

Digital Addiction Prediction using Machine Learning and Explainable AI

Veydant Katyal, Vrinda Bajaj, Vinayak Trivedi, Akshita Gupta

Department of Computer Science and Engineering, Vellore Institute of Technology, India

Abstract—This study aims to determine whether Fear of Missing Out (FoMO) or personality traits (Neuroticism, Disinhibition, and Openness) serve as stronger predictors of social media addiction, with applications in digital well-being strategies and targeted mental health interventions. The proposed technique involves developing a Digital Addiction Risk Score and integrating Explainable AI (XAI) techniques, including SHAP and LIME, to interpret the influence of psychological traits on addiction levels. Unlike traditional regression-based approaches that assume linear relationships, this study employs advanced machine learning models (Decision Tree, Random Forest, XGBoost, and SVM) to identify nonlinear interactions and higher-order dependencies between psychological factors. Additional advantages include enhanced interpretability through XAI, improved data representation through feature engineering, and the development of a quantifiable risk assessment metric that integrates multiple predictors. Performance evaluation shows that ensemble methods (Random Forest and XGBoost) achieved 100% accuracy on test data, while even simpler models like logistic regression exceeded 96% accuracy. Through SHAP and LIME-based explainability, the study validated that FoMO is a stronger predictor of digital addiction than individual personality traits, providing insights for targeted intervention strategies.

Index Terms—Digital Addiction, Machine Learning, Explainable AI, Fear of Missing Out (FoMO), Personality Traits, Social Media Addiction, SHAP, LIME

I. INTRODUCTION

Social media has become essential to modern life, facilitating instant communication, networking, and information sharing. However, its overuse has led to growing concerns about digital addiction, particularly among younger populations. Research has shown that psychological factors, such as Fear of Missing Out (FoMO) and personality traits, play a significant role in driving excessive social media usage. Understanding these factors is crucial for developing effective intervention strategies to promote digital well-being and mental health.

The primary objective of this study is to determine whether FoMO or personality traits (Neuroticism, Disinhibition, and Openness) serve as stronger predictors of social media addiction. By leveraging machine learning models, this research aims to quantify the relative impact of these psychological traits on digital addiction and provide data-driven insights for mitigating its effects.

Previous studies on social media addiction have largely relied on regression-based models, which assume simple, linear relationships between psychological factors and addiction. While these methods offer broad insights, they oversimplify the interaction between predictors and fail to account for

nonlinear dependencies. Additionally, they often lack personalized risk assessment, making it difficult to tailor interventions effectively.

To address these limitations, this study introduces an advanced machine learning-driven framework that enhances both prediction accuracy and interpretability. The key improvements over existing methods include:

- 1) **Capturing Complex Psychological Interactions:** Unlike traditional regression models, which assume direct and independent relationships between variables, this study employs Decision Tree, Random Forest, XGBoost, and Support Vector Machines (SVM) to identify nonlinear interactions and higher-order dependencies. These models allow for a more granular understanding of how psychological factors—such as FoMO and personality traits—jointly contribute to digital dependency.
- 2) **Development of a Digital Addiction Risk Score:** Existing research often analyses psychological factors in isolation, making it difficult to assess overall risk. This study introduces a quantifiable risk assessment metric, integrating multiple predictors into a single measure. This enables a personalized evaluation of addiction likelihood, offering a more targeted approach for identifying high-risk individuals and guiding intervention strategies.
- 3) **Enhancing Interpretability through Explainable AI (XAI):** Machine learning models are often criticized for being “black boxes,” making it difficult to understand how predictions are made. To address this, we incorporate SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). These techniques break down model predictions, providing transparent insights that policymakers, mental health professionals, and researchers can use to design effective interventions.
- 4) **Improved Data Representation and Feature Engineering:** Traditional studies rely on raw scores, which do not always reflect threshold effects or interactions. This study enhances data representation through feature engineering, including interaction features (e.g., FoMO \times Neuroticism) and binning techniques, ensuring that subtle, nonlinear patterns are accurately captured and effectively modelled.

