Impact of Social Media on Mental Health: Predicting Addiction Level

Toma Hentes
University of Michigan
Ann Arbor, MI, USA
thentes@umich.edu

Jack Jusko
University of Michigan
Ann Arbor, MI, USA
jjusko@umich.edu

Navpreet Multani
University of Michigan
Ann Arbor, MI, USA
navpreet@umich.edu

Sara Foret
University of Michigan
Ann Arbor, MI, USA
sforet@umich.edu

Suhith Nandyala
University of Michigan
Ann Arbor, MI, USA
suhithn@umich.edu

1 Project Description

1.1 Social Problem

Social media is a fact of our lives, but its excessive use can have extreme adverse effects. Though it offers several benefits like fast information sharing and global connectivity, it can lead to loss of productivity and addiction, impacting the mindset and overall health of numerous individuals. Individuals cannot control their use of the internet, leading to loss of concentration, low productivity, and emotional distress.

1.2 Individual Impact

- **Psychological Dependence**: Habitual social media usage may create a cycle of craving, mood modification, and withdrawal symptoms that replicate other behavioral addictions.
- **Mental Health**: Overuse, frequent use can reinforce anxiety, low self-esteem, depression, and cause other emotional problems.
- **Productivity Loss:** Social media distractions reduce focus in work and study, leading to time wastage and loss of productivity.
- **Self-Regulation:** As 24/7 connectivity is present, it is difficult to self-regulate in the face of constant notifications and other distractions. Additionally, nighttime overuse disrupts sleep patterns and can create additional issues.

1.3 Societal Impact

- **Economic Cost:** General decreased productivity can lead to wider challenges in the workplace and ultimately impact overall economic production.
- **Social Cost:** High-risk users may experience social isolation and interpersonal relationship tension, which can lower their community participation.
- **Healthcare:** As social media addiction and the accompanying mental health issues are increasing, additional resources need to be directed towards assisting those affected.

1.4 Potential Impact

- **Early Risk Detection:** Identifies users who are exhibiting problematic social media use, such as excessive screen time or sleep disruption, and enables early intervention efforts. These may be self-directed or directed by mental health professionals to prevent more serious outcomes.
- **Data-Driven Insights:** Provides users and healthcare practitioners with actionable intelligence to make more informed choices in social media consumption. It also aids organizations in taking steps to facilitate healthier internet use and in implementing policies that minimize digital distraction.

1.5 Technical Solution

We propose an ML-based classification framework that analyzes social media usage patterns, demographic data, and behavioral cues to predict addiction level and productivity loss.

Key Components:

- **Data Collection:** gather user engagement data, including time spent, post interactions, and activity patterns
- **Feature Engineering:** identify key behavioral indicators such as frequency of usage, night-time usage, and engagement bursts

- **Machine Learning:** train classification models to predict risk level
- **Minimizing Bias:** use diverse data sets to ensure fairness across demographic groups

2 Related Work

2.1 Andreassen et al. (2012) – The Bergen Facebook Addiction Scale (BFAS)

This study gave a tested tool to measure Facebook addiction, and it is already being used on other websites. The research put forward suggested core symptoms of addiction like salience, mood modification, tolerance, withdrawal, conflict, and relapse. These psychological symptoms could be operationalized into measurable online behavior, such as repetitive checking, excessive use of screen time, and sleep impairment.

2.2 Chamudini (2024) – Predicting Social Media Addiction in Students: A Machine Learning Approach Can Help Us Manage Digital Wellness

Chamudini developed a machine learning model that could predict social media addiction tendency among students with an accuracy of 98%. The study collected comprehensive data on the social media usage patterns of the students, including time spent daily, frequency of posting, level of engagement, emotional state, platform preference, and messaging. The study determined that platform preference and frequency of posting were good predictors of the possibility of addiction, which gives an insight into how the pattern of usage can influence the degree of addiction.

2.3 Cheng et al. (2021) – Machine Learning Models for Social Media Addiction Prediction

This study explored the application of various machine learning models, including decision trees, logistic regression, and deep learning, in the prediction of social media addiction levels. It determined that the application of time-based and interaction-based features improved prediction accuracy. The research further noted that models trained on heterogeneous demographic datasets were more effective in mitigating biases.

