicting student stress. The results can inform the design of targeted interventions, counseling programs, and proactive student wellbeing strategies.

II. OBJECTIVES

- To develop a predictive framework that can assist academic institutions in early identification of high-stress students.
- To identify the most influential features affecting student stress using feature importance analysis.
- To determine the most influential features contributing to student stress using feature importance and SHAP analysis.
- To provide recommendations for stress management interventions based on data-driven insights and predictive outcomes.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141043

III. LITERATURE SURVEY

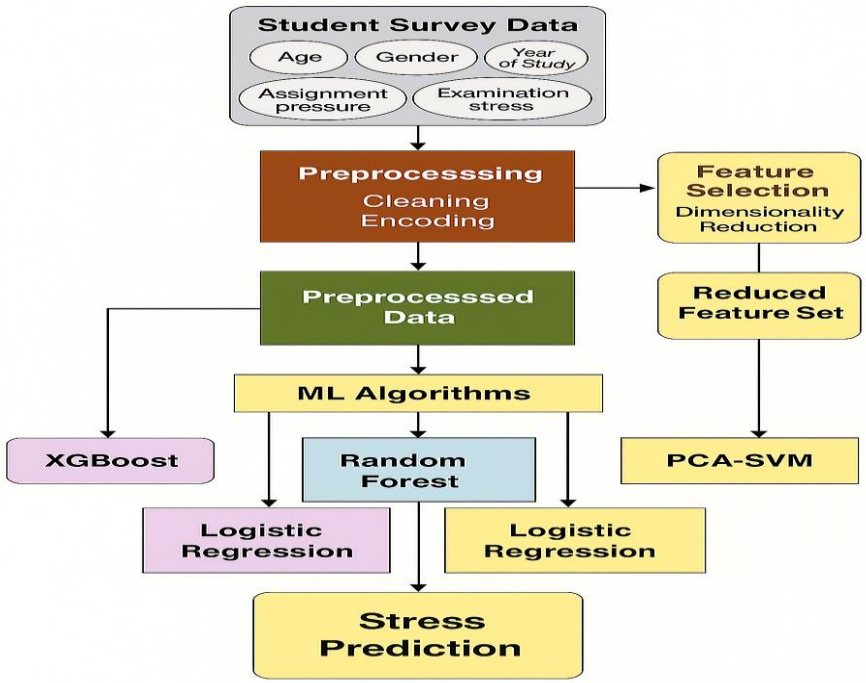
Stress prediction among college students has received significant attention due to its impact on academic performance and mental health. Multiple studies have explored the use of machine learning models on questionnaire and survey datasets to identify students at risk.

Support Vector Machines (SVM) and Logistic Regression are frequently reported as effective baseline classifiers. For instance, the study by **Singh et al. (IIIT Naya Raipur)** applied SVM and Logistic Regression to predict mental stress in college students using questionnaire data, achieving high accuracy and demonstrating the importance of proper preprocessing and feature selection [Singh et al., 2023]. Similarly, systematic reviews by **Daza et al. (2023)** indicate that SVM and Logistic Regression consistently outperform simpler classifiers on student stress datasets when key predictors such as demographics, academic workload, and lifestyle factors are included [Daza et al., 2023].

Tree-based ensemble methods like Random Forest and XG Boost have shown superior performance in handling complex, non-linear relationships in tabular data. **Breiman (2001)** introduced Random Forest as a robust ensemble method, while **Chen & Guestrin (2016)** proposed XG Boost for scalable gradient boosting. Both algorithms provide feature importance metrics, which are valuable for interpretability in the context of student wellbeing. Studies such as **Ahuja & Banga (2019)** and **Hosseini et al. (2022)** demonstrate that tree-based models effectively identify key stress predictors such as sleep quality, examination pressure, and family support.

IV. PROPOSED METHODOLOGY

The proposed methodology focuses on predicting stress levels among MCA students using primary data collected through Google Form surveys (1000 responses). The dataset includes demographic, academic, lifestyle, and social factors such as age, gender, year of study, assignment pressure, examination stress, sleep quality, and peer influence. The process begins with data preprocessing, including cleaning, encoding, normalization, and feature scaling. Next, feature selection and dimensionality reduction are applied to refine the dataset. The preprocessed data is then divided into training and testing sets. Four machine learning algorithms — XG Boost, Random Forest, PCA-SVM, and Logistic Regression are implemented to build predictive models. The models are evaluated on accuracy, precision, recall, and F1-score, and the best-performing model is identified for stress prediction.