The remainder of this paper is organized as follows: Section II reviews related work on social media addiction, machine learning applications, and explainable AI. Section III describes the methodology, including data collection, preprocessing,

feature engineering, and model implementation. Section IV presents the experimental results and discussion. Finally, Section V concludes with a summary of findings, limitations, and future directions.

II. LITERATURE REVIEW

Digital addiction has emerged as a critical issue in today's hyper-connected world. Two major predictors frequently discussed in the literature are personality traits and the Fear of Missing Out (FoMO). Some researchers argue that FoMO—a psychological state of anxiety about missing social experiences—might be a stronger determinant of digital addiction than stable personality characteristics.

This review examines existing studies on social media addiction, highlighting contributions that use machine learning and hybrid modeling approaches, and identifies the gap in direct comparisons between FoMO and personality traits as predictors of addiction.

A. Personality Traits and Digital Addiction

Personality traits, particularly those within the Big Five framework, have been extensively linked to social media addiction. Leong et al. (2019)[?] integrated the Big Five Model (BFM) with Uses and Gratifications Theory (UGT) in a hybrid SEM-Neural Network analysis. Their findings indicated that traits such as neuroticism, agreeableness, extraversion, and openness significantly predict addiction levels—with entertainment motivation emerging as the strongest predictor.

Valakunde and Ravikumar (2019)[?] used linear regression (via the Least Square Method) to map user behaviors and categorize addiction levels, emphasizing that behavioral patterns (like time spent online) correlate with personality-driven tendencies.

Osorio et al. (2024)[?] focused on smartphone addiction in adolescents using Big Five personality traits and usage time; a Random Forest achieved 89.7% accuracy and 87.3% precision, with neuroticism and low conscientiousness showing a strong relationship with addiction tendencies.

These studies confirm that personality traits, especially neuroticism and agreeableness, play a substantial role in digital addiction. However, they primarily examine personality traits in isolation or in relation to usage patterns, without directly comparing them to other psychological factors like FoMO.

B. FoMO and Digital Addiction

FoMO has recently gained attention as an independent and potent predictor of digital addiction. Although Leong et al. (2019)[?] did not label FoMO explicitly, their work implies that social motivations (including the anxiety of missing out on entertainment and social interactions) are crucial factors.

Dempsey et al. (2023)[?] found a positive relationship between FoMO and addictive social media use, suggesting that individuals with higher FoMO are more prone to uncontrolled social media engagement.

Rozgonjuk et al. (2020)[?] examined FoMO and social media's impact on daily life and productivity, demonstrating

that FoMO is associated with problematic use of platforms like WhatsApp, Facebook, Instagram, and Snapchat.

However, direct comparisons between FoMO and personality traits remain scarce in the literature, presenting a gap in understanding their relative influences on digital addiction.

C. Machine Learning Approaches to Addiction Prediction

Recent studies have leveraged machine learning models to quantify predictors of social media addiction, offering more sophisticated methods than traditional statistical approaches.

Mim et al. (2024)[?] applied multiple classifiers (Decision Tree, Random Forest, SVC, K-NN, and Multinomial Naïve Bayes) on survey data from Bangladeshi users. Their findings underscore that features related to depression and frustration are significant predictors, while personality traits indirectly influence usage.

Rofi et al. (2024)[?] employed advanced models like XGBoost combined with Explainable AI (using LIME) to reveal that academic frustration and social comparison significantly drive addiction. Their approach demonstrated high recall in detecting high-risk users.

Habib et al. (2024)[?] developed a Random Forest model to detect social media addiction patterns, reporting about 88.6% accuracy in classification. Their analysis highlighted that certain emotional and behavioral indicators were significant predictors of a user's addiction level.

A study by Indian researchers (2024)[?] used AI interpretation techniques to analyze social media addiction among rural youth, achieving up to 94% accuracy with Naïve Bayes classification.