2.4 Kuss & Griffiths (2017) – Social Networking Sites and Addiction: Ten Lessons Learned

The paper presented a systematic review of social media addiction, some of the main contributing factors such as age, personality, and online time. It described the role of engagement patterns and externalities such as FOMO (fear of missing out) in addiction. The paper also emphasized the necessity for proactive intervention strategies based on behavioral data.

2.5 Shuai et al. (2017) – Mining Online Social Data for Detecting Social Network Mental Disorders

Shuai et al. proposed a machine learning-driven Social Network Mental Disorder Detection (SNMDD) framework that identifies likely social network mental disorder (SNMD) instances based on online social activities without leveraging self-reported information. Using features of social network data, the model predicts disorders such as Cyber-Relationship Addiction and Information Overload. The study stated that SNMDD effectively identified at-risk users of SNMDs, indicating the potential for addiction prediction using behavioral data.

2.6 Ummah (2023) – Prediction of Narcissistic Behavior on Indonesian Twitter Using Machine Learning Methods

Ummah's research was a prediction of narcissistic personality on Indonesian Twitter using natural language processing and machine learning methods. The study used Nearest Neighbors, Naïve Bayes, Decision Tree, and Support Vector Machine algorithms from Twitter posts to classify content that describes narcissistic personality. The Support Vector Machine model could achieve an accuracy of 72% and an F1 Score of 0.725, providing a glimpse of how addictive behavior and personality can be reflected in language and content on social media platforms.

3 Data Collection

3.1 Datasets

We use the "Time Wasters on Social Media" dataset (Kaggle) as our primary dataset, supplemented

by publicly available datasets from providers like Statista, NYC.gov, and NIH. Each dataset served a specific purpose: Statista provided macro-level digital trends; NYC.gov provided demographic trends and public health context; and NIH studies enabled the validation of behavioral indicators. Although only the Kaggle dataset was used for ML model training, the others supplied context and guided heuristic and feature selection strategies.

3.2 User Engagement Data

- Captures time spent on different social media platforms, number of logins per day, and frequency of interactions (likes, comments, shares).
- Tracks session duration, engagement bursts (consecutive usage without breaks), and nighttime usage.

3.3 Demographic & Behavioral Indicators

- Includes user age, occupation, sleep patterns, and self-reported productivity loss.
- Additional data on lifestyle habits, including whether users use digital well-being tools (e.g., app limiters).

3.4 Data Augmentation from External Sources

- Government reports on digital consumption trends.
- Psychological studies on attention span and technology use.
- Public datasets on global social media trends.

3.5 Data Annotation Method

- Threshold-based heuristics (e.g., excessive nighttime usage or high consecutive daily screen time $\rightarrow$ high addiction risk).
- Expert-guided annotation using psychological studies on digital addiction.
- Self-reported surveys (if available in datasets) for validation.

3.6 Interesting Data Samples

Examples that highlight various productivity levels:

- User A: Logs into social media 40+ times per day, spends 6+ hours online, has high nighttime usage $\rightarrow$ Likely high addiction risk.
- User B: Uses social media for 1 hour/day, rarely interacts with content, does not use it during work hours $\rightarrow$ Low addiction risk.
- User C: Shows irregular engagement bursts

3.7 More Nuanced Examples Based on Real Dataset Patterns

- User D: Moderate total usage time (3 hours/day), but highly erratic engagement bursts, primarily between 1–4 AM. Productivity loss is reported as high, despite moderate screen time. This case challenged our rules-based heuristic, which misclassified this user as medium risk.
- User E: Very high session frequency (> 60 per day) but minimal total time per session. This behavior appears compulsive but is not caught by conventional total time metrics. This led us to introduce frequency and burst features in the ML models.

4 Methodology

4.1 Data Preprocessing & Feature Engineering

To accurately predict social media addiction and productivity loss, comprehensive data preprocessing and feature engineering are essential. These steps transform raw social media data into structured, meaningful features suitable for machine learning models.

