Brain Rot Detection in Users: Analyzing Behavioral Insights from Large-Scale User Interactions

Kein Yamada
Graduate School of Fundamental Science and Engineering, Waseda University
Tokyo, Japan
stslm738.ymd@toki.waseda.jp

Hiroshi Watanabe
Graduate School of Fundamental Science and Engineering, Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—Social media are extensively utilized by individuals across all age groups, greatly influencing modern society. However, certain types of content can lead to severe user addiction, often colloquially referred to as "brain rot." We propose an approach for detecting users with strong social media addiction by analyzing user interaction data and the categorical characteristics of consumed media content. Our method investigates behavioral patterns associated with high levels of addiction using a dataset that captures detailed user engagement metrics. Utilizing SHapley Additive exPlanations for feature analysis, we apply decision tree classifiers and random forest to identify users exhibiting addictive behavior. Our research aims to clarify the factors contributing to social media addiction, providing actionable insights for the social media management of consumers.**

Keywords*—internet addiction, social media, metadata, machine learning, random forest, social media.

I. Introduction

Social media platforms are deeply embedded in daily life, offering connectivity and entertainment to users of all ages. However, concerns have grown over excessive and compulsive usage, often labeled "brain rot." Remarkably, brain rot was selected as Oxford Word of the Year 2024 [1], underscoring rising awareness of social media overexposure.

Detecting such addiction-like behavior is challenging due to the complex nature of user interactions and content diversity. There has been medical and psychological methods [2]–[4] for confronting social media addiction but these are not easily accessible for social media operators. Furthermore, it is not practical for every social media users to wear devices that checks their brain activity. Similarly, traditional methods like surveys and interviews [5]–[7] are also limited in scalability and objectivity. In contrast, the increasing availability of user interaction logs and behavioral metadata presents an opportunity for identifying potential cases of addiction in a data-driven way.

We propose a machine learning-based approach for perceiving users that show strong signs of social media addiction. By analyzing user interaction patterns and media content types from a comprehensive recommendation dataset, we extract behavioral profiles and utilize interpretable models to classify addiction risk. Furthermore, we will identify what factors of personal features lead or correlate with brain rot symptoms.

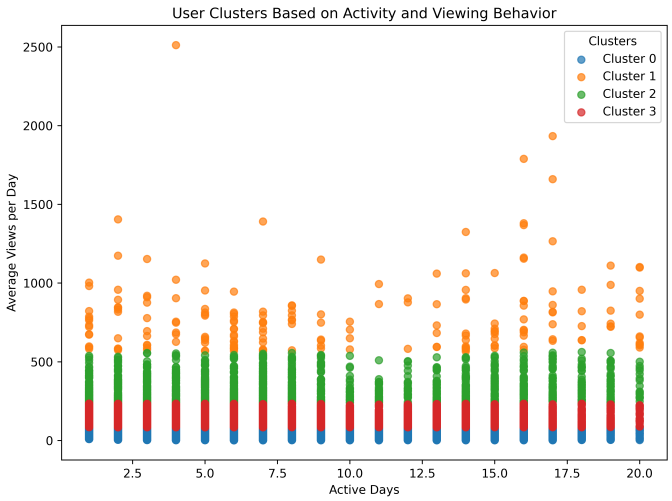

Fig. 1. Social media user distributed into 4 clusters mainly based on average views per day. Cluster 1 in this graph will be labeled as "brain rot" while the rest are labeled as "non-brain rot."

II. Related Work

Several studies have explored social media addiction using machine learning techniques; however, most rely on limited or self-reported datasets. For instance, Akter *et al.* [6] employed a dataset collected through questionnaires and direct interviews involving 504 users, which, while insightful, lacked behavioral granularity and omitted features readily obtainable from social media platforms. Similarly, Mardiah *et al.* [7] utilized a publicly available dataset from Kaggle, comprising 481 entries, also based on questionnaire responses. These approaches, though valuable, are constrained by sample size and subjective bias.

