International Journal of
Applied Data Science in Engineering and Health
https://ijadseh.com

Predicting Mental Health Outcomes: A Machine Learning Approach to Depression, Anxiety, and Stress

Faezeh Norouzi[1,*], Bruna Lino Modes Santos Machado[2]

[1] *Psychiatry and Behavioral Science, Isfahan University of Medical Science, Isfahan, Iran (Fanorouzi8@gmail.com)*

[2] *Instituto Luterano de Ensino Superior de Itumbiara, Brazil*

Received date: September 23, 2024; Accepted date: October 30, 2024

Abstract

Depression, anxiety, and stress are prevalent mental health disorders with profound effects on individuals and society. Early and accurate predictions of these conditions can significantly improve treatment outcomes. In this study, we applied three machine learning models—K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest—to predict severity levels of these disorders based on Depression Anxiety Stress Scales (DASS) responses. Among the models, SVM demonstrated the highest performance, achieving 99% accuracy across all datasets, followed closely by Random Forest, particularly on the depression and stress datasets. These results highlight machine learning's potential in enhancing mental health diagnostics, with SVM proving the most effective for accurate classification.

Keywords*:* Depression; Anxiety; Stress; Machine Learning.

Introduction

Depression, anxiety, and stress are prevalent mental health conditions, each with unique characteristics but often overlapping in their effects on well-being. Depression is marked by persistent sadness, a sense of hopelessness, and a loss of interest in activities once enjoyed. It affects mood, energy, sleep, appetite, and concentration and, in severe cases, can lead to suicidal thoughts, making it a leading cause of global disability. Anxiety, in contrast, is characterized by excessive worry, fear, or nervousness that may interfere with daily functioning. People with anxiety often experience physical symptoms such as a racing heart, muscle tension, and restlessness, which can lead to avoidance of situations that trigger these feelings. Stress, while a natural response to challenging situations, can become harmful when prolonged or intense, affecting physical and mental health by straining the body's adaptive systems. Chronic stress, often linked to work, relationships, or financial concerns, can weaken immunity and heighten vulnerability to both depression and anxiety, creating a complex cycle of emotional distress. Addressing and healing these conditions is essential for enhancing individual well-being and societal productivity. [1-3]

Machine learning (ML) is a branch of artificial intelligence that enables computers to learn from data patterns without explicit programming, making it highly valuable for complex data analysis across many medical fields, including radiology [4-6], diagnostic [7-9], treatment [10-12], and mental health [13-16]. In mental health care, ML can process large and varied datasets—such as clinical records, wearable device data, and self-reported assessments—to identify patterns that would be challenging or time-consuming for human analysis alone. By learning from these patterns, ML algorithms can develop predictive models that offer insights into a person's mental health trajectory, potentially identifying subtle risk factors for conditions like depression, anxiety, and stress [17-19]. In terms of practical applications, ML can predict an individual's likelihood of developing severe symptoms based on behavior changes, social interaction frequency, or even physiological data like sleep and heart rate [20]. It can also assist in personalizing treatment approaches [21], matching patients to therapies that have the highest likelihood of effectiveness for their specific needs. Additionally, machine learning models are being used to develop digital tools for real-time mental health monitoring [22], enabling clinicians to track patient progress and adjust treatments as necessary. By enhancing diagnostic accuracy, offering predictive capabilities, and personalizing interventions, ML has the potential to transform the way mental health conditions are managed, ensuring proactive, data-driven care that can improve outcomes for individuals experiencing depression, anxiety, and stress.

In this paper, we develop predictive models for depression, anxiety, and stress using three machine learning algorithms: K-nearest neighbors (KNN), Support Vector Machine (SVM), and Random Forest. Each of these algorithms offers unique advantages in handling diverse data types and complexities associated with mental health prediction. KNN is effective for analyzing proximity-based relationships within the data, SVM is known for its robust classification capability even with high-dimensional data, and Random Forest provides ensemble learning benefits, reducing overfitting and enhancing prediction accuracy. By comparing these models, we aim to identify the most effective approach for accurately predicting mental health outcomes and to provide insights into the application of machine learning for improving mental health diagnosis and management.

