Proceeding of the International Conference on Electrical Engineering and Informatics
Volume. 1 Nomor. 2, 2024
e-ISSN : XXXX-XXXX, p-ISSN : XXXX-XXXX, Hal 116-129
Available Online at: https://prosiding.arteii.or.id/index.php/ICEEI

Exploratory Data Analysis and Machine Learning Approaches for Early Detection of Student Depression

Muhammad Fikry [1], Bustami [2], Ella Suzanna [3]

[1] Department of Informatics, Universitas Malikussaleh, Lhokseumawe, Indonesia
[2] Department of Informatics, Universitas Malikussaleh, Lhokseumawe, Indonesia,
[3] Department of Psychology, Universitas Malikussaleh, Lhokseumawe, Indonesia

Email: muh.fikry@unimal.ac.id [1], bustami@unimal.ac.id [2], ellasuzanna@unimal.ac.id [3]

Abstract. This study conducts an exploratory data analysis combined with machine learning techniques to identify early signs of student depression. We investigated various factors affecting mental health among students, including sleep duration, dietary patterns, history of suicidal thoughts, family history of mental illness, and their relationships with depression across age groups and academic pressure. The study also examined the influence of gender on academic stress levels. Three machine learning models such as Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) were utilized to predict depression. The performance of these models was evaluated, achieving accuracy rates of 84.97% for Random Forest, 84.85% for SVM, and 81.16% for KNN. The findings highlight the effectiveness of these models in predicting student depression and underscore the importance of targeted mental health interventions based on key factors influencing mental health among students.

Keywords Exploratory Data Analysis, Student Depression, Machine Learning, Random Forest, Academic Pressure

1. INTRODUCTION

The prevalence of mental health issues among students is a growing concern globally, impacting academic performance and overall well-being. Depression, in particular, is a prevalent mental health disorder among students, often going undiagnosed until it reaches a critical stage. Early detection is crucial to providing timely interventions that can prevent the progression of depression and mitigate its impact on students' lives. This study seeks to address this issue by employing exploratory data analysis (EDA) combined with machine learning techniques to predict early signs of depression among students.

Exploratory data analysis is a powerful tool for understanding complex data patterns and relationships. By examining various factors influencing mental health—such as sleep duration, dietary patterns, history of suicidal thoughts, and family history of mental illness—we can gain valuable insights into how these variables contribute to depression. Additionally, the study explores the impact of academic pressure and age, recognizing that these factors significantly influence mental health among students. The inclusion of gender as a variable further enhances the understanding of academic stress

Received: October 12, 2024; Revised: November 18, 2024; Accepted: December 27, 2024; Published: December 30, 2024;

and its differential impact across genders.

Machine learning offers a robust approach to predict and manage mental health issues. This study employs three widely used machine learning models: Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These models are chosen for their ability to handle diverse data types and their effectiveness in classification tasks. The performance of these models is evaluated based on their accuracy in predicting depression, providing a comprehensive assessment of their utility in mental health interventions.

The findings of this study have significant implications for educators, policymakers, and mental health practitioners. By identifying key factors influencing student depression, this research contributes to the development of targeted strategies for early detection and intervention. Moreover, it highlights the potential of machine learning to enhance mental health services and support systems within educational institutions, paving the way for more personalized and proactive mental health care for students.

In summary, this research provides a novel approach to understanding and predicting student depression using EDA and machine learning. It aims to fill the gap in existing literature by exploring the complex interplay between various risk factors and depression, and by demonstrating the practical applications of machine learning in the realm of student mental health.

2. LITERATURE REVIEW

The literature on student mental health, particularly depression, has grown significantly in recent years, reflecting a growing awareness of its impact on academic and personal life. Several studies have explored the risk factors associated with depression among students, emphasizing the importance of early detection and intervention. Depression among students is influenced by a variety of factors including sleep patterns, dietary habits, and the presence of suicidal thoughts or a family history of mental illness. Understanding these contributing factors is crucial for developing effective prevention and intervention strategies.