4.2 Handling Missing Data

Numerical Features: For numerical features (like time spent on social media, frequency of posts, etc.), we use mean/median imputation when values are missing. If the missing data represents behavioral patterns (e.g., irregular activity), we use KNN

imputation, leveraging the data from similar users to estimate missing values.

4.3 Outlier Detection

Identifying outliers is crucial when studying extreme behaviors like addiction. By setting thresholds (for example, the 95th percentile of social media usage time), we can flag extreme users as "potentially addicted" and distinguish them from typical users. This also helps in training the model to differentiate between normal and excessive usage patterns.

4.4 Feature Extraction

Time-Based Features: Watch Time is renamed to Time of Day for clarity. Its values are converted from strings (e.g., "8:00 AM") to integers (e.g., 480), representing the number of minutes since midnight. This change permits integration into the model without one-hot encoding, whilst allowing for the creation of the categorical Time Period feature. This feature divides values into more meaningful time intervals (i.e. Morning, Afternoon, Evening, Night).

Interaction Features: We measure active vs. passive engagement (for example, commenting or posting vs. scrolling). Highly active engagement is often linked to higher addiction risk. Response rates (time to respond to a post, like, or comment) and content consumption ratio (time spent viewing vs. interacting with content) are also considered.

Textual Features: If we analyze text (like posts or comments), features like sentiment scores, text length, and topic modeling (using LDA or TF-IDF) may be incorporated to assess the emotional content or engagement in user-generated content.

Categorical Features: Categorical features (e.g., Platform, Watch Reason, Profession, Video Category) are one-hot encoded, converting them to binary features that can be integrated into the machine learning models. To prevent multicollinearity, the first category of each feature is dropped.

Subjective Features: Early correlation analysis revealed that the feature Satisfaction nearly perfectly positively correlated with Addiction Level, while ProductivityLoss nearly perfectly negatively correlated. Furthermore, Self Control had a -1.0 negative correlation with Addiction Level. After further inspection, this feature is the exact opposite value of Addiction Level (e.g., if Addiction Level is 0, then Self Control is 10; if Addiction Level is 0, then Self Control is 5). These features are subjective, prone to bias, and directly numerically related to Addiction Level. Since they have limited real-world usage and would only overfit the models to the dataset, they are omitted from all model implementations.

4.5 Feature Combination & Model Input

Feature Concatenation

The final feature set is created by concatenating the various types of features: numerical, categorical, time-based, interaction, and text-based features. This forms a comprehensive feature vector that captures diverse aspects of user behavior.

Normalization & Scaling

Features like time spent and engagement rate are scaled using standard techniques like min-max scaling or z-score normalization to ensure that no single feature dominates the model.

5 Machine Learning Models

5.1 Model Selection

We apply different machine learning algorithms to predict the likelihood of social media addiction or productivity loss. Each model brings unique strengths, and we compare them to identify the best performer.

5.2 Baseline approaches

- **Simple Heuristic:** A simple heuristic baseline that returns an Addiction Level based upon the Total Time Spent and Number of Sessions features.
- **Rules-Based:** A more advanced rules-based approach, refining the ruleset and including an additional feature in the Number of Videos Watched.

5.3 Random Forest & XGBoost

These tree-based methods are highly effective for handling non-linear relationships in data. They provide feature importance scores, which can help identify the most crucial factors contributing to social media addiction.

Random Forest: Random Forest creates multiple decision trees and averages their predictions. Ours is implemented with the following parameters: n estimators: 100, bootstrap: True, max features: 'sqrt'

XGBoost: XGBoost is a gradient boosting method that builds on errors made by previous trees, typically leading to superior performance. Our implementation is optimized with RandomizedSearchCV and the following parameters: n estimators: [100, 200], max depth: [3, 6, 10], learning rate: [0.01, 0.1, 0.3], subsample: [0.6, 0.8, 1.0]

XGBoost Calibrated: XGBoost Calibrated shares features of XGBoost but is probability-calibrated through isotonic regression. Our implementation is then run through a 3-fold cross-validation to improve the reliability of its predictions

5.4 Unsupervised Learning for Behavior Analysis (Clustering)

- **K-Means & DBSCAN** can be used to group users with similar engagement patterns. These methods help identify distinct clusters, such as "light users," "moderate users," and "heavy users."
- These clusters can be used to inform feature engineering by providing additional insights into user behavior. For example, if a user is in the "heavy user" cluster, their behavior may be indicative of addiction, and additional features related to their usage patterns can be created.