Methods and Materials

Dataset

In this study, we utilize a dataset comprising responses to the Depression Anxiety Stress Scales (DASS), a widely used psychological assessment tool. Collected from 39,775 participants, this dataset includes detailed questions, responses, and relevant metadata, providing a rich source of information for predicting mental health outcomes related to depression, anxiety, and stress. The dataset was hosted on OpenPsychometrics.org, a nonprofit organization dedicated to public psychology education and research data collection. The extensive data size and range of responses enable robust machine learning model training, aiming to enhance the predictive capabilities of our approach in identifying key patterns and risk factors in mental health assessments. [23]

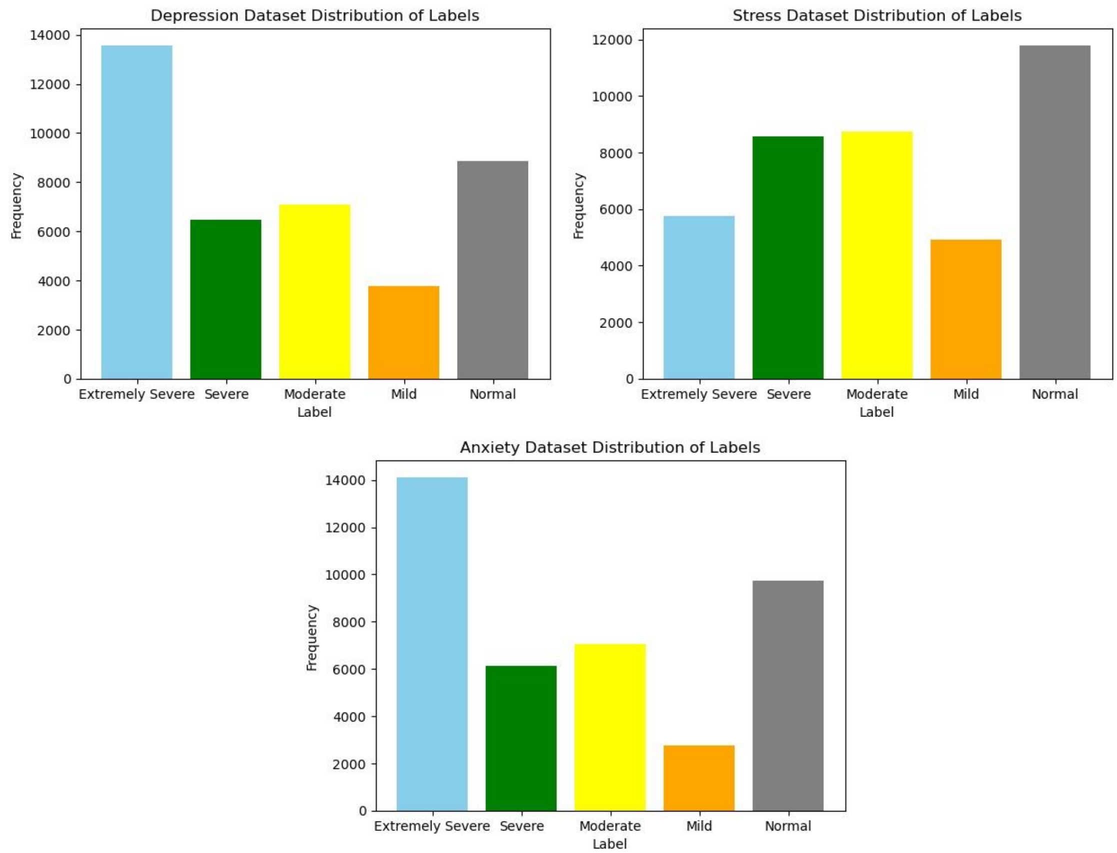

Figure 1 Dataset distribution [23]