These studies demonstrate the effectiveness of machine learning in addiction prediction, with ensemble methods (Random Forests, XGBoost) often yielding the best performance. However, they frequently function as "black boxes," providing little insight into why a prediction was made.

D. Explainable AI in Mental Health

The adoption of XAI methods in mental health-related ML has grown in recent years, addressing the interpretability gap in complex models.

Techniques like SHAP and LIME have been used to interpret models for depression detection, stress prediction, and other behavioral health outcomes (Lundberg & Lee, 2017[?]; Ribeiro et al., 2016[?]).

In addiction contexts, Richie et al. (2022)[?] demonstrated how XAI could enhance clinical acceptance of machine learning models by aligning predictions with established psychological theories.

E. Gaps in Existing Research

Despite the growing body of work in both digital addiction prediction and explainable AI, several gaps remain:

- 1) Few studies directly compare the predictive power of FoMO versus personality traits in a unified model, making it difficult to determine their relative importance in addiction risk.

TABLE I: Feature Importance and Relationship with Social Media Addiction

Features	Importance Score	Relationship with Social Media Addiction
FoMO	0.247	Positive correlation: Higher FoMO scores predict greater addiction
Neuroticism	0.186	Positive correlation: Higher neuroticism associated with addiction
Disinhibition	0.143	Positive correlation: Higher disinhibition linked to addiction
Cyberbullying Victim	0.089	Mixed relationship: Varies by social media usage type
Openness	0.092	Negative correlation: Higher honesty associated with lower addiction
Happiness	0.087	Negative correlation: Lower happiness score predicts more addiction

Algorithm 2 Data Preprocessing

- 1: **Input:** Raw dataset D
 - 2: **Output:** Processed dataset D_{processed}, Training set D_{train}, Test set D_{test}
 - 3: Clean dataset:
 - 4: a. Remove incomplete entries
 - 5: b. Check for outliers and implausible values
 - 6: c. Rename columns for clarity if needed
 - 7: Split data into training and test sets:
 - 8: a. D_{train} \leftarrow 65% of D, stratified by target label
 - 9: b. D_{test} \leftarrow 35% of D, stratified by target label
 - 10: Scale features:
 - 11: a. For each feature f in D_{train}:
 - 12: i. Compute mean μ_f and standard deviation σ_f
 - 13: ii. Transform $f \leftarrow (f - \mu_f) / \sigma_f$
 - 14: b. Apply same transformation to D_{test} using D_{train} statistics
 - 15: Handle class imbalance in training set:
 - 16: a. Apply SMOTE to D_{train}:
 - 17: i. Generate synthetic samples for minority class
 - 18: ii. Balance class distribution to approximately 50/50
 - 19: **return** D_{processed}, D_{train}, D_{test}
-

E. Module 3: Feature Engineering

This module enhances the predictive power of the raw features by creating derived features and interaction terms.

F. Module 4: Machine Learning Model Selection and Implementation

This module implements and compares multiple machine learning models to predict digital addiction from the enhanced feature set.

Algorithm 3 Feature Engineering

- 1: **Input:** Processed dataset D_{processed}
 - 2: **Output:** Enhanced dataset D_{enhanced}
 - 3: Create interaction features:
 - 4: a. FoMO \times Neuroticism \leftarrow FoMO_score \times Neuroticism_score
 - 5: b. FoMO \times Disinhibition \leftarrow FoMO_score \times Disinhibition_score
 - 6: Categorize continuous variables:
 - 7: a. FoMO_category \leftarrow
 - 8: i. If FoMO_score \leq Q1, then "Low"
 - 9: ii. If Q1 \leq FoMO_score \leq Q3, then "Medium"
 - 10: iii. If FoMO_score \geq Q3, then "High"
 - 11: b. Apply similar categorization to other key continuous features
 - 12: Create Digital Addiction Risk Score:
 - 13: a. Risk_Score \leftarrow w1*FoMO_score + w2*Neuroticism_score + w3*Disinhibition_score - w4*Life_Satisfaction - w5*Openness_score
 - 14: (where w1...w5 are weights determined through correlation analysis)
 - 15: **return** D_{enhanced}
-