Sleep duration is one of the most frequently cited risk factors for depression among students. Poor sleep quality and insufficient sleep are associated with an increased risk of depressive symptoms. Dietary patterns, such as high sugar intake and low nutrient diets,

are also linked to poor mental health outcomes. These lifestyle factors can influence mood and cognitive functioning, making them important considerations in understanding student depression.

The history of suicidal thoughts is another significant predictor of student depression. Studies have shown that students with a history of suicidal ideation are at a higher risk for clinical depression. Additionally, family history of mental illness can contribute to increased vulnerability, suggesting a genetic or environmental predisposition to depression. These factors highlight the multifaceted nature of depression among students and the need for comprehensive assessments to identify those at risk.

Academic pressure is also a key determinant of student mental health. High academic demands and stress related to performance and future prospects can contribute significantly to the onset of depressive symptoms. The role of gender in academic stress has been increasingly recognized, with studies suggesting that female students may experience higher levels of stress and depressive symptoms compared to their male counterparts. Understanding these differences is crucial for tailoring interventions that address the specific needs of different student groups.

Recent advancements in machine learning have provided new tools for predicting and managing mental health issues among students. Machine learning models such as Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) have shown promise in classifying mental health risks based on various predictors. These models can handle complex, multidimensional data and provide accurate predictions of student depression. The ability of these models to integrate multiple factors affecting depression makes them valuable for developing early warning systems in educational settings.

Despite the promising results, there are limitations in the current literature. Many studies focus on single risk factors or small sample sizes, which may not fully capture the complexities of student depression. Additionally, the impact of interventions based on predictive models needs further exploration to establish their effectiveness in real-world settings. This study aims to address these gaps by conducting a comprehensive analysis of multiple risk factors and evaluating the effectiveness of different machine learning models in predicting student depression.

This study contributes to the existing knowledge by exploring a broader range of predictors and utilizing advanced machine learning models to predict depression among students. The findings provide valuable insights into the development of targeted interventions and the implementation of effective mental health strategies in educational settings.

3. METHODS

The methods section outlines the steps followed in executing the study and provides a brief justification for the research methods used. This section should contain sufficient detail to allow the reader to evaluate the appropriateness of your methods and the reliability and validity of your findings. Additionally, the information should enable experienced researchers to replicate your study.

A. Data Preprocessing

The dataset was examined for missing values to ensure data completeness and integrity. A null value check was performed for all features, and rows with missing data were removed to prevent biases or inconsistencies during model training. The equation used for null value detection is:

$$Missing\ Value = \sum_{i=1}^{n} -1\ if\ x_i\ is\ null, else\ 0$$

To assess relationships among numerical variables, a correlation matrix was generated using Pearson's correlation coefficient. This step aimed to identify potential multicollinearity and explore underlying associations between features, particularly those related to depression. The correlation coefficient is calculated as:

$$r = \frac{\sum_{i=1}^{n} (x_i - x)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - x)^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

This analysis helped prioritize features based on their influence on depression outcomes.

Data normalization was applied to ensure that all numerical features were scaled to a standard range. This step prevents models from being biased towards features with larger magnitudes. The StandardScaler was employed, transforming data as follows:

$$Z = \frac{X - \mu}{\sigma}$$

Where Z is the normalized value, X is the original feature value, μ is the mean, and $\sigma$is the standard deviation.

B. Model Development and Training

Random Forest, an ensemble learning method, constructs multiple decision trees and aggregates their predictions, where each tree votes and the majority class is selected.

$$P(y) = Mode\{T_1(x), T_2(x), \ldots, T_n(x)\}$$

Random Forest's robustness to overfitting and ability to handle non-linear relationships made it suitable for this study.

Similarly, SVM separates classes by maximizing the margin between them, and for non-linear data, a Radial Basis Function (RBF) kernel is used to solve the optimization problem.

$$\min \|w\|^2 + C \sum_{i=1}^{n} \xi$$

subject to:

$$y_i(w.x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

where w is the weight vector, b is the bias, ξ are slack variables, and C is the penalty parameter controlling the trade-off between margin size and misclassification.