Figure 1 illustrates the distribution of severity labels for depression, anxiety, and stress within the dataset, each dataset showing unique patterns that reflect variations in symptom severity among participants. In Depression, there is a clear skew towards the Extremely Severe category, with nearly 14,000 cases, indicating that a substantial portion of respondents reported severe depressive symptoms. The Normal category follows with approximately 9,000 cases, representing individuals without depressive symptoms. The Severe and Moderate categories are both well-represented, with around 6,000 cases each, while the Mild category has the fewest cases, at around 4,000. This imbalance, with a high concentration in the Extremely Severe and Normal categories, could potentially impact the performance of machine learning models by making it more challenging to accurately distinguish between intermediate severity levels. In contrast, the Stress shows a more balanced distribution across severity levels. The Normal category has the highest frequency, with nearly 12,000 cases, suggesting that a large portion of respondents report low-stress levels. The Severe and Moderate categories are close in frequency, each with around 9,000 cases, reflecting a substantial population with moderate to high stress. Meanwhile, the Extremely Severe and Mild categories are less frequent, each below 7,000, indicating fewer participants at the extreme ends of the stress scale. This distribution provides a broader range of severity levels for model training, which could aid in developing models capable of more nuanced classification between stress levels. Anxiety mirrors the depression dataset’s pattern, with a pronounced skew towards the Extremely Severe category, which includes around 14,000 cases, and the Normal category, with approximately 9,000 cases. The intermediate Severe and Moderate categories each have a moderate number of cases, while Mild is the least represented.

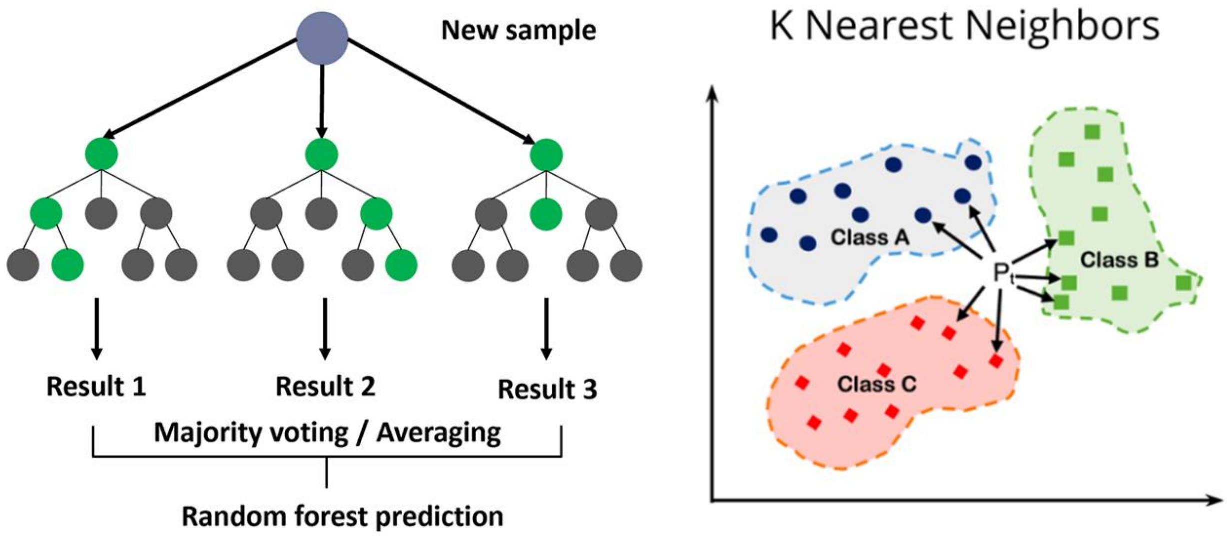

Figure 2 left: Random Forest [24], right: KNN [25]

This distribution, heavily weighted toward high and low extremes, presents similar challenges to those in the depression dataset, as models may struggle to differentiate intermediate levels due to the smaller representation of mild and moderate cases.

Overall, these datasets present unique distribution patterns, with both the depression and anxiety datasets showing pronounced imbalances toward the Extremely Severe and Normal categories, while the stress dataset displays a more even distribution across severity levels. This variability in label distribution will likely influence the performance of predictive models, potentially necessitating techniques like class balancing or weighting adjustments to ensure accurate predictions across all severity categories.

Machine Learning Models

- K-Nearest Neighbors (KNN): KNN is a non-parametric, instance-based learning algorithm that classifies a data point based on the majority label of its 'k' nearest neighbors, where 'k' is a user-defined parameter. The algorithm calculates the distance between points, commonly using Euclidean distance, though other distance metrics (such as Manhattan or Minkowski) can be applied depending on the dataset characteristics. KNN’s simplicity lies in its lazy learning approach—it stores all training instances and makes predictions only at the time of classification. This method allows it to adapt to complex decision boundaries, but it is computationally intensive for large datasets as it requires calculating distances to all training points for each new prediction. Additionally, KNN can be sensitive to the choice of 'k' and the scale of features, which often necessitates normalization or standardization.