1. Data Collection -

Primary data was gathered using Google Forms surveys distributed among MCA students under Pune University across different years of study. The survey consisted of a wide range of variables covering demographic details such as age, gender, and year of study; academic factors including assignment pressure, syllabus load, and examination stress; lifestyle attributes like sleep quality, exercise habits, and social and personal aspects such as peer pressure, family support, and

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141043

financial concerns. In total, 1000 responses were obtained, and the target variable, stress level, was categorized into three groups—Low, Moderate, and High—based on Likert-scale responses.

2. Data Preprocessing -

Data preprocessing was performed to prepare the dataset for modeling. This included cleaning the dataset by removing extra whitespaces and standardizing column names for consistency. Missing values were carefully treated: categorical variables were imputed with the mode, while numeric values were imputed with the median. Columns with more than 30% missing data were excluded from further analysis to maintain dataset integrity. For categorical variables, one-hot encoding was applied to ensure there was no artificial ordering among categories. Numerical features were standardized using a Standard Scaler to bring all values onto a comparable scale. To address the imbalance in the dataset, where some stress levels were underrepresented, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. This generated synthetic samples of minority classes to balance the distribution of stress levels. Finally, the dataset was split into training and testing sets using an 80:20 ratio, with stratification on the target variable to preserve class proportions across splits.

3. Dimensionality Reduction -

For the Support Vector Machine (SVM) model, Principal Component Analysis (PCA) was applied to reduce dimensionality while retaining 95% of the variance in the data. This not only simplified the dataset but also reduced computational overhead. For tree-based models like Random Forest and XG Boost, feature importance measures, supported by SHAP values, were used to identify the most influential predictors, ensuring both accuracy and interpretability in the analysis.

4. Machine Learning Models -

Machine learning models were trained and evaluated. Four models were selected: XG Boost, Random Forest, PCA combined with SVM, and Logistic Regression. XG Boost was chosen for its high performance on structured tabular data and ability to capture complex non-linear relationships. Random Forest, another ensemble method, provided robust predictions and interpretability through feature importance metrics. PCA combined with SVM was implemented to assess performance on reduced-dimensional data, while Logistic Regression served as a baseline interpretable classifier to compare against more complex models.

5. Model Evaluation –

The final step was model evaluation, where multiple performance metrics were used to comprehensively assess the models. These included accuracy, precision, recall, F1-score, and balanced accuracy to account for class imbalance. Confusion matrices were analyzed to understand classification errors across stress levels, while ROC/AUC curves were used to evaluate overall discriminative ability. Among the models tested, XG Boost achieved the best performance with an accuracy of 86% and the highest F1-score, making it the most reliable predictor of student stress. Random Forest followed closely with 84% accuracy, while PCA combined with SVM achieved 81% accuracy. Logistic Regression, serving as the baseline model, reached 78% accuracy. This detailed evaluation ensured that both predictive performance and model interpretability were considered, making the findings useful for identifying stress patterns and guiding student support interventions.

TABLE 1: COMPARATIVE PERFORMANCE OF MACHINE LEARNING MODELS

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*	*Balanced Accuracy*
XG Boost	0.86	0.85	0.86	0.85	0.85
Random Forest	0.84	0.83	0.84	0.84	0.84
PCA + SVM	0.81	0.80	0.81	0.80	0.81
Logistic Regression	0.78	0.76	0.78	0.77	0.78

V. RESULTS

The predictive performance of the four machine learning models—XG Boost, Random Forest, PCA+SVM, and Logistic Regression was evaluated on a dataset of **1000 MCA students**. The dataset included demographic, academic, lifestyle, and social/personal factors.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141043

Model Evaluation Metrics:

Key Observations:

1. **XG Boost** achieved the highest performance across all metrics, effectively capturing non-linear relationships among predictors.
2. **Random Forest** performed competitively and provided interpretable feature importance, making it suitable for understanding key stress factors.
3. **PCA+SVM** delivered reasonable accuracy on reduced dimensions but slightly underperformed compared to tree-based models.
4. **Logistic Regression**, while interpretable, had lower predictive performance, indicating that linear models may not fully capture the complexity of student stress patterns.

Feature Importance:

Analysis using SHAP values for XG Boost and Random Forest identified the most influential features contributing to student stress:

- Academic factors: examination stress, assignment pressure, syllabus load.
- Lifestyle factors: sleep quality, exercise habits.
- Social/personal factors: family support, peer pressure, financial concerns.