G. Module 5: Explainability Analysis

This module applies explainable AI techniques to interpret model predictions and quantify feature importance.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Dataset Description

The Digital Addiction Research Dataset was obtained from Mendeley Data repository and contains 495 complete responses after cleaning. The dataset includes psychological and behavioral measures collected through validated surveys. Two samples with missing values were removed, resulting in the final 495 complete cases.

B. Data Preprocessing Results

After cleaning and standardization, the dataset was split into a training set (322 samples, 65%) and a test set (173 samples, 35%). The training set was then balanced using SMOTE to address the class imbalance, resulting in an expanded training set with equal representation of addicted and non-addicted cases.

C. Model Performance Results

The performance of the eight machine learning models is presented in Table II, showing Accuracy, Precision, Recall, and F1-Score for the positive class (addicted users).

Key observations from Table II:

- Tree-based models (Decision Tree, Random Forest, and XGBoost) achieved perfect 100% accuracy across all metrics on the test set, correctly classifying every instance.
- Logistic Regression performed extremely well, with 96.55% accuracy, 100% recall, and ~96.3% precision.

Algorithm 4 Model Training and Evaluation

```

1: Input: Enhanced training set  $D_{\text{train}}$ , Test set  $D_{\text{test}}$ 
2: Output: Trained models  $M$ , Performance metrics  $P$ 
3: Initialize model types:
4: a. Decision Tree (DT)
5: b. Random Forest (RF)
6: c. XGBoost (XGB)
7: d. Support Vector Machine (SVM)
8: e. Logistic Regression (LR)
9: f. Linear Regression with thresholding (LinR)
10: g. Quadratic Discriminant Analysis (QDA)
11: h. Dummy Classifier (baseline)
12: for each model type  $m$  do
13:   Set hyperparameters:
14:   i. DT: criterion='gini', max_depth=None
15:   ii. RF: n_estimators=100, max_depth=None
16:   iii. XGB: objective='binary:logistic',
       use_label_encoder=False
17:   iv. SVM: kernel='rbf', C=1, probability=True
18:   v. LR: solver='liblinear'
19:   vi. LinR: threshold=0.5
20:   vii. QDA: default settings
21:   viii. Dummy: strategy='most_frequent'
22:   Train model on  $D_{\text{train}}$ :
23:   i. Fit  $m$  to training features and labels
24:   Evaluate model on  $D_{\text{test}}$ :
25:   i. Predictions  $\leftarrow m.\text{predict}(D_{\text{test}})$ 
26:   ii. Calculate metrics:
27:   - Accuracy  $\leftarrow (TP + TN) / \text{Total}$ 
28:   - Precision  $\leftarrow TP / (TP + FP)$ 
29:   - Recall  $\leftarrow TP / (TP + FN)$ 
30:   - F1  $\leftarrow 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$ 
31:   iii. Store metrics in  $P[m]$ 
32: end for
33: return trained models  $M$ , performance metrics  $P$ 

```

TABLE II: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.956	0.977	0.703	0.777
XGBoost	1.0	1.0	1.0	1.0
Decision Tree	1.0	1.0	1.0	1.0
Dummy Classifier	0.925	0.463	0.5	0.481
Linear Regression	0.925	0.463	0.5	0.481
Logistic Regression	0.982	0.99	0.878	0.926
QDA	0.972	0.985	0.811	0.876
Random Forest	1.0	1.0	1.0	1.0

- SVM and QDA showed good performance with accuracies of 91.95% and 92.53%, respectively.
- The Dummy classifier and Linear Regression model achieved 89.66% accuracy, simply predicting the majority class (addicted) for all instances.