- Support Vector Machine (SVM): SVM is a supervised learning algorithm that aims to find an optimal hyperplane that maximally separates classes in the feature space. In its simplest form, SVM is a linear classifier; however, it can handle non-linear separations using the "kernel trick," where data is mapped into a higher-dimensional space to find a linear separation in this transformed space. Common kernels include linear, polynomial, radial basis function (RBF), and sigmoid, each providing different capabilities to capture complex relationships. SVM maximizes the margin—the distance between the hyperplane and the nearest data points from each class, known as support vectors—enhancing generalization. The choice of regularization parameter (C) and kernel parameters, such as gamma for the RBF kernel, are critical in balancing the trade-off between maximizing the margin and minimizing classification error, especially in cases where data is not linearly separable.

- Random Forest: Random Forest is an ensemble learning algorithm that combines the predictions of multiple decision trees to improve accuracy and reduce overfitting. Each tree is trained on a bootstrap sample of the dataset, and at each node, a random subset of features is considered for splitting, promoting diversity among the trees. This approach helps Random Forest to be resilient against noise and provides reliable predictions even in high-dimensional, complex datasets. The number of trees (n_estimators) and the maximum depth of each tree are hyperparameters that impact the model's performance and computational efficiency. Random Forest also provides feature importance scores by measuring how much each feature decreases impurity across the forest, enabling insight into the contribution of individual variables in the classification process. This robustness and interpretability make Random Forest particularly useful for applications requiring both accuracy and explainability.

Results

The four metrics—accuracy, precision, recall, and F1-score—each provide a unique perspective on model performance, particularly in classification tasks. Accuracy measures the proportion of correctly classified instances out of all predictions, offering a general overview of a model's effectiveness. However, accuracy alone can be misleading, especially with imbalanced datasets, where high accuracy may not indicate strong performance across all classes. Precision is the ratio of true positive predictions to the total positive predictions, reflecting the model's ability to avoid false positives. This metric is crucial when the cost of false positives is high, as it ensures that predictions labeled as positive are indeed correct. Recall (or sensitivity) measures the model's capacity to identify all relevant cases by calculating the ratio of true positives to the total actual positives, focusing on minimizing false negatives. High recall is essential when it's critical to capture all positive cases, even if it means including some false positives. F1-score combines precision and recall into a single metric by taking their harmonic mean, offering a balanced measure especially useful when precision and recall are equally important, as it accounts for both false positives and false negatives. Together, these metrics give a comprehensive view of model performance, allowing for nuanced evaluations of classification effectiveness across datasets.

The results for each disorder highlight varying levels of effectiveness for the models used: KNN, SVM, and Random Forest. On the Depression dataset, SVM demonstrated the highest performance, achieving near-perfect scores with accuracy, precision, recall, and a F1-score of 99%. This indicates that SVM effectively classified depression severity with minimal misclassification. KNN and Random Forest also performed well on the depression dataset, each reaching an accuracy of 93% and similar precision, recall, and F1 scores (around 90%). While these results are strong, they fall short of SVM's performance, likely due to SVM's strength in handling complex boundaries between classes. For the Stress dataset, SVM again outperformed the other models, achieving 99% across all metrics, indicating high reliability in stress classification.

