

Article

Not peer-reviewed version

Predicting Student Depression Using Machine Learning Techniques

Noorulain Naimatullah and [Noor Ul Amin](#) *

Posted Date: 20 November 2025

doi: 10.20944/preprints202511.1478.v1

Keywords: depression; prediction model; machine learning



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Predicting Student Depression Using Machine Learning Techniques

Noorulain Naimatullah and Noor Ul Amin *

Department of Software Engineering, University of Sialkot, Sialkot, Pakistan

School of Computer Science, Taylor's University, Subang Jaya, Malaysia

* 1230200243@uskt.edu.pk, nooraminnawab@gmail.com

Abstract

Depression is a well-known health issue and is the third leading cause of disability, following cardiac and respiratory problems. Several research findings indicate that university students are particularly prone to depression, despite being considered a relatively privileged group. However, the recorded prevalence rates show significant variability across different settings, which can negatively impact their academic performance, social relationships, and overall lifestyle. To address this, a machine learning model has been developed using various algorithms to train and predict depression in students based on relevant parameters. The algorithm with the highest accuracy has been proposed for this prediction task.

Keywords: depression; prediction model; machine learning

1. Introduction

There has been increasing interest in the topic, and findings from various studies consistently conclude that the level of depression among college students is alarmingly high and cannot be overlooked[1]. It often leads to increased cases of social and academic difficulties, as well as suicide. Students undergoing the transitional stage from adolescence to adulthood may experience significant stress due to academic pressure, the desire to fit in, and concerns about future planning. Several factors contribute to students' mental health[2]. In our study, responses related to the social construction of failure were found to be significantly positively correlated with depressive moods. Stress also plays a critical role in the onset of depression; the transition from childhood to adulthood is inherently stressful, compounded by peer pressure, academic expectations, and future uncertainties[3,4].

It is natural for students to occasionally feel sad, angry, moody, or even all three at once. However, when such negative moods persist for weeks, months, or even years, and hinder a student's ability to study effectively, it may indicate the presence of clinical depression[5]. This study confirms that students are not exempt from depression; rather, it often exists in a form that goes undetected or untreated [11]. The structure and flow of this study are illustrated in Figure 1, which outlines the key components and methodology of the paper.

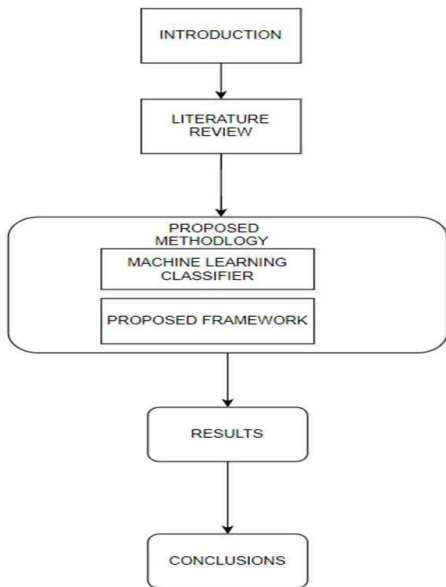


Figure 1. Overall flow of the study.

2. Literature Review

Depression among students is rising rapidly, making it one of the most prevalent mental health conditions globally [12]. It affects students’ academic achievements, social interactions, behavior, and overall daily life functioning [13]. Depression is now recognized as one of the most emerging psychological disorders among university students, as supported by studies such as those by Lyubomirsky et al. (2003) and Vredenburg et al. (1988). Epidemiological findings further emphasize that depression is a multifaceted disorder, contributing to dysfunction in interpersonal, social, and occupational domains [14].

The essence of depression lies in the absence of positive emotions and is often accompanied by disrupted sleep patterns, poor appetite, difficulty concentrating, anxiety, and insomnia [14]. A study published in the Journal of the American Medical Association reveals that the prevalence of depression among college and university students is alarmingly high, with at least one in three students experiencing severe depressive symptoms [1]. According to this study, university campuses urgently require comprehensive mental health support systems, including workshops and counseling sessions, as depression remains a hidden yet deadly condition [15].