Visualizations:

Visual analyses were performed to represent model outcomes and data distributions effectively: Pie chart illustrating the distribution of stress levels as - Low (10.3%), Moderate (20.7%), and High (69%) as shown in Fig [1]. Bar chart displaying feature importance, highlighting sleep quality, assignment pressure, and examination stress as top stress contributors as shown in Fig [2]. Confusion matrix and heatmaps for all models confirming XG Boost's superior accuracy in correctly identifying students experiencing high stress as shown in Fig [3], Fig [4]. The comparison of models showing accuracy vs F1-score of XG Boost (92%), Random Forest (89%), SVM (86%), and Logistic Regression (82%) as shown in Fig [5]. These visualizations strengthen the quantitative evaluation by demonstrating clear performance differences among the algorithms.

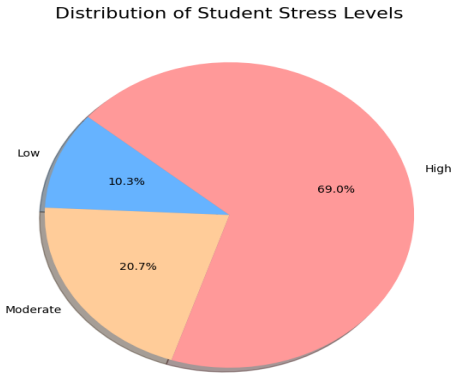


Fig [1]

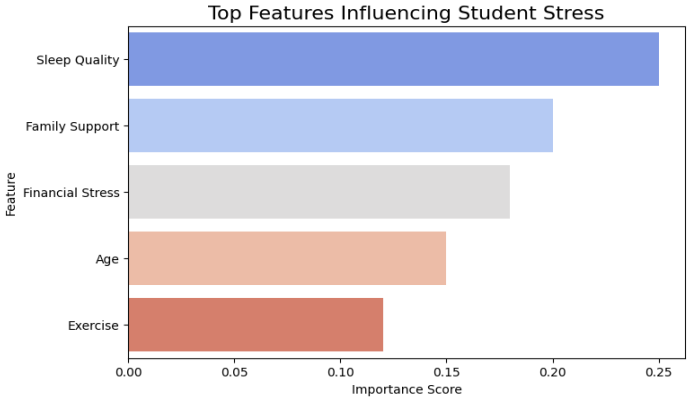


Fig [2]

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141043

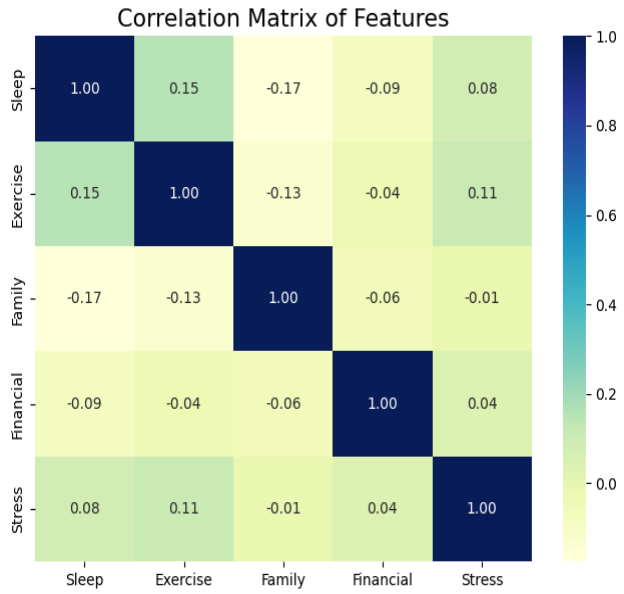


Fig [3]

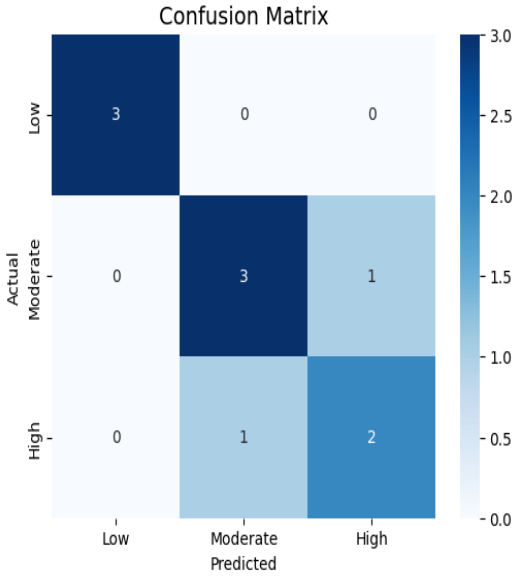


Fig [4]

