

Article

Not peer-reviewed version

Digital Behavior and Mental Health Prediction Through Explainable AI

[Talin Arya](#) *

Posted Date: 30 December 2025

doi: 10.20944/preprints202512.2611.v1

Keywords: mental health; digital behavior; reproducibility; explainable AI; psychology; XGBoost; SHAP



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

Digital Behavior and Mental Health Prediction Through Explainable AI

Talin Arya

School of Analytics, Harrisburg University of Science and Technology; tarya@my.harrisburgu.edu

Abstract

Digital habits such as screen time, notifications, and social media engagement have increasingly influenced mental health and overall well-being. This research examines the link between these digital practices and mental health outcomes by utilizing an explainable AI framework from a public dataset containing 500 anonymized entries that combine behavioral metrics with self-reported measures. Building on initial logistic regression analyses, this study employs gradient boosting with XGBoost, enhanced by Shapley Additive Explanations (SHAP), to strengthen both predictive accuracy and interpretability. To evaluate reproducibility, models were trained with five random seeds, and performance was assessed using root mean square error (RMSE) and the coefficient of determination (R^2). The outcomes demonstrated consistent predictive performance (RMSE ranged from 5.8 to 6.8; R^2 ranged from 0.25 to 0.31) and consistently highlighted sleep hours, notification count, and focus score as the most significant predictors. SHAP analysis revealed low variance across seeds, reaffirming the reliability of these features. These findings highlight how behavioral data can inform digital wellness initiatives. This research contributes to a transparent, reproducible analytical framework that bridges the gap between computational modelling and psychological research, supporting the application of explainable AI in mental health research.

Keywords: mental health; digital behavior; reproducibility; explainable AI; psychology; XGBoost; SHAP

1. Introduction

Mental health and wellness have increasingly become important topics for research in both academic and public fields, with stress, anxiety, and mood-related disorders having become increasingly common on a global scale. Mental health conditions represent a significant health concern, affecting quality of life and productivity. At the same time, daily life has become deeply intertwined with digital technologies, creating an environment where mental well-being is influenced not only by traditional social and biological factors but also by patterns of digital behavior.

The widespread adoption of smartphones, social media platforms, and on-the-go digital services has shaped how individuals communicate, work, and manage their tasks. Although these technologies provide undeniable benefits, including enhanced social connections, access to information, and flexible work opportunities, growing evidence suggests that excessive or poor management of digital engagement impacts mental health. Understanding how specific digital behaviors relate to mental well-being has therefore become a critical research priority.

A substantial body of literature has linked increased screen time and digital media use to adverse mental health outcomes such as anxiety, depression symptoms, loneliness and reduced subjective well-being (Twenge & Campbell, 2018; Przybylski & Weinstein, 2017). Sleep disruption represents a crucial mechanism in this relationship. Late-night screen exposure and mobile device usage have been associated with shorter sleep duration, poor sleep quality and elevated stress levels (Hale & Guan, 2015; Carter et al., 2016; Levenson et al., 2017). These sleep-related effects are concerning given the well-established relationships between sleep quality and emotional regulation, cognitive functioning, and mental health.

Beyond total screen time, patterns of digital interaction play a critical role. Engagement with social media platforms has been associated with increased anxiety, depressive symptoms, and perceived social isolation, particularly when usage is characterized by passive consumption or frequent social comparison (Primack et al., 2017; Keles et al., 2020). Additionally, other behaviors like frequent app switching, notifications count, and interruptions have been shown to increase cognitive load, low attention span, and elevated stress levels (Mark et al., 2007; Kushlev et al., 2016; Stothart et al., 2015). These findings suggest that these small behavior patterns of digital engagement can be important indicators of mental health outcomes.

Despite these findings, much of the existing research relies heavily on self-reported measures of both digital behavior and psychological outcomes. While self-reported findings are valuable, they are subject to bias and temporal resolution. Moreover, many prior studies have employed traditional statistical approaches that may have struggled to capture nonlinear relationships and interactions among behavioral variables. This leads to reproducibility and interpretability challenges in research.

In recent years, dependence on machine learning methods has increased for mental health prediction tasks, offering the potential to model complex relationships within behavioral data. However, some studies have emphasized predictive accuracy at the expense of model transparency and reproducibility. This limits clinical trust and raises concerns about the reliability and generalizability of reported findings, particularly in sensitive domains such as mental health.