If left untreated, depression can lead to serious and dangerous consequences in both students’ personal and professional lives. Academic pressure plays a significant role, as students often place high expectations on themselves, while also facing pressure from family members[16-18]. This can result in feelings of loneliness, anxiety, and helplessness. Several factors contribute to depression among students, including cultural, environmental, biological, and lifestyle-related elements [6].

From a biological perspective, genetic predisposition and neurochemical imbalances can significantly increase the risk of depression [6]. Additionally, personality traits such as low self-esteem or introversion can isolate students from social support systems. Many students also engage in social comparison on platforms like Instagram or TikTok, further affecting their self-worth. Environmental factors, such as academic overload, conflicts with peers, and bullying, also contribute substantially to the worsening of students’ mental health [4], [10]. Unhealthy lifestyle choices such as poor diet, irregular sleep (less than 6–8 hours), and lack of physical activity further exacerbate the condition, particularly during high-stress academic periods. The COVID-19 pandemic, which began in 2020, has significantly worsened these issues, with depression and anxiety rates among students increasing dramatically worldwide [1]–[4].

Recent advancements in machine learning offer promising tools for early detection and prediction of student depression. These models can analyze complex and large datasets with high accuracy, identifying at-risk individuals with significant precision [5], [7], [8], [9], [11]. Some studies have also explored the use of biomarkers, such as cortisol levels, to detect stress and depression at early stages [10]. Till to date, no pharmaceutical treatment has been universally effective for addressing depression among college and university students [6].

Precautionary measures can help reduce the risk of depression. These include maintaining a healthy lifestyle, engaging in regular physical activity, eating nutritious food, and ensuring adequate sleep[19-22]. Moreover, financial stress, family pressure, and lack of emotional support must also be addressed within university frameworks [6], [13].

Effective intervention requires continued research to identify root causes and develop tailored solutions. Universities must provide dedicated support systems to help students manage their mental health proactively. As illustrated in Figure 1, the flow of the study includes the identification of risk factors, data collection, analysis through machine learning models, and interpretation of the results.

3. Proposed Methodology

Depression, as a mental health condition, has been widely studied and modeled using machine learning techniques. In this study, a machine learning-based algorithm is developed to predict the likelihood of depression among students. The implementation is carried out using the RapidMiner platform, which enables efficient handling of data preprocessing, model training, and evaluation. The proposed methodology involves systematic workflow comprising data preprocessing (handling missing values), feature optimization, and classification using various machine learning algorithms. The key classifiers used in this research include KNN, Naïve Bayes, Random Forest, and Decision Tree algorithms. Figure 2 shows the training dataset being clustered into two groups within the feature space during the initial phase of the model development.

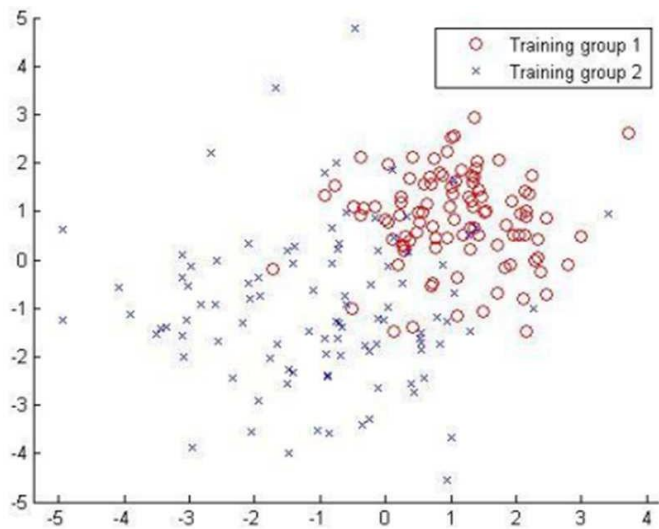


Figure 2. Feature space clustering of training dataset into Group 1 and Group 2 during initial training.

3.1. Framework

The proposed framework includes all essential steps of supervised machine learning, beginning with data cleaning and preprocessing, followed by optimization, training, and testing. Each operation is designed to improve the model’s prediction accuracy by selecting relevant features and discarding noisy or irrelevant ones.