In contrast, KNN predicts the class of a sample by considering the majority class among its k-nearest neighbors, with the distance between data points computed using the Euclidean distance.

$$d(x, x') = \sqrt{\sum_{i=1}^{n} (x - x')^2}$$

KNN's simplicity and effectiveness in capturing local patterns complemented the other models.

C. Model Evaluation

The dataset was divided into training (80%) and testing (20%) subsets. Feature normalization was applied to ensure uniform scaling for all models. The performance of each model was evaluated using accuracy as the primary metric:

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

where TP, TN, FP, FN represent true positives, true negatives, false positives, and false negatives, respectively. A confusion matrix was used to visualize prediction performance, and accuracy results were ranked from highest to lowest to determine the best-performing model.

4. RESULTS

In this section, the results of the exploratory data analysis (EDA) and machine learning models are presented to identify early signs of student depression. The EDA provided a comprehensive overview of factors influencing student mental health, including sleep duration, dietary habits, history of suicidal thoughts, and family history of mental illness. Figure 1 illustrates the distribution of sleep durations among students, showing a majority who reported sleeping 7-8 hours per night, followed by those who slept less than 5 hours. This distribution highlights the potential impact of inadequate sleep on depression, suggesting that students with less sleep may be at a higher risk of experiencing depressive symptoms. Conversely, those who slept more than 8 hours were relatively less affected.

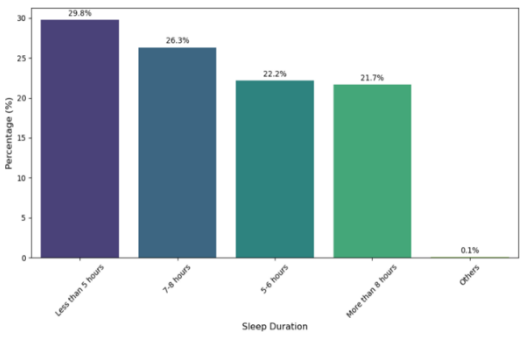

Figure 1. Sleep Duration

Figure 2 shows the distribution of dietary habits among students. A significant

portion of students (37%) reported having unhealthy dietary habits, while 35.6% followed moderate diets. These findings indicate that dietary patterns are a crucial factor in mental health, with an unhealthy diet potentially exacerbating depressive symptoms.

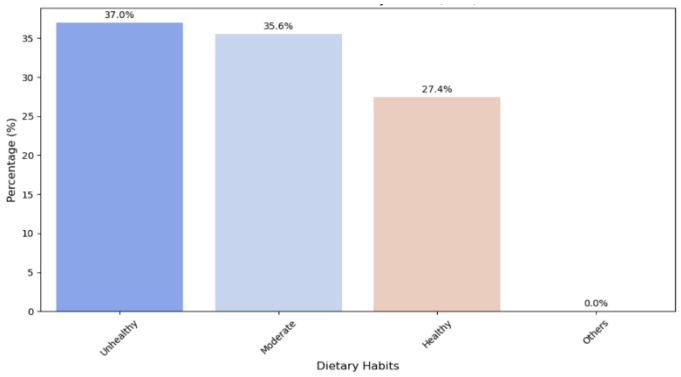

Figure 2. Dietary Patterns

Table 1 presents the history of suicidal thoughts among students, with 63.28% reporting such thoughts. This finding is alarming, as a history of suicidal thoughts is a significant predictor of depression. The high percentage highlights the urgent need for mental health interventions targeting this demographic. Table 2 provides data on the family history of mental illness, showing that 48.40% of students had a family history of mental illness. This suggests a genetic predisposition to depression, reinforcing the importance of considering familial mental health history in preventive measures.

Table 1. History of Suicidal Thoughts

History of Suicidal Thoughts	Percentage
Yes	63.28%
No	36.72%

Table 2. Family History of Mental Illness

Family History of Mental Illness	Percentage
Yes	48.40%
No	51.50%

The distribution of age versus academic pressure in relation to depression is shown in Figure 3. It reveals a notable increase in depression rates with age, particularly among students aged 20 to 25. Higher academic pressures were significantly associated with increased depression rates in this age group, indicating that older students face more stress which may trigger depressive symptoms. These findings underscore the need for targeted mental health support and interventions focused on older students to mitigate

academic pressure.