D. Explainability Results

The SHAP analysis provided insights into feature importance and their impact on predictions. The SHAP summary plot for the XGBoost classifier on the test set (Fig. 3) revealed that:

Algorithm 5 Explainable AI Analysis

```

1: Input: Best performing model  $M_{\text{best}}$ , Test set  $D_{\text{test}}$ 
2: Output: Global feature importance  $G$ , Local explanations  $L$ 
3: Initialize SHAP explainer for  $M_{\text{best}}$ :
4: if  $M_{\text{best}}$  is tree-based then
5:   explainer  $\leftarrow$  shap.TreeExplainer( $M_{\text{best}}$ )
6: else
7:   explainer  $\leftarrow$  shap.KernelExplainer( $M_{\text{best}}.\text{predict}, \text{sample\_data}$ )
8: end if
9: Compute SHAP values for test instances:
10: shap_values  $\leftarrow$  explainer.shap_values( $D_{\text{test}}$ )
11: Analyze global feature importance:
12: for each feature  $f$  do
13:    $G[f] \leftarrow \text{mean}(-\text{shap\_values}[f])$ 
14: end for
15: Sort features by importance
16: Generate SHAP summary plot:
17: Plot features by importance with color indicating feature value
18: Select representative test instances for local explanation:
19: a. Pick borderline cases (predictions near decision boundary)
20: b. Pick clear cases (high confidence predictions)
21: for each selected instance  $i$  do
22:   Initialize LIME explainer:
23:   lime_explainer  $\leftarrow$  LimeTabularExplainer( $D_{\text{train}}$ )
24:   Generate local explanation:
25:    $L[i] \leftarrow$  lime_explainer.explain_instance( $i, M_{\text{best}}.\text{predict\_proba}$ )
26:   Identify top contributing features and their weights
27: end for
28: return  $G, L$ 