3.2. Dataset Description

The dataset used in this study is publicly available on **Kaggle** and is tailored to analyze and predict depression among students. It includes features such as:

- 1) Demographics: age, gender, academic grade
- 2) Lifestyle factors: sleep habits, physical activity, and social life
- 3) Clinical history: past mental health conditions
- 4) Survey scores: results from standardized depression questionnaires

These attributes allow for a comprehensive analysis of factors associated with depression in the student population. Figure 3. illustrates dataset attributes used for depression prediction, including demographics, lifestyle, and clinical history.

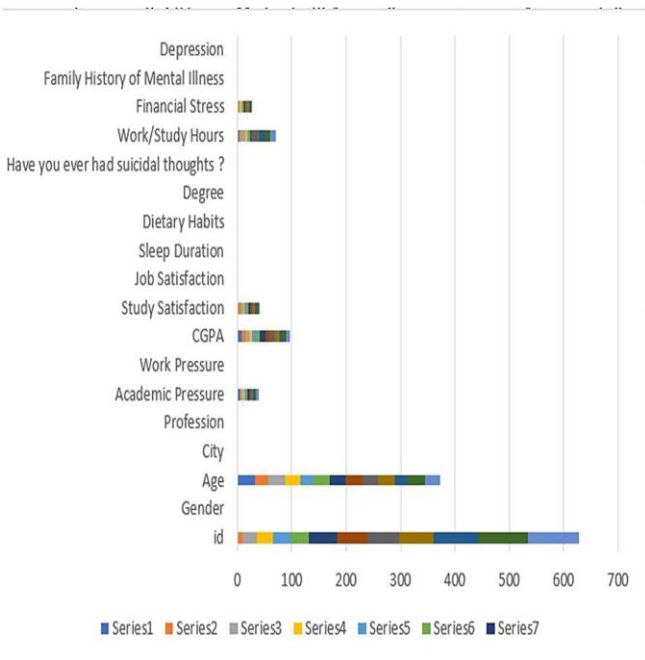


Figure 3. Dataset Attributes.

3.3. Handling Missing Values

Missing values in the dataset were handled using the "Replace Missing Values" operator in RapidMiner Studio. Suitable substitute values were used to fill in gaps, ensuring consistency and reliability in subsequent modeling phases[23-27].

3.4. Optimization and Feature Selection

After handling missing values, feature selection and optimization were performed using the "Optimize Selection" operator. This process removes irrelevant features and improves the model's performance and training efficiency. The dataset was then divided into two subsets:

- a) **Training set** (70%) for model learning
- b) **Testing set** (30%) for model evaluation

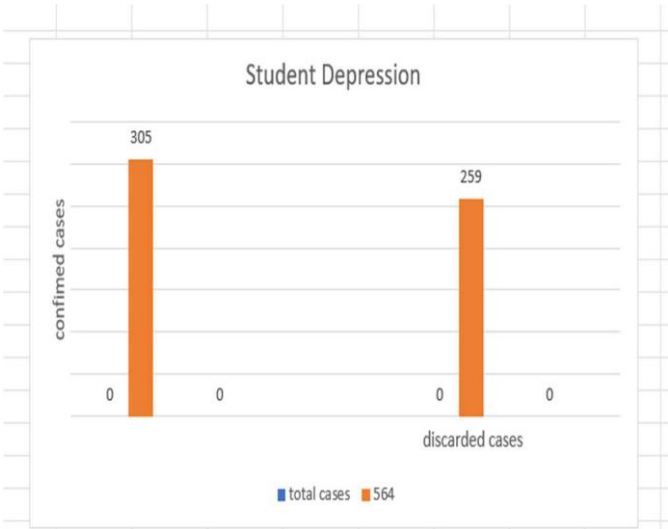


Figure 4. Class distribution of student depression cases showing 305 confirmed and 259 discarded cases among 564 total instances.

This classification highlights the imbalanced nature of the dataset. While the dataset identifies 305 "Yes" (confirmed) cases and 259 "No" (discarded) cases, the source of confirmation or criteria used remains unclear, as such metadata is not publicly available. Figure 5 shows the optimization process in RapidMiner: Operators for preprocessing, selection, classification, and evaluation.