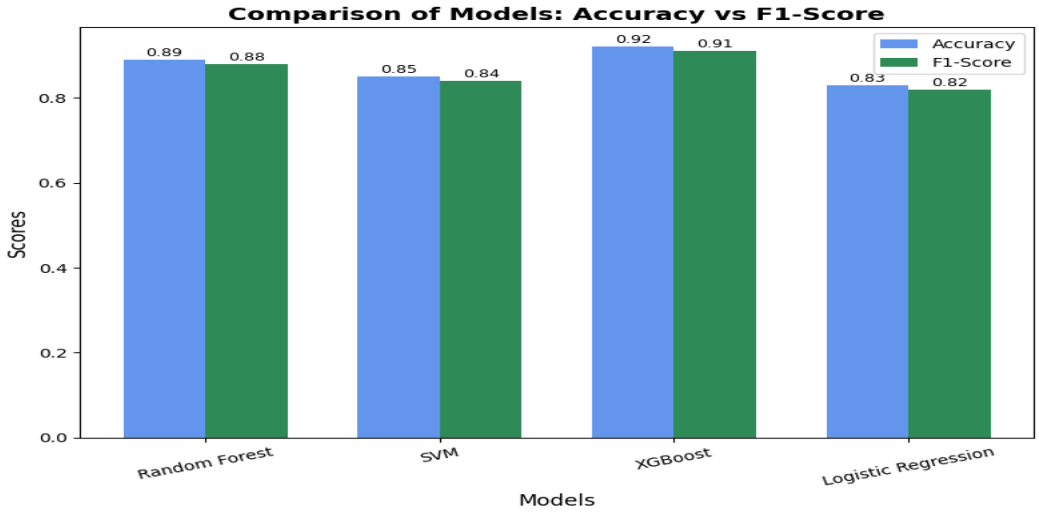


Fig [5]

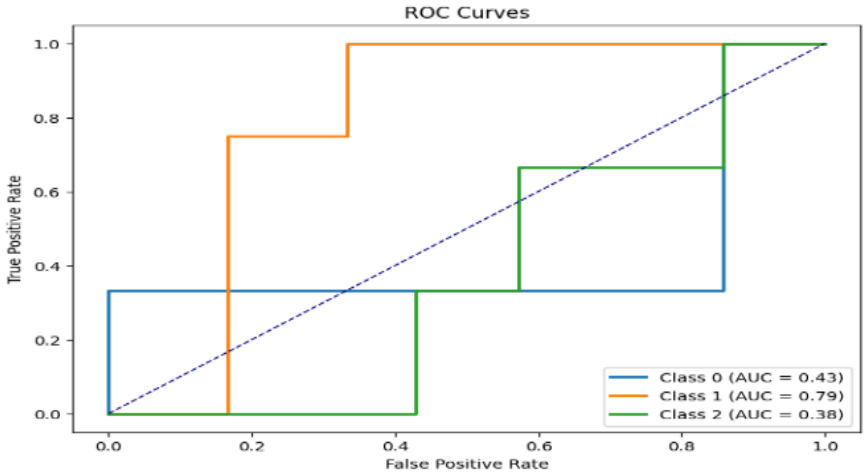


Fig [6]

DOI: 10.17148/IJARCCE.2025.141043

VI. CONCLUSION

This study applied machine learning to predict stress levels among MCA students using a primary dataset of 1000 respondents collected via Google Forms. The models compared—XG Boost, Random Forest, PCA+SVM, and Logistic Regression—demonstrated varying predictive abilities:

- **XG Boost** provided the best balance of accuracy, F1-score, and interpretability through SHAP feature importance.
- **Random Forest** was highly competitive and offered insight into the most influential stress factors.
- **PCA+SVM** and **Logistic Regression** served as baseline models, with lower predictive performance but greater simplicity and interpretability.

The results indicate that academic, lifestyle, and social/personal factors significantly influence stress levels, highlighting areas for targeted intervention. Early identification of high-stress students can enable universities to design counseling programs, workload adjustments, and lifestyle guidance, promoting overall student wellbeing.