III. Proposed Method

A. Brain Rot Labeling

While prior studies have primarily relied on self-reported or limited-scale datasets, few have relied on large-scale social media interaction logs, which can provide more accurate, scalable, and objective insights into user addiction. Hence, we utilized the KuaiSAR dataset [8] for brain rot analysis. This dataset commonly serves as a benchmark for evaluating

979-8-3315-2415-9/25/$31.00 ©2025 IEEE

recommendation and search functionalities on social media platforms. This dataset is well-suited for our purpose, as it offers numerous features directly tied to user behavior and platform services. However, this dataset has one major concern, which is the lack of ground truth labels for users with addiction.

Since we lack the label important for classifying and analyzing brain rot users, we set a hypothetical ground truth through two approaches. For the first approach, we consider the use of the pre-labeled "recommend active level," provided by the authors of the KuaiSAR dataset. This indicator has a range of 0 to 3, where the values are based on the active days of the users. In other words, this label for brain rot will heavily rely on active days. However, we assume that this approach may face an issue of not grasping the correct users. For instance, there could be a user that possibly stops using the social media after watching 100 short videos, averaging 100 views per day. In another case, there could be a different user that does the same but watches 1000 short videos. If these two users have the same active days, it may possibly present conflicts with analyzing results.

Seeing this as the potential issue, we present the second approach of utilizing both active days and average views per day for labeling brain rot users. In this approach, we place the emphasis more towards the average views per day. Hence, we achieve the distribution shown in Fig. 1 through the k-mean clustering. As it can be seen in Fig. 1, there are indeed several viewers that have been actively using social media but are low in average view counts per day (cluster 0). We also see that there are several groups of users that have a relatively high average view counts per day, despite the low active days. By taking these two approaches, we expect to gain a better understanding and a clearer insight on the data-driven clues for brain rot.

B. Exploring User Interaction Features

KuaiSAR contains extensive records of user interactions, including actions such as likes, follows, searches, and clicks, alongside metadata such as item category, item duration, and user playing time. These features enable a detailed analysis of user engagement patterns, supporting the detection of addiction-related behavior.

To further improve the model's interpretability and performance, we applied SHapley Additive exPlanations (SHAP) [9] to evaluate the contribution and impact of each feature on the prediction of social media addiction. Additionally, we engineered new features from the original dataset such as categorical ratio (cat_ratio), watch efficiency, session count, engagement ratio, favorite category per user, night owl ratio, average actions per day, and other features to better capture behavioral patterns associated with high addiction levels. We expect these modifications to improve classification accuracy and provide a more thorough understanding of excessive social media behavior. The descriptions and equations about the new features are listed below:

Interaction Rates: These include like_rate, follow_rate, and search_rate, each defined as the proportion of a specific action to total click count. For example, like_rate equals to (like count) divided by (click count). These rates provide normalized indicators of how often specific behaviors occur relative to general activity, helping control for individual differences in overall usage volume.

$$IR_u^{(a)} = \frac{\sum_{i=1}^{N_u} a_{u,i}}{\sum_{i=1}^{N_u} \text{click}_{u,i}}, \quad a \in \{\text{like}, \text{follow}, \text{search}\}. \tag{1}$$

Watch Efficiency: Defined as the ratio of playing_time to duration_ms, this feature quantifies how thoroughly users watch content. Users with low watch efficiency might be rapidly skipping through content, while high-efficiency users may be more engrossed.

$$\text{WatchEfficiency}_u = \frac{\sum_{i=1}^{N_u} \text{playing\_time}_{u,i}}{\sum_{i=1}^{N_u} \text{duration\_ms}_{u,i}}. \tag{2}$$

Engagement Ratio: This composite metric sums the like and follow counts and normalizes them by the click count. It reflects how interactive a user is per content viewed, offering a richer signal of affective or social engagement.