To address these gaps, this research integrates gradient boosting using XGBoost with Shapley Additive Explanations (SHAP) to examine the relationship between digital behaviors and mental health outcomes by evaluating model performance across multiple random seeds, this study assesses reproducibility and stability of feature importance estimates. This approach allows for both predictive performance and transparent interpretation of how specific digital behaviors contribute to mental health predictions.

This research provides valuable insights. It identifies key digital behavioral predictors: Sleep duration, notification count, and focus related measures that are associated with mental health outcomes. It demonstrates reproducible results by examining stability across multiple runs. As this study tries to bridge the gap between prior research and move closer to real-world applications, it addresses the following research question: **Which digital behaviors are associated with mental health outcomes, and how consistent are these associations across different model runs?**

2. Literature Review

Sleep has long been recognized as a crucial indicator of both cognitive performance and emotional regulation. Research demonstrates that insufficient sleep impairs executive functioning, decision making, and emotional control (Killgore, 2010). These impairments are relevant to mental health since they are closely linked to mood instability, stress, and reduced coping capacity. Long-term sleep deprivation has also been linked with psychological consequences. (Walker 2017) highlights sleep’s central role in emotional resilience, showing that sleep disruption contributes to mood disorders and diminished psychological well-being. Meta-analytic evidence further confirms that sleep loss produces cumulative, dose-response effects on performance (Van Dongen et al., 2003) and widespread declines in cognitive functioning (Pilcher & Huffcutt, 1996). Together, these findings establish sleep as a fundamental component of psychological well-being and justify its inclusion as a key predictor in mental health research.

Digital behavior has also been shown to play an important role in shaping mental health outcomes. Prior studies indicate that frequent digital interruptions can increase stress levels, even when task efficiency appears to improve (Mark et al., 2008). Smartphone notifications have been found to disrupt sustained attention and increase cognitive demand, contributing to mental fatigue and stress (Kushlev et al., 2016; Stothart et al., 2015). Research on media multitasking further suggests that individuals who frequently switch between applications exhibit weaker cognitive control and reduced attentional capacity (Ophir et al., 2009). In addition, overall screen time has been associated with mental health in a nonlinear manner, with moderate use linked to well-being and excessive use

associated with negative outcomes, often described as a “Goldilocks effect” (Przybylski & Weinstein, 2017).

Focus and attention are equally relevant when examining the relationship between digital behavior and mental health. Theoretical frameworks of attention highlight the role of attentional networks in emotional regulation and self-control (Posner & Rothbart, 2007). Individuals with stronger working memory capacity are better able to resist distractions and maintain goal-directed behavior (Kane & Engle, 2002). Attentional lapses, by contrast, have been associated with reduced cognitive performance and increased vulnerability to stress (Robertson et al., 1997). These findings support the inclusion of focus-related measures as potential protective factors in predictive models of mental health.

More recently, concerns regarding transparency and reproducibility have gained prominence in machine learning research. XGBoost has emerged as an effective method for modeling complex relationships in tabular data due to its efficiency and ability to capture nonlinear interactions (Chen & Guestrin, 2016). SHAP provides a unified framework for interpreting model predictions by assigning feature contributions at both global and local levels (Lundberg & Lee, 2017). Calls for interpretable machine learning emphasize the importance of human-understandable explanations, particularly in high-stakes domains such as mental health (Doshi-Velez & Kim, 2017). Reproducibility has also been identified as a critical issue, as inconsistent results across model runs undermine confidence in computational findings (Pineau et al., 2021). Prior work demonstrates that reproducibility can be systematically improved through controlled experimental design and repeated evaluations (Bouthillier et al., 2019).

Despite these advances, relatively few studies integrate behavioral digital data with interpretable machine learning approaches while explicitly addressing reproducibility. This gap motivates the research’s focus on combining XGBoost with SHAP and evaluating model stability across multiple runs.

3. Methods

3.1. Dataset and Variables

The dataset used in this study was obtained from a publicly available Kaggle repository and consists of approximately 500 anonymized observations. Each entry includes both behavioral metrics and self-reported psychological measures. Behavioral variables include **sleep duration**, **notification count**, **social media usage time**, **screen-to-sleep ratio**, **app switches**, and **total daily screen time**. Self-reported measures include **mood score** and **focus score**. This combination allows for the examination of objective digital behavior alongside subjective mental health indicators.