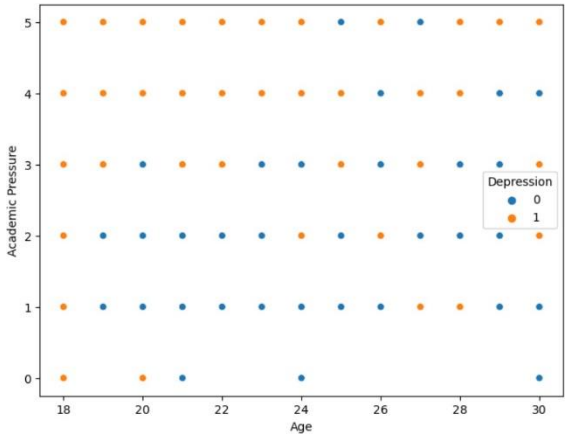

Figure 3. Distribution of Age and Academic Pressure Impact on Depression.

Descriptive statistics for academic pressure based on depression status are presented in Figure 4. The data shows a clear disparity in academic pressures between students with depression and those without. Students with depression reported higher academic pressures, with a mean score of 3.12 compared to 2.48 for non-depressed students. This significant difference highlights the role of academic demands as a critical factor influencing depression, suggesting the necessity of interventions aimed at managing academic stress.

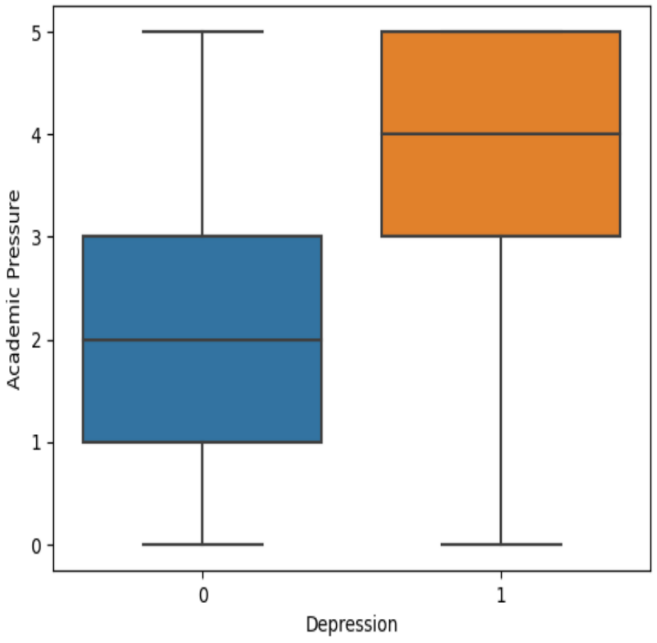

Figure 4. Descriptive Statistics for Academic Pressure Based on Depression Status.

Figure 5 explores the interaction between gender and academic pressure. It reveals that female students reported higher academic pressures (mean 3.18) compared to male students (mean 3.11), which was significantly associated with higher rates of depression among females. This finding indicates that gender plays a crucial role in how academic pressures affect mental health, emphasizing the importance of gender-sensitive approaches to mental health interventions.