$$ER_u = \frac{\sum_{i=1}^{N_u} (\text{like}_{u,i} + \text{follow}_{u,i})}{\sum_{i=1}^{N_u} \text{click}_{u,i}}. \tag{3}$$

Session Count: To model temporal clustering of activity, we define a session as a group of clicks separated by less than 30 minutes. Session count thus reflects the number of distinct viewing bursts and can distinguish between habitual short bursts versus prolonged single sessions.

$$\text{SessionCount}_u = \left| \{ s \in \mathcal{S}_u \mid \text{gap}(s_j, s_{j-1}) \geq 30 \text{ min} \} \right|. \tag{4}$$

Night Owl Ratio: This metric quantifies the proportion of user interactions that occur between midnight and 5:00 AM. Elevated values may suggest irregular usage habits or possible signs of dependency, especially if combined with high frequency.

$$NOR_u = \frac{\sum_{i=1}^{N_u} \mathbb{1}_{\{00:00 \leq t_i < 05:00\}}}{N_u}. \tag{5}$$

Top Category Ratio: By examining the most frequently watched content category and calculating its viewing frequency over all interactions, this metric assesses content diversity. A high top category ratio may suggest obsessive consumption of particular content genres.

$$TCR_u = \frac{\max_{c \in \mathcal{C}} \left(\sum_{i=1}^{N_u} \mathbb{1}_{\{\text{cat}_{u,i} = c\}} \right)}{N_u}. \tag{6}$$

User Category Ratio: This equation defines the ratio $r_{u,c}$ of interactions that user u has with content category c. The nu-