3.2. Preprocessing

Data preprocessing involved handling missing values through imputation and normalizing continuous variables to ensure comparability across participants. Categorical variables were encoded where required. Outliers were reviewed but retained, as extreme values may reflect meaningful variation in digital behavior rather than data errors.

One self-reported variable, anxiety level, was excluded prior to modeling. Initial analyses indicated that anxiety level acted as a dominant predictor and overshadowed the influence of behavioral features. Since the goal of this study was to evaluate the contribution of digital behaviors rather than self-reported psychological states, anxiety level was removed to improve interpretability and maintain focus on behavioral predictors.

3.3. Modeling Approach

The modeling framework employed gradient boosting using XGBoost, which is well suited for tabular data and capable of capturing nonlinear relationships and feature interactions. The dataset

was split into training, validation, and test sets using a 60:20:20 ratio. The training set was used for model fitting, the validation set for hyperparameter tuning, and the test set for final evaluation.

Hyperparameter tuning was conducted using Optuna, a Bayesian optimization framework that efficiently explores the parameter space. GPU acceleration was enabled to improve computational efficiency during training.

3.4. Reproducibility

To assess reproducibility, the entire modeling pipeline was executed using five different random seeds. These seeds controlled data splitting, model initialization, and training procedures. Running the model across multiple seeds allowed for evaluation of stability in both predictive performance and feature importance, ensuring that results were not dependent on a single initialization.

3.5. Interpretability

Model interpretability was achieved using Shapley Additive Explanations (SHAP). SHAP provides both global explanations, showing which features consistently influence predictions across the dataset, and local explanations, illustrating how individual features contribute to specific predictions. Feature importance rankings were compared across all five random seeds to assess stability.

Key behavioral features were mapped to psychological constructs to support interpretability. Sleep duration reflects sleep hygiene and resilience, notification count represents attentional and stress load, focus score reflects sustained attention, application switching captures multitasking behavior, and screen-to-sleep ratio represents balance between digital engagement and rest.

3.6. Evaluation Metrics

Performance was evaluated using root mean squared error (RMSE) and the coefficient of determination (R^2). These metrics were calculated across all five random seeds to assess consistency in predictive accuracy. In addition, variance in SHAP values across seeds was examined to evaluate the stability of feature importance estimates.

4. Results

4.1. Model Performance

Across five random seeds, the XGBoost regression model demonstrated consistent predictive performance. As shown in Figure 1, RMSE values ranged from 5.8 to 7.0, while R^2 values ranged from 0.25 to 0.31 across model runs. Box plots of RMSE and R^2 indicate a narrow spread of values, suggesting limited variation in performance across different random initializations.

To further examine predictive behavior, scatter plots comparing actual versus predicted mental health scores and residuals versus predicted values were generated for the final seed (Figures 2 and 3). The actual-versus-predicted plot shows clustering around the ideal prediction line, while the residuals plot exhibits a symmetric distribution centered around zero, with no strong evidence of heteroscedasticity. These visualizations complement the numerical performance metrics and indicate stable model behavior across the range of predicted values.