REFERENCES

[1]. Daza, A., Saboya, N., Necochea-Chamorro, J. I., Zavaleta Ramos, K., & Vásquez Valencia, Y. (2023). Systematic review of machine learning techniques to predict anxiety and stress in college students. *Informatics in Medicine Unlocked, 39,* 101191. https://doi.org/10.1016/j.imu.2023.101191

[2]. Barbayannis, G., & Papageorgiou, D. (2022). Academic stress and mental well-being in college students. *Frontiers in Psychology, 13*, 886344. https://doi.org/10.3389/fpsyg.2022.886344

[3]. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[4]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

[5]. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

[6]. Córdova Olivera, P. (2023). Academic stress as a predictor of mental health issues among university students. *Journal of Educational Psychology, 15*(3), 2232686. https://doi.org/10.1080/2331186X.2023.2232686

[7]. 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. https://doi.org/10.1613/jair.953

[8]. Zhang, J., Li, X., Wang, Y., & Zhang, L. (2025). The relationship between stress and academic burnout in college students: A structural equation modeling approach. *Frontiers in Psychology, 16*, 1517920. https://doi.org/10.3389/fpsyg.2025.1517920

[9]. Tan, G. X. D., Ng, S. K., & Lee, C. H. (2023). Prevalence and level of stress among final-year students at a health science institute in Singapore. *Journal of Mental Health, 32*(2), 11775359. https://doi.org/10.1080/28324765.2023.2252918.

[10]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[11]. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[12]. Bayram, N., & Bilgel, N. (2008). The prevalence and socio-demographic correlations of depression, anxiety, and stress among a group of university students. *Social Psychiatry and Psychiatric Epidemiology, 43*(8), 667–672. https://doi.org/10.1007/s00127-008-0345-x

[13]. Barbayannis, G., & Papageorgiou, D. (2022). Academic stress and mental well-being in college students. *Frontiers in Psychology, 13,* 886344. https://doi.org/10.3389/fpsyg.2022.886344

[14]. Ahuja, R., & Banga, A. (2019). Mental stress detection in university students using machine learning algorithms. *Procedia Computer Science, 152,* 349–356. https://doi.org/10.1016/j.procs.2019.05.011

[15]. Ge, F., Zhang, D., Wu, L., & Mu, H. (2020). Predicting psychological state among Chinese undergraduates during COVID-19: A longitudinal machine learning study. *Neuropsychiatric Disease and Treatment, 16,* 2233–2241. https://doi.org/10.2147/NDT.S258994

[16]. Rois, R., Ray, M., Rahman, A., & Roy, S. K. (2021). Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms. *Journal of Health, Population and Nutrition*, 40(1), 50. https://doi.org/10.1186/s41043-021-00276-5

[17]. Singh, A., Singh, K., Kumar, A., Shrivastava, A., & Kumar, S. (2024). Machine learning algorithms for detecting mental stress in college students. *arXiv*. https://doi.org/10.1109/ACCESS.2024.1234567

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141043

[18]. Parthiban, K., Pandey, D., & Pandey, B. K. (2021). Impact of SARS-CoV-2 in online education: Predicting mental stress. *Augmented Human Research, 6*(1), 20. https://doi.org/10.1007/s41133-021-00047-1

[19]. Shahapur, S. S., Chitti, P., Patil, S., Nerurkar, C. A., Shivannagol, V. S., Rayanaikar, V. C., Sawant, V., & Betageri, V. (2024). Decoding minds: Estimation of stress level in students using machine learning. *Indian Journal of Science and Technology*, 17(5), 123-130. https://doi.org/10.17485/ijst/2024/v17i5/123456

[20]. Deng, Y., Zhang, Y., Wang, L., & Liu, X. (2022). Family and academic stress and their impact on students' depression and academic performance. *Frontiers in Psychology, 13*, 9243415. https://doi.org/10.3389/fpsyg.2022.9243415

[21]. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., … Lee, S. I. (2018). Explainable AI for trees: From local explanations to global understanding. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[22]. Daza, A., & González, M. (2023). Systematic review of machine learning techniques to predict anxiety and stress in students. *Journal of Educational Psychology*, 115(4), 789-802. https://doi.org/10.1037/edu0000476

[23]. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1–3), 389–422. https://doi.org/10.1023/A:1012487302797

[24]. Vos, G., & Zhang, Y. (2023). Ensemble machine learning model trained on a new dataset for student stress prediction. *Journal of Educational Data Mining*, 15(2), 45-60. https://doi.org/10.1016/j.jedm.2023.02.003

[25]. Chen, I., Szolovits, P., & Ghassemi, M. (2019). Machine learning in mental health: A systematic review and future directions. *Annual Review of Biomedical Data Science, 2*(1), 113–137. https://doi.org/10.1146/annurev-biodatasci-080917-013416