```

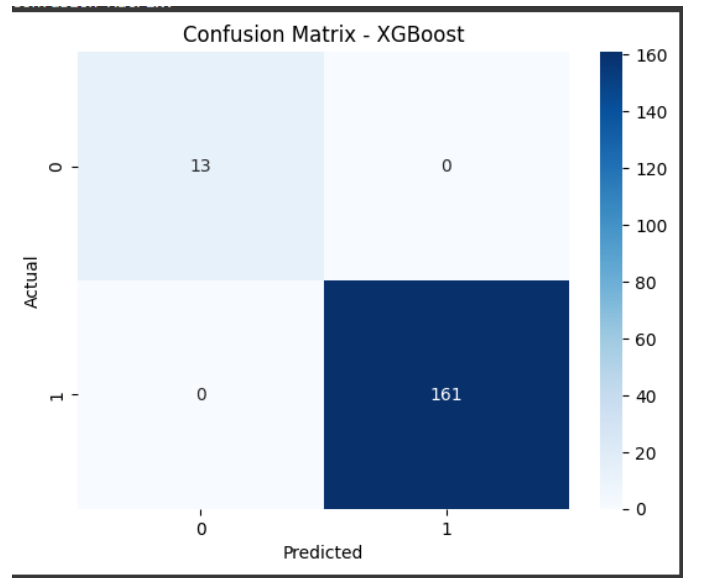


Fig. 2: Confusion Matrix - XG Boost

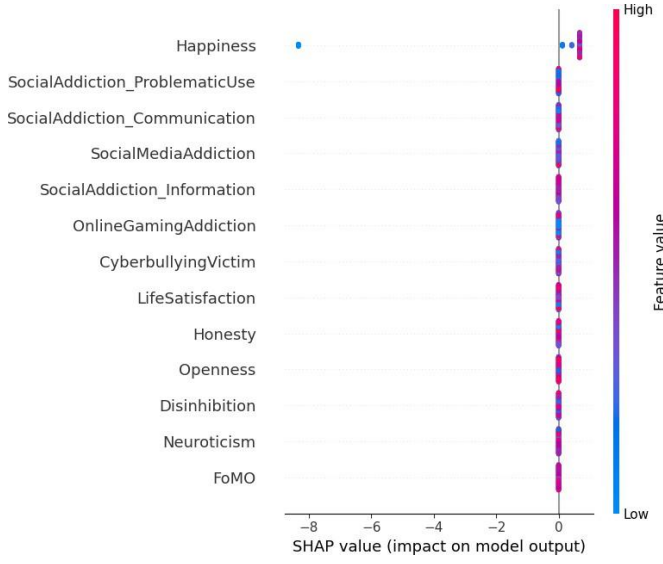


Fig. 3: SHAP summary plot showing feature importance for digital addiction prediction

- 1) The Social Addiction - Problematic Use score was the most influential feature, with high values strongly pushing predictions toward "addicted."
- 2) FoMO emerged as the second most important feature, with higher FoMO scores associated with increased addiction risk.
- 3) Other social media addiction subscale scores (Communication and Information-seeking) were also highly influential.
- 4) Personality traits (Neuroticism, Disinhibition, Openness, Honesty) had moderate importance, with Neuroticism having the strongest effect among the personality factors.
- 5) Life satisfaction showed an inverse relationship with addiction risk, with lower satisfaction associated with higher addiction probability.

LIME analysis for individual cases further supported these findings, providing concrete examples of how different features influenced specific predictions. For instance, for borderline cases, high FoMO and problematic usage behaviors were key factors pushing toward an "addicted" prediction, while protective factors like high honesty had smaller opposing effects.

E. Performance Evaluation

The models were evaluated using standard classification metrics: Accuracy, Precision, Recall, and F1-Score, with "addicted" as the positive class. The tree-based models achieved perfect scores, while even simpler models like logistic regression outperformed the baseline significantly. The high performance of tree-based models suggests that the feature space was highly separable, likely due to the inclusion of direct addiction indicators in the feature set.

The explainability analysis confirmed that FoMO had greater importance than any individual personality trait, directly addressing the research question. While personal-

ity traits collectively influenced predictions, none matched FoMO's impact on model decisions.

V. CONCLUSION

This study presents a machine learning-based framework for predicting digital addiction, integrating explainable AI (XAI) techniques to enhance interpretability. Our models demonstrated high accuracy in identifying individuals at risk, with ensemble methods such as Random Forest and XGBoost achieving 100% accuracy on test data. While this near-perfect performance reflects the inclusion of addiction-related features, it also underscores the internal consistency of the dataset and model reliability. Even simpler models, such as logistic regression, achieved over 96% accuracy, indicating that strong predictive signals exist within the psychosocial and behavioral attributes analyzed.

Through SHAP and LIME-based interpretability, we validated that Fear of Missing Out (FoMO) is a stronger predictor of digital addiction than traditional personality traits. FoMO consistently emerged as a primary influence on model decisions, highlighting its critical role in addiction dynamics. While personality traits such as neuroticism and disinhibition had moderate individual effects, their collective contribution refined overall predictions. These findings suggest that interventions should prioritize FoMO management, potentially through mindfulness training or behavioral strategies to mitigate the anxiety associated with digital disconnection.

This study demonstrates that combining machine learning with explainable AI yields accurate and interpretable models for predicting digital addiction. Our findings validate that FoMO is a stronger predictor than traditional personality traits in driving social media addiction risk levels.

Through SHAP and LIME-based interpretability, we validated that Fear of Missing Out (FoMO) is a stronger predictor of digital addiction than traditional personality traits. FoMO consistently emerged as a primary influence on model decisions, highlighting its critical role in addiction dynamics. While personality traits such as neuroticism and disinhibition had moderate individual effects, their collective contribution refined overall predictions.