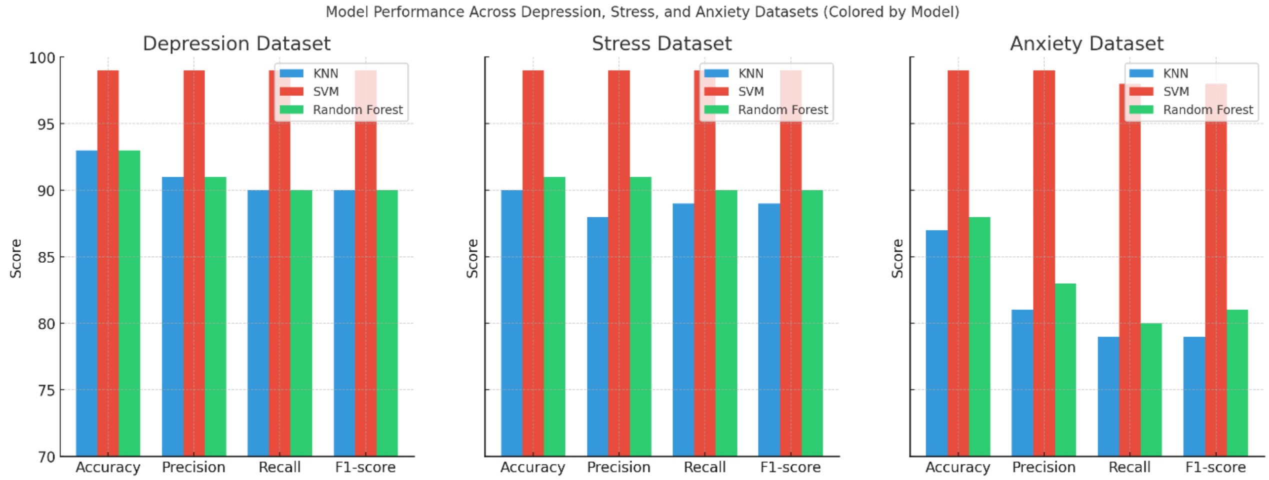

Figure 3 Performance metrics of each model for each dataset

Random Forest followed closely with 91% accuracy and balanced metrics (precision, recall, F1-score) around 90%, showing that it was also able to capture stress severity effectively. KNN achieved slightly lower performance on the stress dataset, with an accuracy of 90%, precision of 88%, recall of 89%, and F1-score of 89%. This suggests that while KNN can effectively classify stress severity, it may be less robust than SVM and Random Forest for this task, potentially due to its sensitivity to the dataset's distribution. In the Anxiety dataset, SVM remained the top performer with an accuracy of 99%, a precision of 99%, and a slightly lower recall and F1 score (98%). This minor decrease in recall suggests that SVM might have missed some instances of anxiety and severity but still maintained high overall accuracy. Random Forest performed moderately well, with an accuracy of 88%, precision of 83%, recall of 80%, and F1-score of 81%. KNN showed the lowest performance on the anxiety dataset, with 87% accuracy and an F1-score of 79%, indicating it may struggle with the nuances of anxiety classification. These results highlight SVM's overall superiority across all three datasets, likely due to its margin-maximizing capability, which enhances its precision and generalization. Random Forest also performed well, particularly for stress and depression, due to its ensemble nature, while KNN was the least effective, particularly for anxiety, possibly due to its sensitivity to noisy or overlapping data points.

Conflict of interest

The authors declared no conflict of interest.

References

[1] Howard LM, Khalifeh H. Perinatal mental health: a review of progress and challenges. World Psychiatry. 2020 Oct;19(3):313-27.

[2] Bertolote J. The roots of the concept of mental health. World psychiatry. 2008 Jun;7(2):113.

[3] Koutsimani P, Montgomery A, Georganta K. The relationship between burnout, depression, and anxiety: A systematic review and meta-analysis. Frontiers in psychology. 2019 Mar 13;10:284.

[4] Zadeh FS, Molani S, Orouskhani M, Rezaei M, Shafiei M, Abbasi H. Generative Adversarial Networks for Brain Images Synthesis: A Review. arXiv preprint arXiv:2305.15421. 2023 May 16.

[5] Yasaka K, Hatano S, Mizuki M, Okimoto N, Kubo T, Shibata E, Watadani T, Abe O. Effects of deep learning on radiologists' and radiology residents' performance in identifying esophageal cancer on CT. The British Journal of