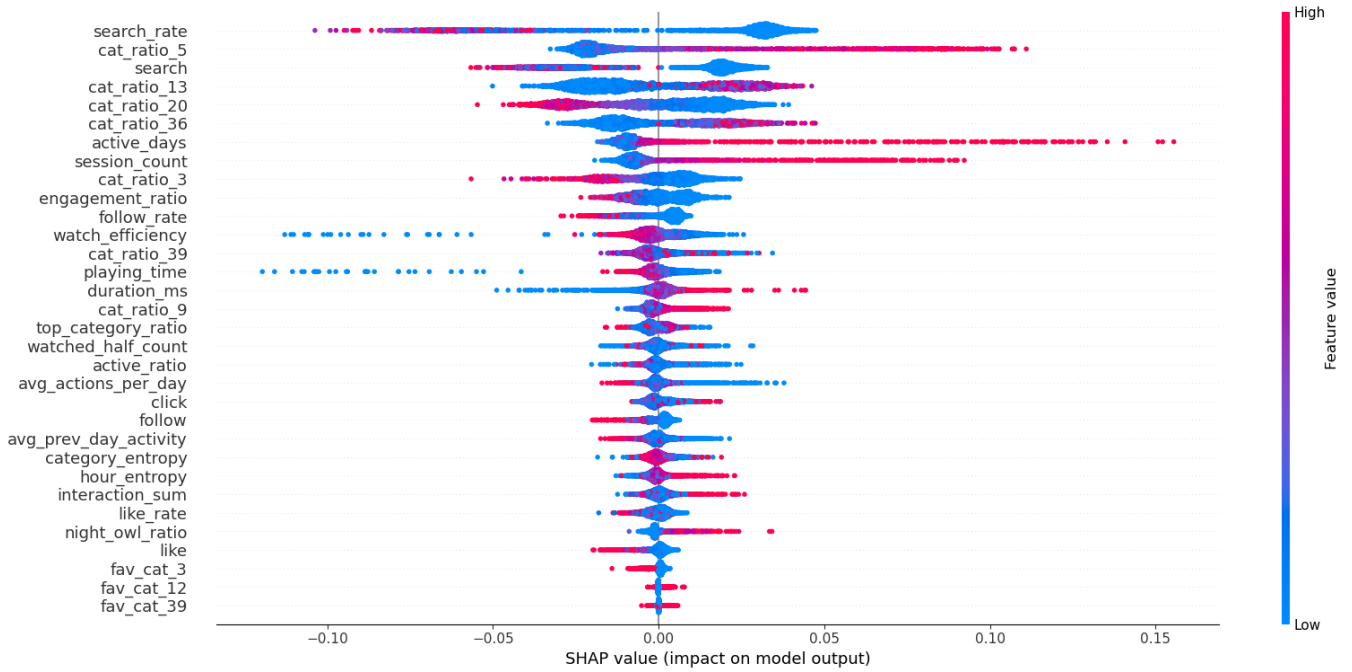

Fig. 2. Visualization of user-centric action and behavioral features extracted from social media interaction data, illustrating their relative contributions to the detection of brain rot levels. Features are displayed in descending order of importance, as determined by their impact on the predicting brain rot levels.

merator $n_{u,c}$ represents the number of interactions in category c, normalized by the total interactions across all categories $\mathcal{C}$.

$$r_{u,c} = \frac{n_{u,c}}{\sum_{c' \in \mathcal{C}} n_{u,c'}}. \tag{7}$$

Category Entropy: This equation defines the category entropy $H_u^{(cat)}$ for a user u, which measures the diversity of content categories the user engages with. The term $p_u(c)$ represents the probability (or proportion) of interactions from user u within category c, and $\mathcal{C}$ denotes the set of all categories. A small constant 10^{-9} is added inside the logarithm to prevent numerical instability when $p_u(c) = 0$. Higher entropy values indicate a more diverse viewing pattern, while lower values suggest a strong preference for specific categories.

$$H_u^{(cat)} = -\sum_{c \in \mathcal{C}} p_u(c) \log_2 \left(p_u(c) + 10^{-9} \right). \tag{8}$$

Hour Entropy: This equation defines the hour entropy $H_u^{(hour)}$ for a user u, which quantifies the diversity of the user's activity across the 24 hours of a day. The term $p_u(h)$ represents the probability of user u being active during hour h. A small constant 10^{-9} is added inside the logarithm to avoid numerical issues when $p_u(h) = 0$. A higher value of $H_u^{(hour)}$ indicates that the user's activity is spread across many hours, while a lower value suggests activity concentrated within specific hours.

$$H_u^{(hour)} = -\sum_{h=0}^{23} p_u(h) \log_2 \left(p_u(h) + 10^{-9} \right). \tag{9}$$

Watched Half Count: This equation defines the *watched_half_count* for a user u, which represents the total number of videos that the user watched at least halfway. For each video i in the set of N_u videos, the indicator function $\mathbb{1}(\cdot)$ counts 1 if the ratio of playing time to video duration is at least 0.5.

$$WHC_u = \sum_{i=1}^{N_u} \mathbb{1}\left(\frac{\text{playing\_time}_{u,i}}{\text{duration\_ms}_{u,i} + 1} \geq \frac{1}{2} \right). \tag{10}$$

Active Ratio: Calculated as the total number of boolean actions divided by total playing_time, this feature captures action density during consumption. A user with a high active ratio is constantly interacting with the platform, which could signal compulsive behavior.

$$AR_u = \frac{\sum_{i=1}^{N_u} (\text{l}_{u,i} + \text{fl}_{u,i} + \text{s}_{u,i} + \text{c}_{u,i} + \text{fw}_{u,i})}{\sum_{i=1}^{N_u} \text{playing\_time}_{u,i}}. \tag{11}$$

IV. Experiment

A. Evaluation method

For our evaluation, we employ a subset of 25,877 users from the KuaiSAR dataset. Each user in this subset is assigned a brain rot level ranging from 0 to 3, determined by their activity intensity metrics, such as the number of active days or average views per day. To simplify the classification task, we treat the highest level (level 3) as "brain rot," while grouping all other levels (0, 1, and 2) under the category of "non-brain rot." This