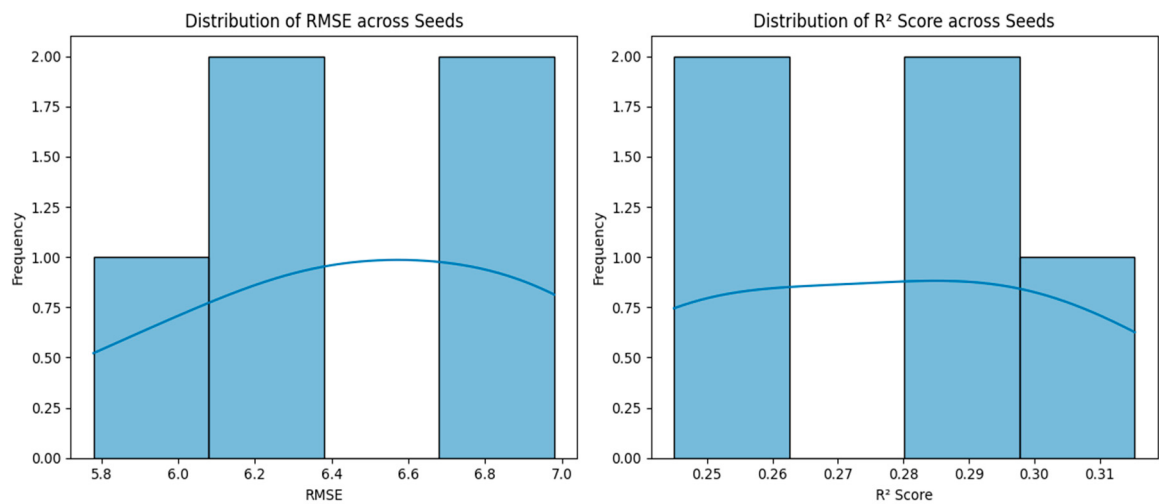


Figure 1. Box plots of RMSE and R² values across five random seeds for XGBoost model.

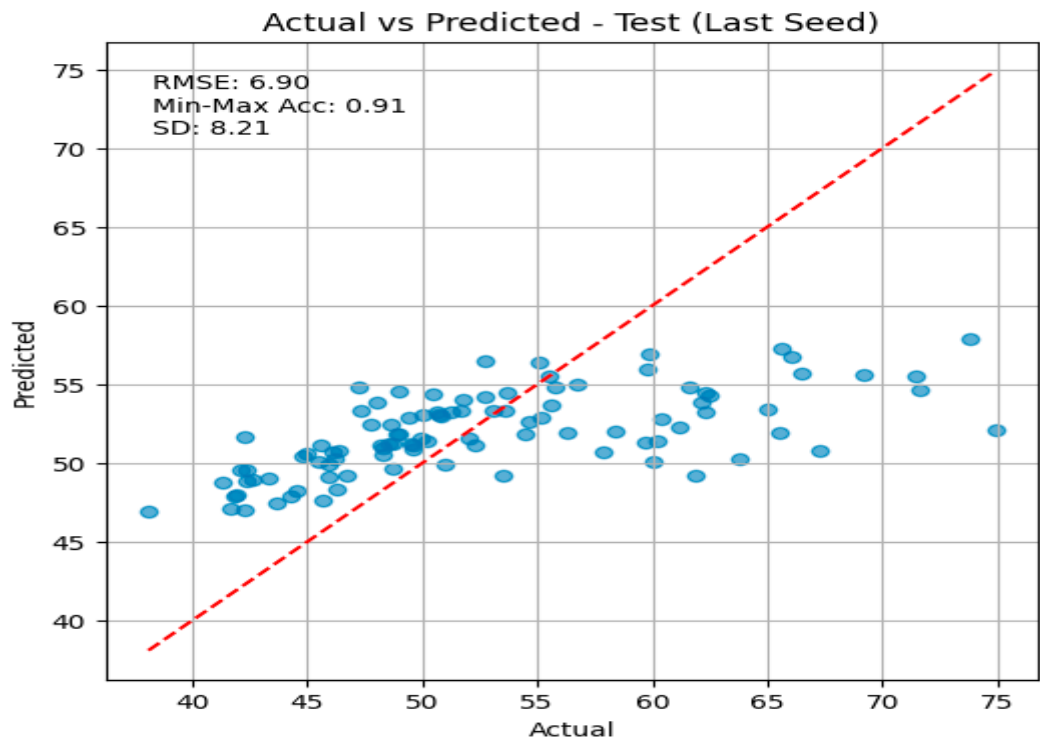


Figure 2. Scatter plot of Actual VS Predicted mental health scores.