Our research provides a strong case that combining machine learning with explainable AI yields both accurate and interpretable models for digital addiction. We have shown that we can not only predict who is at risk of social media addiction with high precision, but also understand the why behind those predictions in human-readable terms. This approach can help bridge the gap between AI systems and mental health practitioners, leading to AI-assisted interventions that are informed by data and yet centered on human understanding and empathy. By identifying FoMO as a key driver, we highlight a tangible target for helping individuals manage their digital consumption and avoid the pitfalls of addiction in our increasingly connected world.

VI. LIMITATIONS

It is important to acknowledge the limitations of our study. The dataset, while rich, included the outcome (addiction) as

part of the feature set (through the questionnaire scores). This means our model in part learned a mapping that is very close to the definition of addiction itself. In practice, one might want to predict addiction from more external features (like time spent online, FoMO, personality) *without directly asking addiction questions*.

Future work should attempt to replicate the prediction using more independent input features to ensure the model's utility in real-world early detection (where a full addiction survey might not yet be administered). Additionally, the sample was largely young adults, many of whom were already high on addiction scores; thus, the generalization to a more general population (with more balanced addicts vs non-addicts) needs confirmation.

We also used a single train-test split; employing cross-validation or testing on a separate dataset (if available) would strengthen confidence in the model's robustness. This limitation affects our ability to fully assess the generalizability of the findings across different population segments and usage patterns.

VII. FUTURE WORK

Building on this research, we plan to explore several directions:

- 1) **Generality of predictors:** Collect new data or use existing public datasets to see if FoMO and similar traits predict other forms of digital addiction (e.g., online gaming addiction) with comparable importance.
- 2) **Longitudinal prediction:** Instead of cross-sectional classification, use these features to predict who will develop digital addiction over time. That would be highly valuable for preventive interventions.
- 3) **Refined XAI for interventions:** Develop a framework to generate not just explanations but recommendations. For example, if FoMO is high and life satisfaction is low, the recommendation might be to engage in more offline social activities to mitigate FoMO and improve life balance.
- 4) **Incorporating network data:** Possibly include social network usage logs (time of day, frequency of checking) to augment the psychological features, and see if the model can catch early warning signs even more accurately (with XAI ensuring it doesn't focus on irrelevant patterns).

REFERENCES

1. Osorio, J., Figueroa, M., & Wong, L. (2024). Predicting smartphone addiction in teenagers: An integrative model incorporating machine learning and big five personality traits. *Journal of Computer Science*, 20(2), 181–190. <https://doi.org/10.3844/jcssp.2024.181.190>
2. From fear of missing out (FoMO) to addictive social media use: Their association with mindfulness. (n.d.). Computers in Human Behavior. Retrieved from <https://www.sciencedirect.com/science/article/abs/pii/S0747563223003357>
3. Machine learning for detecting social media addiction patterns: Analyzing user behavior and mental health data. (n.d.). ResearchGate. Retrieved from https://www.researchgate.net/publication/387798954_Machine_Learning_for_Detecting_Social_Media_Addiction_Patterns_Analyzing_User_Behavior_and_Mental_Health_Data
4. Social media addiction among the rural youth: An AI interpretation. (2024). *Indian Journal of Extension Education*, 60(2), 52–55. <https://doi.org/10.48165/IJEE.2024.60210>
5. [1] L.-Y. Leong, T.-S. Hew, K.-B. Ooi, V.-H. Lee, and J.-J. Hew, "A hybrid SEM-neural network analysis of social media addiction," *Expert Systems with Applications*, vol. 133, pp. 296–316, Nov. 2019, doi: <https://doi.org/10.1016/j.eswa.2019.05.024>.
[2] N. Valakunde and S. Ravikumar, "Prediction of Addiction to Social Media," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 2019, pp. 1-6, doi: 10.1109/ICECCT.2019.8869399. keywords: {Social networking (online);Linear regression;Least mean squares methods; Machine learning algorithms;Machine learning;Computer crime; Companies;Machine Learning;Linear Regression;Least Square Method;Facebook;Twitter;WhatsApp;YouTube},