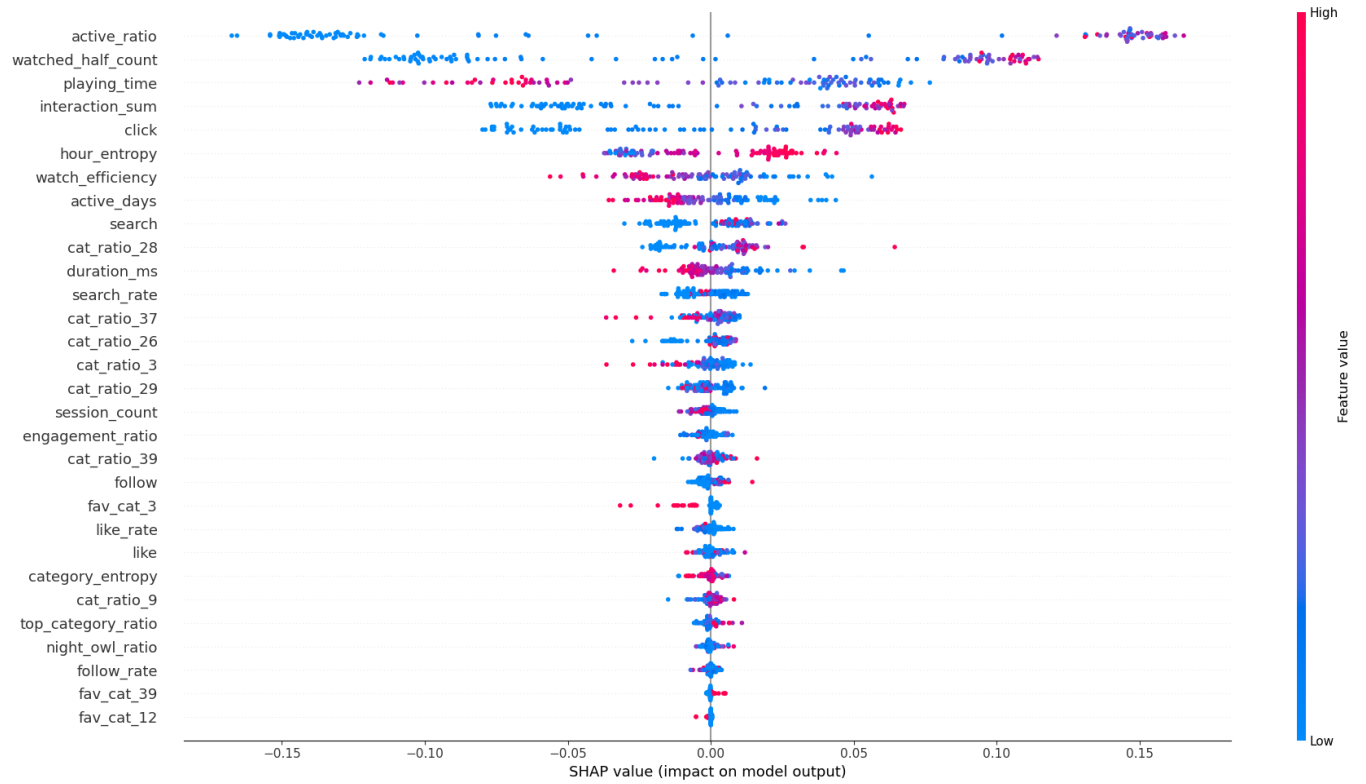

Fig. 3. Visualization of user-centric action and behavioral features extracted from social media interaction data through different labeling approach, illustrating their relative contributions to the detection of brain rot levels. Features are displayed in descending order of importance, as determined by their impact on the predicting brain rot levels.

binary classification approach allows us to focus on identifying distinct behavioral indicators that differentiate highly active users from the rest.