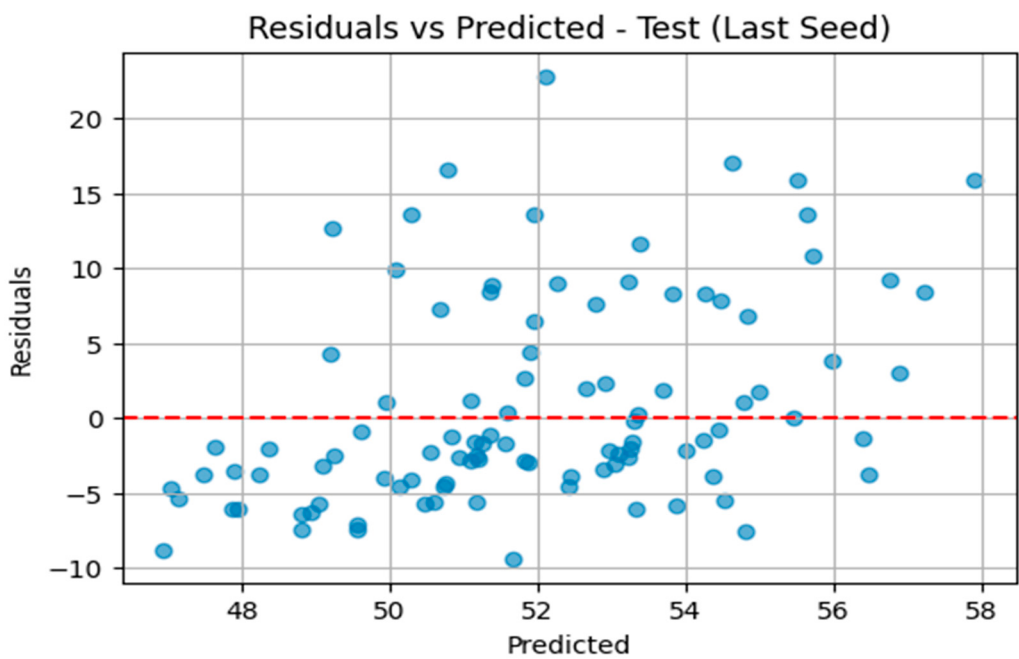


Figure 3. Residuals plotted against predicted mental health scores (Final Seed).

4.2. Feature Importance

Global interpretability analysis using SHAP summary plots identified sleep hours, notification count, and focus score as the most influential predictors of mental health outcomes (Figure 4). Higher sleep duration was associated with more positive predicted outcomes, while notification count and screen-to-sleep ratio also exhibited strong SHAP values. Other features, including application switching and mood score, displayed more dispersed SHAP values, indicating comparatively weaker or less consistent influence across participants.

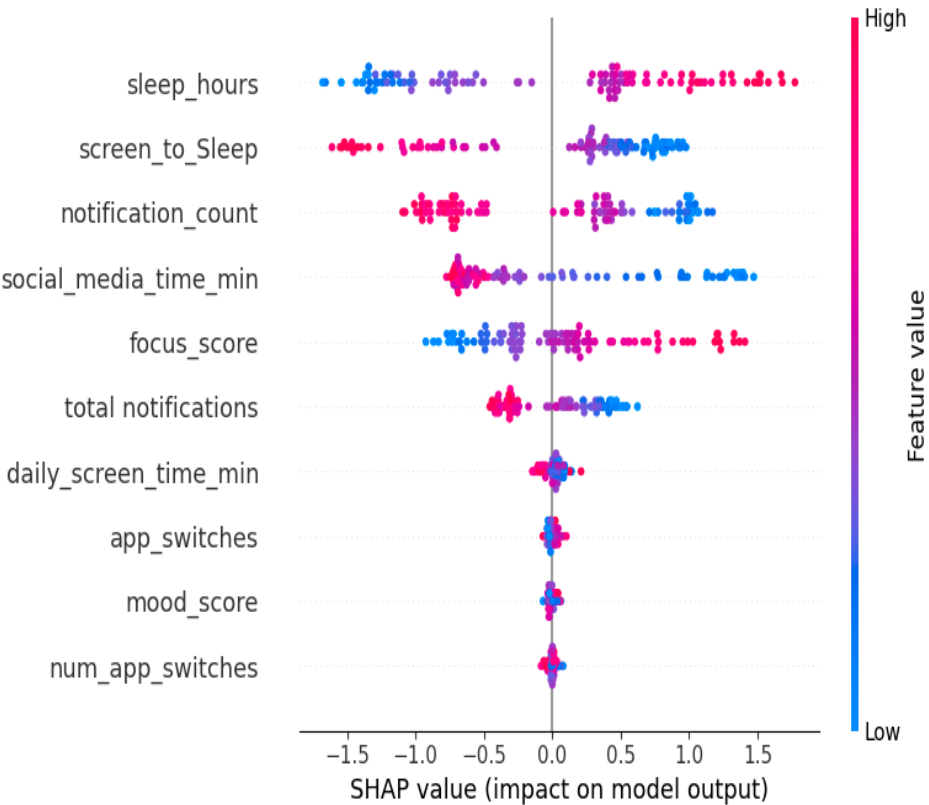


Figure 4. SHAP summary plot illustrating feature importance for mental health prediction.