Radiology. 2023 Oct 1;96(1150):20220685.
[6] Liu C, Tian Y, Song Y. A systematic review of deep learning-based research on radiology report generation. arXiv preprint arXiv:2311.14199. 2023 Nov 23.
[7] Jiao T, Li F, Cui Y, Wang X, Li B, Shi F, Xia Y, Zhou Q, Zeng Q. Deep learning with an attention mechanism for differentiating the origin of brain metastasis using MR images. Journal of Magnetic Resonance Imaging. 2023 Nov;58(5):1624-35.
[8] Chen H, Kim S, Hardie JM, Thirumalaraju P, Gharpure S, Rostamian S, Udayakumar S, Lei Q, Cho G, Kanakasabapathy MK, Shafiee H. Deep learning-assisted sensitive detection of fentanyl using a bubbling-microchip. Lab on a Chip. 2022;22(23):4531-40.
[9] Degadwala S, Krishnamurthy VN, Vyas D. DeepSpine: Multi-Class Spine X-Ray Conditions Classification Using Deep Learning. In2024 3rd International Conference on Sentiment Analysis and Deep Learning (ICSADL) 2024 Mar 13 (pp. 8-13). IEEE.
[10] Ahmad A, Tariq A, Hussain HK, Gill AY. Revolutionizing Healthcare: How Deep Learning is poised to Change the Landscape of Medical Diagnosis and Treatment. Journal of Computer Networks, Architecture and High Performance Computing. 2023 Jul 11;5(2):458-71.
[11] Chen X. AI in Healthcare: Revolutionizing Diagnosis and Treatment through Machine Learning. MZ Journal of Artificial Intelligence. 2024 Jul 24;1(2).
[12] HaghighiKian SM, Shirinzadeh-Dastgiri A, Vakili-Ojarood M, Naseri A, Barahman M, Saberi A, Rahmani A, Shiri A, Masoudi A, Aghasipour M, Shahbazi A. A Holistic Approach to Implementing Artificial Intelligence in Lung Cancer. Indian Journal of Surgical Oncology. 2024 Sep 5:1-22.
[13] Abd Rahman R, Omar K, Noah SA, Danuri MS, Al-Garadi MA. Application of machine learning methods in mental health detection: a systematic review. Ieee Access. 2020 Oct 6;8:183952-64.
[14] Milne-Ives M, Selby E, Inkster B, Lam C, Meinert E. Artificial intelligence and machine learning in mobile apps for mental health: A scoping review. PLOS digital health. 2022 Aug 15;1(8):e0000079.
[15] Srividya M, Mohanavalli S, Bhalaji N. Behavioral modeling for mental health using machine learning algorithms. Journal of medical systems. 2018 May;42:1-2.
[16] Jain T, Jain A, Hada PS, Kumar H, Verma VK, Patni A. Machine learning techniques for prediction of mental health. In2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) 2021 Sep 2 (pp. 1606-1613). IEEE.
[17] Gil M, Kim SS, Min EJ. Machine learning models for predicting risk of depression in Korean college students: identifying family and individual factors. Frontiers in Public Health. 2022 Nov 17;10:1023010.
[18] Qasrawi R, Polo SP, Al-Halawa DA, Hallaq S, Abdeen Z. Assessment and prediction of depression and anxiety risk factors in schoolchildren: machine learning techniques performance analysis. JMIR formative research. 2022 Aug 31;6(8):e32736.
[19] Li H, Li Y, Duan Y, Wang S, Liu M, Luo Y, Wang J, Chen Z, Yang P, Xie J. Risk factors for depression in China based on machine learning algorithms: A cross-sectional survey of 264,557 non-manual workers. Journal of Affective Disorders. 2024 Dec 15;367:617-22.
[20] Fiorini L, Cavallo F, Dario P, Eavis A, Caleb-Solly P. Unsupervised machine learning for developing personalised behaviour models using activity data. Sensors. 2017 May 4;17(5):1034.
[21] Boukhechba M, Baglione AN, Barnes LE. Leveraging mobile sensing and machine learning for personalized mental health care. Ergonomics in design. 2020 Oct;28(4):18-23.
[22] VENKATARAMANAIAH B, REDDY KG, RAMAKRISHNAN K, SATHISH M, RAJESH S. Real Time Mental Health Monitoring System using Machine Learning. In2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI) 2024 Aug 28 (pp. 806-811). IEEE.
[23] https://www.kaggle.com/datasets/lucasgreenwell/depression-anxiety-stress-scales-responses/data
[24] https://medium.com/@roiyeho/random-forests-98892261dc49
[25] https://medium.com/@sachinsoni600517/k-nearest-neighbours-introduction-to-machine-learning-algorithms-9dbc9d9fb3b2