We begin by training a random forest classifier using the default interaction-based features provided within the KuaiSAR recommendation service: playing time, like, follow, search, click, and item duration. These features serve as the baseline indicators of user engagement. To evaluate the added value of more granular behavioral insights, we subsequently retrain the model using an expanded feature set containing more than 20 features in total. This optimized feature set includes a variety of aggregated, temporal, and activity-based metrics designed to capture deeper patterns of user interaction.

The choice of features is carefully controlled based on the hypothetical ground truth labels used in each experimental setup. We explicitly remove features that are directly tied to the definition of the labels to prevent data leakage and ensure a fair evaluation. For example, in the setup illustrated in Fig. 2, where the label is derived from active days, the feature active_days is excluded from the model. Similarly, when the classification is based on average views per day, as shown in Fig. 3, features such as avg_actions_per_day and avg_prev_day_activity are removed to avoid overly optimistic performance results.

While the random forest classifier provides traditional performance metrics such as accuracy, precision, and recall, we do not place significant emphasis on these predictive metrics for this study. The primary reason is that the labels we use are hypothetical and not empirically validated. Consequently, the baseline for accuracy and other metrics is uncertain and lacks definitive validity. Rather than evaluating the model solely based on these conventional measures, we focus on the examination of feature contributions and behavioral patterns revealed by the model. This perspective shifts the purpose of classification from predicting brain rot to using the classifier as a tool for interpreting which features or behaviors are most indicative of high engagement patterns. In this sense, our analysis prioritizes interpretability and feature-level insights over raw classification performance.

B. Results and discussion

1) Analysis based on Active Days: The SHAP analysis in Fig. 2 highlights several strong behavioral indicators. A low search rate and low search count are highly associated with brain rot users, suggesting that these users engage less in active content discovery. Additionally, a high category ratio 5 (style) is prominent among brain rot users, while category ratio 20 (pixiv) is comparatively lower. Non-brain rot users tend to show a low category ratio 13 (san nong), indicating different viewing preferences. Brain rot users also have a higher session count, reflecting more fragmented or frequent viewing patterns.

Moderate indicators show a mixed yet meaningful distinction. Watch efficiency and engagement ratio (likes and follows relative to clicks) are generally lower for brain rot users, indicating more passive consumption. Similarly, follow rate is low, while playing time tends to be higher for non-brain rot users and lower for brain rot users. The interaction sum is noticeably higher for brain rot users, indicating a larger total number of actions even if they are less meaningful.

Weak indicators include watched half count, which displays inconsistent trends, and active ratio (actions divided by playing time), which is generally low across both groups. The night owl ratio leans slightly toward brain rot users but with minimal significance, and category entropy is also highly mixed. Furthermore, many base features such as playing time and clicks appear near the bottom of the SHAP rankings, suggesting that relying solely on these default metrics would not be sufficient for reliable classification.

2) Analysis based on Average Views per Day: When using average views per day as the label (Fig. 3), the most notable feature is watched half count, which strongly differentiates brain rot users who frequently watch videos beyond half their duration. Interestingly, playing time is significantly lower for brain rot users and higher for non-brain rot users, implying that highly engaged users consume more content but spend less time per video. Brain rot users show high interaction sum, click count, and hour entropy, indicating a large volume of activity distributed throughout the day. In contrast, watch efficiency is lower for brain rot users, while non-brain rot users exhibit more effective engagement. Active days are also fewer for brain rot users, suggesting that their activity is concentrated over shorter, intense periods. Search count, however, remains mixed and generally low for both groups.

Moderate indicators include category ratios, which show only minor differences, and session count, which is lower among brain rot users, likely due to their tendency to watch for extended sessions rather than starting multiple new ones. Engagement ratio and follow count are similarly mixed, providing limited discriminative value.

Weak indicators are consistent with those in Fig. 2: category entropy is uninformative, and night owl ratio remains weak, with low values across both groups. Top category ratio and like count also appear low in significance, further suggesting that single-content categories and simple interaction counts offer minimal insight into brain rot behavior.