4.3. Interpretability and Reproducibility

Local interpretability was examined using SHAP force plots and ExactExplainer visualizations. These plots illustrate how combinations of features contributed to individual predictions. Across samples, cases with higher sleep duration and stronger focus scores tended to exhibit more favorable predicted outcomes, even when notification counts were elevated.

Variance analysis of SHAP values across five random seeds is shown in Figure 5. Sleep hours exhibited the highest variance, indicating some sensitivity to sample composition, while screen-to-sleep ratio, focus score, and notification count demonstrated moderate variance. Other features, such as total screen time, application switching, and mood score, showed minimal variance, reflecting limited influence on model predictions.

ExactExplainer visualizations (Figure 6) display SHAP-based feature contributions across 20 samples ordered by similarity. Consistent banding patterns across samples indicate stable feature interactions for individuals with similar behavioral profiles.

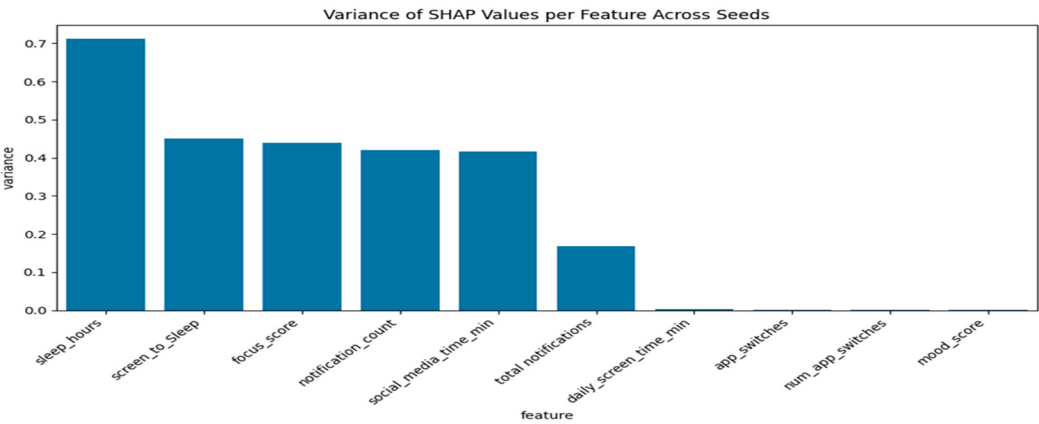


Figure 5. Variance of SHAP values for each feature across five seeds.