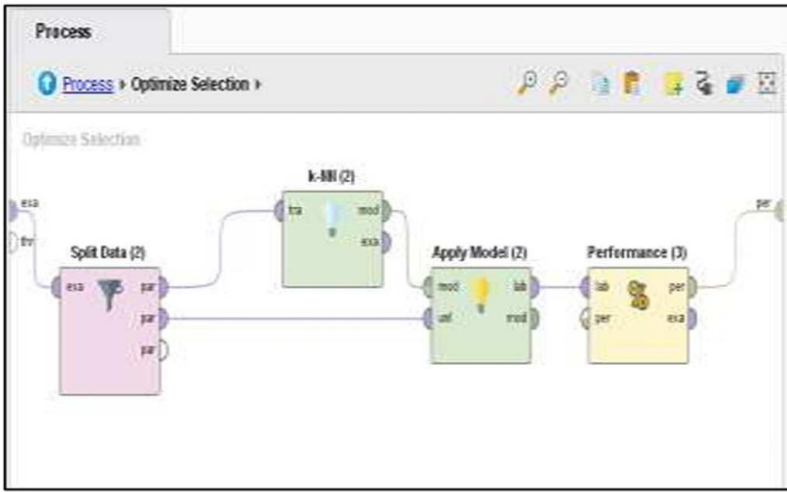


Figure 6. Sample Rapid Miner Workflow.

Operators such as Split Data, Apply Model, and Performance were used. The Split Data operator partitions the dataset into training and testing subsets. The Apply Model operator applies the trained model, and the Performance operator evaluates metrics such as accuracy, precision, and recall. The accuracy for each classifier was calculated, and the one yielding the highest predictive performance was selected for final implementation. Accuracy is defined as the ratio of correctly predicted instances to the total number of predictions made.

4. Results

This section presents the results obtained by applying various machine learning algorithms to predict depression among students. The evaluation primarily focuses on the accuracy of each algorithm, and the confusion matrix for the highest-performing model is also discussed.

To determine the best-performing algorithm, the accuracy of each classifier was computed and compared. The results clearly indicate that the Decision Tree algorithm achieved the highest prediction accuracy among all tested models. Once the highest accuracy model was identified, a confusion matrix was generated to further assess its performance. Figure 7 shows the accuracy comparison.

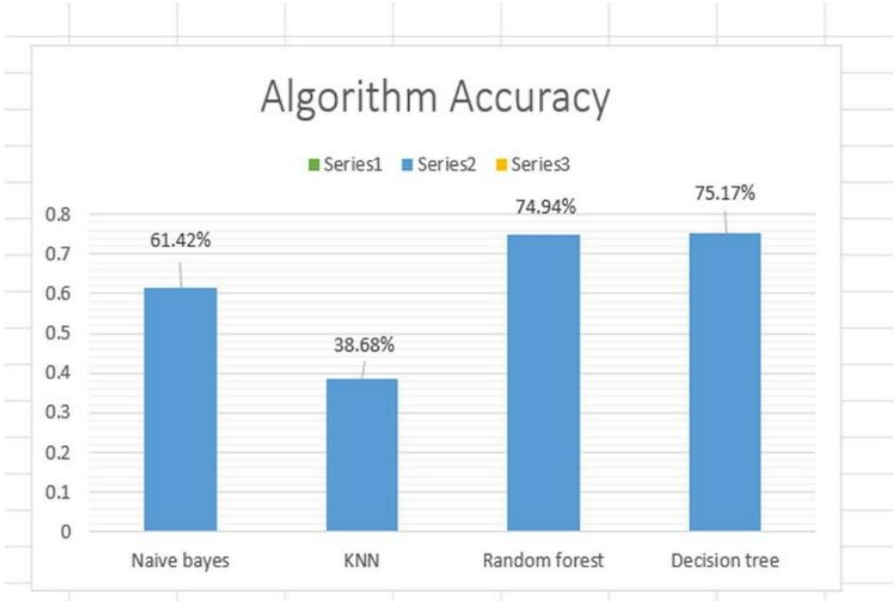


Figure 7. Accuracy of Algorithm.