3) Similarity and Difference in Labeling Methods: The strongest predictors differ significantly depending on how the label is defined. In Fig. 2, search rate, session count, and specific category ratios (e.g., style and pixiv) dominate, whereas in Fig. 3, watched half count, playing time, interaction sum, and hour entropy take precedence. Despite these differences, watch efficiency and engagement ratio are moderately important across both setups, consistently indicating that brain rot users engage less meaningfully despite higher click activity. Notably, playing time shifts from being a weak or mixed indicator in Fig. 2 to a strong signal in Fig. 3, highlighting the influence of the labeling method.

Across both figures, watched half count, interaction sum, playing time, click count, and watch efficiency emerge as the most reliable indicators of brain rot behavior. While content-specific metrics such as category ratios play a role under certain labels (e.g., active days), behavioral intensity and interaction patterns are far stronger predictors overall. This suggests that brain rot is better captured through measures of how users interact—such as session patterns and efficiency—rather than what content they consume.

V. Conclusion

We present a method for analyzing and detecting symptoms of social media addiction, commonly referred to as brain rot, using large-scale user logs and session data. Our approach identifies several key behavioral features that contribute significantly to the classification of highly addicted users, offering potential applicability to similar datasets. For future work, we aim to establish more rigorous criteria for categorizing users with high levels of brain rot. Additionally, seeking for the assistance of psychological experts will be strongly recommended for validating the proper ground truth labels for users with social media addiction.

References

[1] Oxford University Press, "Oxford Word of the Year 2024: Brain rot," *Oxford Languages*, Dec. 2024. [Online]. Available: https://corp.oup.com/word-of-the-year/ [Accessed: Jul. 17, 2025]

[2] Y. Yin, X. Cai, M. Ouyang, S. Li, X. Li, and P. Wang, "FoMO and the brain: Loneliness and problematic social networking site use mediate the association between the topology of the resting-state EEG brain network and fear of missing out," *Comput. Human Behav.*, vol. 141, Art. no. 107624, 2023.

[3] Y. Sun, H. Wang, and S. Bo, "Altered topological connectivity of internet addiction in resting-state EEG through network analysis," *Addict. Behav.*, vol. 95, pp. 49–57, 2019.

[4] Q. He, O. Turel, and A. Bechara, "Brain anatomy alterations associated with social networking site (SNS) addiction," *Sci. Rep.*, vol. 7, Art. no. 45064, 2017. doi: 10.1038/srep45064

[5] T. Ehsan and J. Basit, "Machine learning for detecting social media addiction patterns: Analyzing user behavior and mental health data," *Int. J. Inf. Sci. Technol. (IJIST)*, vol. 6, no. 4, pp. 1789–1807, Oct. 2024.

[6] M. Akter, K. F. Ritu, M. T. Habib, M. S. Rahman and F. Ahmed, "A Machine Learning Approach To Predict Social Media Addiction During COVID-19 Pandemic," in *Proc. Int. Conf. Appl. Artif. Intell. Comput. (ICAAIC)*, Salem, India, 2022, pp. 401-405, doi: 10.1109/ICAAIC53929.2022.9793193.

[7] M. Mardiah and K. Kusnawi, "Analysis of Social Media Addiction: A Comparison of the Performance of Linear Regression and Random Forest Algorithms in Predicting User Behaviour," in *Proc. 12th Int. Conf. Inf. Commun. Technol. (ICoICT)*, Bandung, Indonesia, 2024, pp. 411-418, doi: 10.1109/ICoICT61617.2024.10698019.

[8] Z. Sun *et al*., "KuaiSAR: A unified search and recommendation dataset," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manag. (CIKM)*, New York, NY, USA, 2023, pp. 5407–5411. doi: 10.1145/3583780.3615123.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS*, Red Hook, NY, USA, 2017, pp. 4768–4777. 2017, pp. 4766–4775.