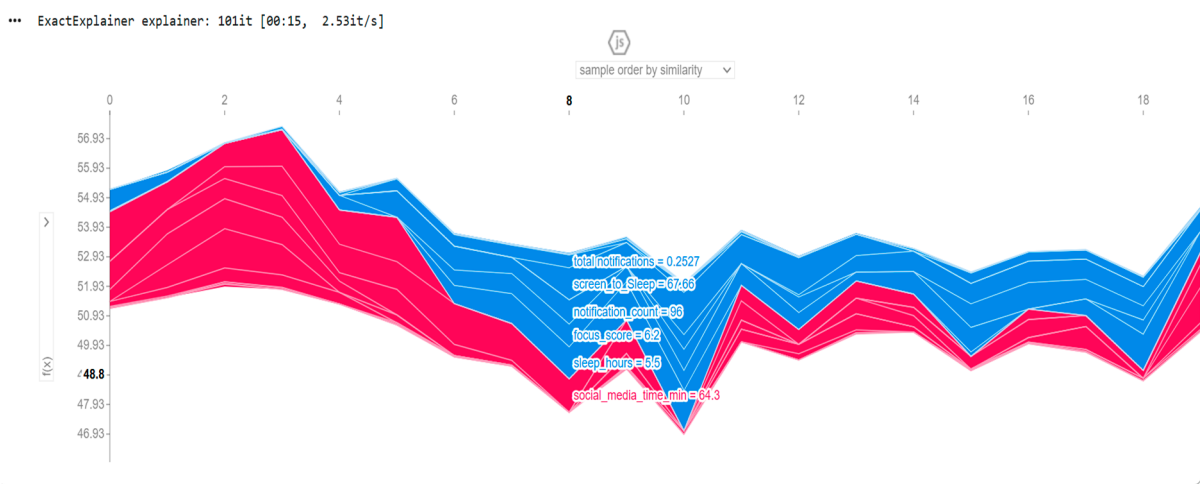


Figure 6. ExactExplainer visualization showing SHAP-based feature contributions across 18 samples, ordered by similarity.

5. Discussion

This study applied XGBoost regression with SHAP interpretability to examine digital behavioral predictors of mental health outcomes. The discussion below summarizes key findings, interprets the

psychological relevance of predictors, highlights methodological contributions related to reproducibility, and addresses limitations and ethical considerations.

5.1. Summary of Key Findings

Across five random seeds, the model demonstrated consistent predictive performance, supporting the stability of the findings. Sleep duration, notification count, and focus score emerged as the most influential predictors of mental health outcomes. SHAP analysis indicated that greater sleep duration and higher focus scores were associated with more favorable predicted outcomes, even when notification activity was high. Variance analysis showed that feature importance rankings remained stable across model runs, confirming reproducibility.

5.2. Interpretation of Digital Behavioral Predictors

Sleep duration emerged as the strongest predictor, aligning with prior research linking sleep quality to emotional regulation, cognitive functioning, and psychological well-being. The negative association between notification count and mental health outcomes supports existing findings on digital overload and attentional fragmentation. Frequent notifications may increase cognitive demand and stress by disrupting sustained attention. Focus score contributed positively, consistent with research on attentional control and self-regulation, suggesting that individuals with stronger focus may better manage digital distractions.

5.3. Reproducibility and Methodological Contribution

A key contribution of this research is its emphasis on reproducibility. By training models across multiple random seeds and examining both performance metrics and SHAP values, the study demonstrates that interpretability results are stable and not dependent on a single model run. This approach addresses concerns regarding transparency and reliability in machine learning research and supports more trustworthy application of computational methods in mental health studies.

5.4. Limitations

Several limitations should be acknowledged. First, the dataset was modest in size and relied on self-reported measures, which may introduce bias and measurement error. Second, the cross-sectional design limits causal inference, making it difficult to determine whether digital behaviors influence mental health or vice versa. Third, the model explained a moderate proportion of variance in outcomes, reflecting the multifactorial nature of mental health. Finally, the dataset was sourced from a public repository, which may limit generalizability to broader populations.

5.5. Ethical Considerations

Ethical considerations are important when applying machine learning to mental health data. Digital behaviors may vary across demographic groups, and models trained on non-representative data risk biased predictions. Transparency is therefore critical. Explainable AI methods such as SHAP support human oversight by allowing researchers and practitioners to understand how predictions are generated. Predictive models should be used to complement, rather than replace, professional judgment within appropriate ethical and clinical frameworks.

6. Conclusion

This study examined the relationship between digital behavioral patterns and mental health outcomes using an interpretable machine learning framework. By integrating XGBoost regression with Shapley Additive Explanations (SHAP), the research aimed to balance predictive performance with transparency while explicitly addressing reproducibility across multiple model runs. The results demonstrated consistent performance across five random seeds and identified sleep duration,

notification count, and focus-related measures as the most influential behavioral predictors of mental health outcomes.

From a methodological perspective, the emphasis on reproducibility strengthens confidence in the reliability of the findings and addresses growing concerns about transparency in machine learning research. Evaluating stability across multiple random seeds and examining variance in SHAP values helped ensure that feature importance estimates were not dependent on a single model initialization. This approach highlights the value of interpretable and reproducible machine learning techniques, particularly in sensitive domains such as mental health.

From a psychological perspective, the findings align with existing literature on sleep hygiene, digital overload, and attentional control. The prominence of sleep duration and focus-related measures supports the importance of restorative behaviors and sustained attention as protective factors, while the influence of notification frequency reflects the potential impact of digital interruptions on well-being. Although the model explained a moderate proportion of variance, this outcome reflects the multifactorial nature of mental health rather than a limitation of the analytical approach.