In the above figure it is quite observable that the Decision tree algorithm recorded the highest accuracy percentage among the classifiers. The number of accurate results which can be achieved by using various algorithms has been summarized below. From this chart, it is stated that the highest algorithm accuracy has been provided by the decision tree altogether 75.17%. For boundaries of decision that are complex decision tree algorithm models such as student depression interconnectivity of features such as (sleep duration, study satisfaction and work duration) do not present simple linear relations. Tree is making split decisions, and the maxims of decision might be based on several features and that results to high accuracy because delicate relationships are preserved. The confusion matrix which is produced by decision tree algorithm is shown Table 1.

Table 1. Confusion matrix for Decision Tree classifier with precision and recall values.

	True: Yes	True: No	Class Precision
Predicted: Yes	429	116	78.72%
Predicted: No	97	216	69.01%
Class Recall	81.56%	65.06%	

5. Conclusion

Multiple supervised learning algorithms were evaluated for the classification task, and the results are presented through comparative analysis. Among all applied models, the Decision Tree algorithm achieved the highest accuracy of 75.17%, followed closely by Random Forest with 74.94%. In contrast, Naïve Bayes and KNN recorded lower accuracies 61.42% and 38.68%, respectively, in the initial fold. These findings highlight the superior performance of tree-based models for this classification task. Future work may involve testing these models on additional datasets to further validate their generalizability.

References

1. Wang, C., Wen, W., Zhang, H., Ni, J., Jiang, J., Cheng, Y., ... & Liu, W. (2021). Anxiety, depression, and stress prevalence among college students during the COVID-19 pandemic: A systematic review and meta-analysis. *Journal of American college health*, 71(7), 2123-2130.
2. Zhang, Y., Bao, X., Yan, J., Miao, H., & Guo, C. (2021). Anxiety and depression in Chinese students during the COVID-19 pandemic: a meta-analysis. *Frontiers in public health*, 9, 697642.
3. Hasanah, U., Fitri, N. L., Supardi, S., & PH, L. (2020). Depression among college students due to the COVID-19 pandemic. *Jurnal Keperawatan Jiwra*, 8(4), 421-424.
4. Holm-Hadulla, R. M., Wendler, H., Baracsi, G., Storck, T., Möltner, A., & Herpertz, S. C. (2023). Depression and social isolation during the COVID-19 pandemic in a student population: the effects of establishing and relaxing social restrictions. *Frontiers in Psychiatry*, 14, 1200643.
5. Qasrawi, R., Polo, S. P. V., Al-Halawa, D. A., Hallaq, S., & Abdeen, Z. (2022). Assessment and prediction of depression and anxiety risk factors in schoolchildren: machine learning techniques performance analysis. *JMIR formative research*, 6(8), e32736.
6. Liu, X. Q., Guo, Y. X., Zhang, W. J., & Gao, W. J. (2022). Influencing factors, prediction and prevention of depression in college students: a literature review. *World journal of psychiatry*, 12(7), 860.
7. Mutalib, S., Shafiee, N. S. M., & Abdul-Rahman, S. (2021). Mental health prediction models using machine learning in higher education institution. *Turkish Journal of Computer and Mathematics Education*, 12(5), 1782-1792.
8. Rani, R., & Gupta, S. (2024, November). Predicting student anxiety and depression using random forest classifiers optimizer. In *2024 Second International Conference Computational and Characterization Techniques in Engineering & Sciences (IC3TES)* (pp. 1-5). IEEE.
9. Huamán-Romani, Y. L., Roque-Tito, E., Bautista-López, L., & Gutiérrez-Aguilar, M. D. (2021, December). Level of depression of college students with binary logistic regression model approximation in Covid-19 times. In *2021 IEEE 1st International Conference on Advanced Learning Technologies on Education & Research (ICALTER)* (pp. 1-4). IEEE.
10. Malik, S. S., & Khan, A. (2023, April). Anxiety, depression and stress prediction among college students using machine learning algorithms. In *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)* (pp. 1-5). IEEE.
11. Iparraguirre-Villanueva, O., Paulino-Moreno, C., Epifanía-Huerta, A., & Torres-Ceclén, C. (2024). Machine Learning Models to Classify and Predict Depression in College Students. *International Journal of Interactive Mobile Technologies*, 18(14).
12. Diwaker C., Tomar P., Solanki A., Nayyar A., Jhanjhi N.Z., Abdullah A., Supramaniam M.A New Model for Predicting Component-Based Software Reliability Using Soft Computing(2019) IEEE Access, 7, art. no. 8864075, pp. 147191 - 147203, DOI: 10.1109/ACCESS.2019.2946862
13. Kok S.H., Abdullah A., Jhanjhi N.Z., Supramaniam M.A review of intrusion detection system using machine learning approach(2019) International Journal of Engineering Research and Technology, 12 (1),pp. 8 - 15,
14. Sindiramutty S.R., Jhanjhi N.Z., Ray S.K., Jazri H., Khan N.A., Gaur L.Metaverse: Virtual meditation(2023) Metaverse Applications for Intelligent Healthcare, pp. 93 - 158, DOI: 10.4018/978-1-6684-9823- 1.ch003
15. Mushtaq M., Ullah A., Ashraf H., Jhanjhi N.Z., Masud M., Alqhatani A., Alnfai M.M. Anonymity Assurance Using Efficient Pseudonym Consumption in Internet of Vehicles (2023) Sensors, 23 (11), art. no. 5217,DOI: 10.3390/s23115217
16. Ahmed, Q. W., Garg, S., Rai, A., Ramachandran, M., Jhanjhi, N. Z., Masud, M., & Baz, M. (2022). Ai-based resource allocation techniques in wireless sensor internet of things networks in energy efficiency with data optimization. *Electronics*, 11(13), 2071.
17. Khan, N. A., Jhanjhi, N. Z., Brohi, S. N., Almazroi, A. A., & Almazroi, A. A. (2022). A secure communication protocol for unmanned aerial vehicles. *CMC-Computers Materials & Continua*, 70(1), 601-618.
18. Muzafar, S., & Jhanjhi, N. Z. (2020). Success stories of ICT implementation in Saudi Arabia. In *Employing Recent Technologies for Improved Digital Governance* (pp. 151-163). IGI Global Scientific Publishing.