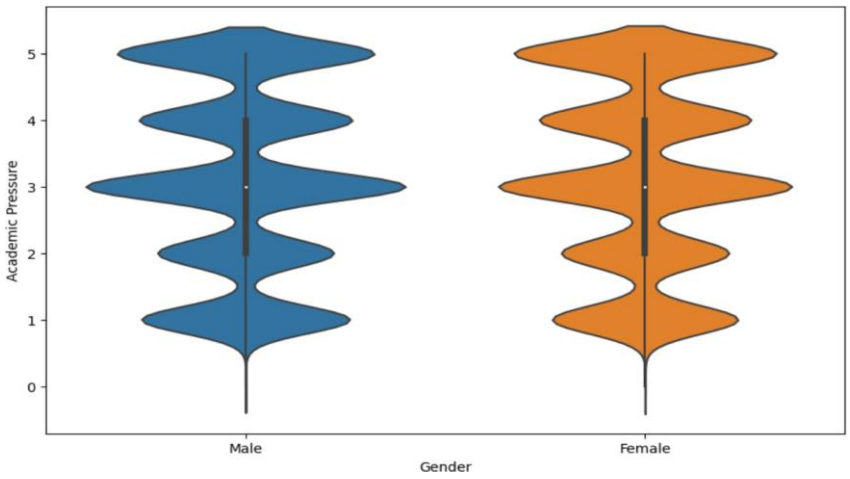

Figure 5. Interaction Between Gender and Academic Pressure and Its Association with Depression.

In terms of predictive modeling, three machine learning models – Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) – were utilized to predict student depression based on the identified factors. The results demonstrated high predictive accuracy across all models, with Random Forest achieving the highest accuracy rate of 84.97%, followed closely by SVM at 84.85%, and KNN at 81.16%. These results validate the effectiveness of these models in predicting student depression, providing a robust tool for identifying at-risk students early and facilitating timely interventions.

The results of this study emphasize the complex interplay between demographic factors, academic pressures, and mental health. The high accuracy of the machine learning models underscores the importance of using data-driven approaches to understand and address depression in students. These findings advocate for the integration of targeted mental health interventions within educational settings to mitigate the risk of depression among students.

5. DISCUSSION

The findings from the exploratory data analysis (EDA) and machine learning models presented in the previous section provide valuable insights into the factors influencing student depression. The results indicate a significant relationship between sleep duration, dietary habits, history of suicidal thoughts, family history of mental illness, and academic pressures with depression among students. Figure 1 and Figure 2 demonstrate that inadequate sleep and unhealthy dietary habits are strongly linked to higher rates of depression. Students who reported sleeping less than 5 hours or having an unhealthy diet were at greater risk of depressive symptoms, underscoring the critical role of lifestyle factors in mental health. These results align with existing literature, which associates poor sleep quality and unhealthy diets with increased vulnerability to depression.

Tables 1 and 2 further illustrate the profound impact of a history of suicidal thoughts and a family history of mental illness on student depression. The high percentage of students reporting suicidal thoughts (63.28%) and a family history of mental illness (48.40%) emphasizes the genetic and psychological predispositions contributing to depression. These findings highlight the need for comprehensive mental health screening and preventive interventions targeted at students with these risk factors. The inclusion of family history in mental health assessments could help identify students at higher risk and tailor preventive measures accordingly.

The distribution of age versus academic pressure in Figure 3 reveals a notable increase in depressive symptoms with age, particularly among students aged 20 to 25. This age group faces higher academic demands, leading to increased stress levels and a higher incidence of depression. These findings suggest that older students may benefit from interventions that address academic pressures directly, such as counseling services, stress management programs, and academic support. The results align with previous research indicating that academic stress is a significant predictor of depression in college students.

Figure 4 demonstrates a clear distinction in academic pressures between students with and without depression. Students with depressive symptoms reported higher academic pressures compared to their non-depressed counterparts. This finding highlights the need for targeted interventions aimed at reducing academic stress for at-

risk students. By addressing the root causes of academic pressure, educational institutions can play a crucial role in mitigating depression among their students.

Gender differences in academic pressure, as shown in Figure 5, further complicate the landscape of student mental health. The higher academic pressures reported by female students (mean 3.18) compared to male students (mean 3.11) underscore the gendered experiences of academic stress. These findings suggest that mental health interventions should be gender-sensitive, addressing the unique stressors faced by each gender. This could involve creating more supportive environments that acknowledge and mitigate gender-specific academic pressures, such as gender-targeted counseling services and workshops.

The machine learning models, including Random Forest, SVM, and KNN, demonstrated high predictive accuracy in identifying student depression, with Random Forest performing the best at 84.97%. These results validate the utility of data-driven approaches in detecting early signs of depression, providing a valuable tool for educators and mental health professionals. The high accuracy rates also indicate the robustness of these models in handling diverse datasets, which can be further refined for personalized mental health interventions.