Several limitations should be considered when interpreting these results. The dataset was modest in size and relied on self-reported measures, which may introduce bias. In addition, the cross-sectional design limits causal inference, and the findings may not generalize to all populations. Future research could extend this work by incorporating larger, more diverse datasets, longitudinal designs, and additional behavioral or clinical variables to further examine the relationship between digital behavior and mental health.

Overall, this research demonstrates that interpretable and reproducible machine learning can provide meaningful insights into how everyday digital behaviors relate to mental health outcomes. By bridging computational modeling and psychological research, the study contributes a transparent analytical framework that can support future research and inform the development of responsible, evidence-based digital wellness interventions.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The dataset used in this study is publicly available and was obtained from Kaggle. All data used in the analysis are anonymized and accessible through the original source.

Code Availability: The code used to preprocess the data, train the XGBoost models, evaluate performance across multiple random seeds, and generate SHAP explanations is publicly available on GitHub at: <https://github.com/tarya145/Final-Thesis>.

References

1. Kaggle. (n.d.). Mental health and digital behavior dataset[2020-2024]. Kaggle.[Mental Health and Digital Behavior \(2020–2024\)](#)
2. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2623–2631. <https://doi.org/10.1145/3292500.3330701>
3. Bouthillier, X., Laurent, C., & Vincent, P. (2019). Unreproducible research is reproducible. Proceedings of the 36th International Conference on Machine Learning, 624–633. <http://proceedings.mlr.press/v97/bouthillier19a.html>
4. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint. <https://arxiv.org/abs/1702.08608>
5. Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, 9(4), 637–671. <https://uncg.edu/~mjkane/pubs/Kane%20&%20Engle%202002,%20WM%20&%20PFC,%20PBR.pdf>

6. Killgore, W. D. S. (2010). Effects of sleep deprivation on cognition. *Progress in Brain Research*, 185, 105–129. <https://doi.org/10.1016/B978-0-444-53702-7.00007-5>
7. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4768–4777. [1705.07874](https://arxiv.org/abs/1705.07874)
8. Mark, G., Gudith, D., & Klocke, U. (2008). The cost of interrupted work: More speed and stress. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 107–110. <https://ics.uci.edu/~gmark/chi08-mark.pdf>
9. Ophir, E., Nass, C., & Wagner, A. D. (2009). Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences*, 106(37), 15583–15587. <https://www.pnas.org/doi/10.1073/pnas.0903620106>
10. Pilcher, J. J., & Huffcutt, A. I. (1996). Effects of sleep deprivation on performance: A meta-analysis. *Sleep*, 19(4), 318–326. <https://doi.org/10.1093/sleep/19.4.318>
11. Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research (A NeurIPS 2019 perspective). *Journal of Machine Learning Research*, 22(164), 1–20. [2003.12206](https://arxiv.org/abs/2003.12206)
12. Posner, M. I., & Rothbart, M. K. (2007). Research on attention networks as a model for the integration of psychological science. *Annual Review of Psychology*, 58, 1–23. https://users.php.ufl.edu/rbauer/cognitive/Articles/posner_rothbart_integration_07.pdf
13. Przybylski, A. K., & Weinstein, N. (2017). Digital screen time limits and young people's psychological well-being: Evidence from population data. *Child Development*, 88(6), 1737–1751. <https://doi.org/10.1111/cdev.12819>
14. Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). "Oops!": Performance correlates of everyday attentional failures in traumatic brain injury and normal subjects. *Neuropsychologia*, 35(6), 747–758. [https://doi.org/10.1016/S0028-3932\(97\)00015-8](https://doi.org/10.1016/S0028-3932(97)00015-8)
15. Van Dongen, H. P. A., Maislin, G., Mullington, J. M., & Dinges, D. F. (2003). The cumulative cost of additional wakefulness: Dose-response effects on neurobehavioral functions and sleep physiology. *Sleep*, 26(2), 117–126. <https://doi.org/10.1093/sleep/26.2.117>
16. Walker, M. (2017). *Why we sleep: Unlocking the power of sleep and dreams*. Scribner. [Why We Sleep: Unlocking the Power of Sleep and Dreams - PDFDrive.com](https://www.pdfdrive.com/why-we-sleep-ebook.html)

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.