19. Jabeen, T., Jabeen, I., Ashraf, H., Jhanjhi, N. Z., Yassine, A., & Hossain, M. S. (2023). An intelligent healthcare system using IoT in wireless sensor network. *Sensors*, 23(11), 5055.
20. Shah, I. A., Jhanjhi, N. Z., & Laraib, A. (2023). Cybersecurity and blockchain usage in contemporary business. In *Handbook of Research on Cybersecurity Issues and Challenges for Business and FinTech Applications* (pp. 49-64). IGI Global.
21. Hanif, M., Ashraf, H., Jalil, Z., Jhanjhi, N. Z., Humayun, M., Saeed, S., & Almuhaideb, A. M. (2022). AI-based wormhole attack detection techniques in wireless sensor networks. *Electronics*, 11(15), 2324.
22. Shah, I. A., Jhanjhi, N. Z., Amsaad, F., & Razaque, A. (2022). The role of cutting-edge technologies in industry 4.0. In *Cyber Security Applications for Industry 4.0* (pp. 97-109). Chapman and Hall/CRC.
23. Humayun, M., Almufareh, M. F., & Jhanjhi, N. Z. (2022). Autonomous traffic system for emergency vehicles. *Electronics*, 11(4), 510.
24. Muzammal, S. M., Murugesan, R. K., Jhanjhi, N. Z., & Jung, L. T. (2020, October). SMTrust: Proposing trust-based secure routing protocol for RPL attacks for IoT applications. In *2020 International Conference on Computational Intelligence (ICCI)* (pp. 305-310). IEEE.
25. Brohi, S. N., Jhanjhi, N. Z., Brohi, N. N., & Brohi, M. N. (2023). Key applications of state-of-the-art technologies to mitigate and eliminate COVID-19. *Authorea Preprints*.
26. Khalil, M. I., Humayun, M., Jhanjhi, N. Z., Talib, M. N., & Tabbakh, T. A. (2021). Multi-class segmentation of organ at risk from abdominal ct images: A deep learning approach. In *Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS 2021* (pp. 425-434). Singapore: Springer Nature Singapore.
27. Humayun, M., Jhanjhi, N. Z., Niazi, M., Amsaad, F., & Masood, I. (2022). Securing drug distribution systems from tampering using blockchain. *Electronics*, 11(8), 1195.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.