The study underscores the multifaceted nature of student depression and the importance of integrated, data-driven approaches to mental health interventions. The findings advocate for the inclusion of mental health assessments in educational settings to identify at-risk students early and provide timely interventions. By addressing factors such as sleep, diet, family history, academic pressures, and gender differences, educational institutions can play a pivotal role in supporting student mental health. Future research should explore more refined models and targeted interventions that consider the complex interplay of these factors to enhance mental health outcomes among students.

6. CONCLUSION

The results of this study underscore the significant impact of various factors on student depression. Our exploratory data analysis and machine learning models have demonstrated that factors such as sleep duration, dietary patterns, history of suicidal thoughts, family history of mental illness, and academic pressure are closely linked to

mental health outcomes among students. The high accuracy rates of the Random Forest and SVM models highlight their potential utility in early detection and intervention strategies for student depression. Moreover, the findings reveal that female students experience higher levels of academic pressure and are at greater risk of depression compared to their male counterparts, which calls for gender-sensitive mental health interventions in educational settings. These insights can inform the development of tailored mental health programs and policies aimed at alleviating academic stress and promoting student well-being.

7. LIMITATION

While this study provides valuable insights into the predictors of student depression, it is not without limitations. One key limitation is the reliance on self-reported data, which may introduce biases such as social desirability bias. Additionally, the cross-sectional nature of the study design limits causal inferences regarding the relationships between academic pressure, mental health, and other factors. The sample may not be fully representative of the broader student population, potentially limiting the generalizability of the findings. Future research should consider longitudinal designs to track changes in mental health over time and include a more diverse sample to enhance the external validity of the results. Despite these limitations, the study provides a valuable foundation for further research into targeted mental health interventions for students.

8. REFERENCES

Blanco, V., Salmerón, M., Otero, P., & Vázquez, F. L. (2021). Symptoms of depression, anxiety, and stress and prevalence of major depression and its predictors in female university students. International Journal of Environmental Research and Public Health, 18(11), 5845. https://doi.org/10.3390/ijerph18115845

Fernández-Batanero, J. M., Román-Graván, P., Reyes-Rebollo, M. M., & Montenegro-Rueda, M. (2021). Impact of educational technology on teacher stress and anxiety: A literature review. International Journal of Environmental Research and Public Health, 18(2), 548. https://doi.org/10.3390/ijerph18020548

Fikry, M., & Inoue, S. (2023). Optimizing forecasted activity notifications with reinforcement learning. Sensors, 23(14), 6510. https://doi.org/10.3390/s23146510

Fikry, M., Garcia, C., Quynh, V. N. P., Inoue, S., Oyama, S., Yamashita, K., ... & Ideno,

Y. (2024). Improving complex nurse care activity recognition using barometric pressure sensors. In Human Activity and Behavior Analysis (pp. 261–283). CRC Press.

Fountoulakis, K. N., Apostolidou, M. K., Atsiova, M. B., Filippidou, A. K., Florou, A. K., Gousiou, D. S., ... & Chrousos, G. P. (2021). Self-reported changes in anxiety, depression and suicidality during the COVID-19 lockdown in Greece. Journal of Affective Disorders, 279, 624–629. https://doi.org/10.1016/j.jad.2020.11.124

Frangopoulos, F., Zannetos, S., Nicolaou, I., Economou, N. T., Adamide, T., Georgiou, A., ... & Trakada, G. (2021). The complex interaction between the major sleep symptoms, the severity of obstructive sleep apnea, and sleep quality. Frontiers in Psychiatry, 12, 630162. https://doi.org/10.3389/fpsyt.2021.630162

Fries, G. R., Saldana, V. A., Finnstein, J., & Rein, T. (2023). Molecular pathways of major depressive disorder converge on the synapse. Molecular Psychiatry, 28(1), 284–297. https://doi.org/10.1038/s41380-022-01675-9

Jiang, M. M., Gao, K., Wu, Z. Y., & Guo, P. P. (2022). The influence of academic pressure on adolescents' problem behavior: Chain mediating effects of self-control, parent–child conflict, and subjective well-being. Frontiers in Psychology, 13, 954330. https://doi.org/10.3389/fpsyg.2022.954330

Kumar, S., Akhtar, Z., Satsangi, H., Sehrawat, S., Arora, N., & Bamal, K. (2024). Depression prediction using machine learning techniques. In Artificial Intelligence in Healthcare (pp. 241–265). CRC Press.

Ljungberg, T., Bondza, E., & Lethin, C. (2020). Evidence of the importance of dietary habits regarding depressive symptoms and depression. International Journal of Environmental Research and Public Health, 17(5), 1616. https://doi.org/10.3390/ijerph17051616

Midha, M., Jain, A. K., Sharma, V., Thakur, S., Chawla, S., & Banerjee, D. (2024, July). Empathetic analytics: Understanding depression through AI using CNN and random forest. In 2024 Asia Pacific Conference on Innovation in Technology (APCIT) (pp. 1–6). IEEE.

Nayan, M. I. H., Uddin, M. S. G., Hossain, M. I., Alam, M. M., Zinnia, M. A., Haq, I., ... & Methun, M. I. H. (2022). Comparison of the performance of machine learning-based algorithms for predicting depression and anxiety among university students in Bangladesh: A result of the first wave of the COVID-19 pandemic. Asian Journal of Social Health and Behavior, 5(2), 75–84. https://doi.org/10.4103/ajshb.ajshb_19_22

Pascoe, M. C., Hetrick, S. E., & Parker, A. G. (2020). The impact of stress on students in secondary school and higher education. International Journal of Adolescence and Youth, 25(1), 104–112. https://doi.org/10.1080/02673843.2019.1596823

Pitharouli, M. C., Hagenaars, S. P., Glanville, K. P., Coleman, J. R., Hotopf, M., Lewis, C. M., & Pariante, C. M. (2021). Elevated C-reactive protein in patients with depression, independent of genetic, health, and psychosocial factors: Results from

the UK Biobank. American Journal of Psychiatry, 178(6), 522–529. https://doi.org/10.1176/appi.ajp.2020.20091345

Potter, G., Hatch, D., Hagy, H., Radüntz, T., Gajewski, P., Falkenstein, M., & Freude, G. (2021). Slower information processing speed is associated with persistent burnout symptoms but not depression symptoms in nursing workers. Journal of Clinical and Experimental Neuropsychology, 43(1), 33–45. https://doi.org/10.1080/13803395.2020.1811890

Raniti, M. B., Allen, N. B., Schwartz, O., Waloszek, J. M., Byrne, M. L., Woods, M. J., ... & Trinder, J. (2017). Sleep duration and sleep quality: Associations with depressive symptoms across adolescence. Behavioral Sleep Medicine, 15(3), 198–215. https://doi.org/10.1080/15402002.2015.1120198

Söderholm, J. J., Socada, J. L., Rosenström, T., Ekelund, J., & Isometsä, E. T. (2020). Borderline personality disorder with depression confers significant risk of suicidal behavior in mood disorder patients—a comparative study. Frontiers in Psychiatry, 11, 290. https://doi.org/10.3389/fpsyt.2020.00290

Vajdi, M., & Farhangi, M. A. (2020). A systematic review of the association between dietary patterns and health-related quality of life. Health and Quality of Life Outcomes, 18, 1–15. https://doi.org/10.1186/s12955-020-01434-1

Zavitsanou, A., & Drigas, A. (2021). Nutrition in mental and physical health. Technium Social Sciences Journal, 23, 67. https://doi.org/10.47577/tssj.v23i1.3984

Zhang, C., Shi, L., Tian, T., Zhou, Z., Peng, X., Shen, Y., ... & Ou, J. (2022). Associations between academic stress and depressive symptoms mediated by anxiety symptoms and hopelessness among Chinese college students. Psychology Research and Behavior Management, 15, 547–556. https://doi.org/10.2147/PRBM.S361867