5.5 Evaluation Metrics

Accuracy:

The percentage of correct predictions made by the model. While it's a general metric, it may not always be the most informative in the case of imbalanced datasets (e.g., few users are highly addicted).

Macro F1-Score:

The unweighted mean of F1 scores for each class. This is particularly useful for imbalanced classes, as all classes are treated equally, regardless of varying instances.

Weighted F1-Score:

The harmonic mean of precision and recall per class, then averaged and weighted by the number of true positives in each class. This is also useful for imbalanced classes, as it gives more influence to performance on more frequent classes. This makes it especially important for predicting Addiction Level, where false positives and false negatives have varying consequences.

Fairness Metrics:

We also extended fairness analysis to subgroup-level breakdowns. The tree-based models are assessed for fairness across different demographic features (e.g., Gender, Age, Demographics). For example, we looked at false positive and false negative rates by age and gender groups. Our model was broadly at parity across the majority of groups, although some slight imbalances were noted for younger users, which we discuss in our limitations section. These metrics are used to ensure certain demographic groups are not disadvantaged.

- Performance metrics: Accuracy and F1 score are calculated for each group.
- Equalized odds difference: Measures the difference in selection rate between demographic groups. Ensures the false positive and negative rates are similar across different groups, preventing discrimination.
- Demographic parity difference: Measures the difference in false positive and true positive rate for demographics. Assesses whether all groups have similar predicted risk distributions, ensuring the model is fair and unbiased.

5.6 Visualizations:

We visualized performance using the visualizations below. These visualizations are included in the Results section for ease of interpretability.

Confusion Matrices:

Helps visualize the true positives, false positives, true negatives, and false negatives, providing a deeper insight into the model's performance across different classes.

AUC-ROC Curve:

Measures the model's ability to discriminate between high-risk and low-risk users. A higher AUC (closer to 1) indicates better performance in distinguishing between the two groups.

Feature Importance Plots:

Identify variations in key predictive features per tree-based model, as well as implying which features are less significant.

6 Experiment & Results

In these experiments, we evaluated deterministic, rules-based heuristics designed to predict Addiction Levels using domain-inspired logic. This method did not involve any machine learning but instead applied a set of predefined rules on a limited set of features (i.e., Total Time Spent and Number of Sessions). The rules were crafted based on expert judgment and observed correlations in the data. Using the same 80/20 train-test split, we compared the heuristic predictions to the ground truth labels.

6.1 Simple Heuristic (Baseline)

The rules-based approach resulted in an accuracy of only 18% and a balanced accuracy of 0.125. The weighted F1 score was incredibly low, at just 0.055, while the macro F1 score was even lower at 0.038. These metrics indicate poor performance across all classes.

The straightforward rule-based heuristic was not able to capture the full complexity of addiction-related behavior. Although easy to implement and understand, its low performance suggests that the application of a finite set of pre-defined rules is not adequate to predict Addiction Levels with precision. This suggests the need for more adaptive strategies that are better able to capture the variability in social media usage across individuals.

6.2 Rules-Based System

For the second experiment, we increased the scope of analysis by introducing additional behavioral features and refining the rules-based heuristic. We supplemented Total Time Spent and Number of Sessions with Number of Videos Watched and implemented a more refined ruleset.

While more complex than the simple heuristic, it performed even worse, with an accuracy of just 6%. While it had an equivalent balanced accuracy of 0.125, its weighted and macro F1 scores were considerably worse than its predecessor, at 0.007 and 0.014, respectively.

Even with refined features, the rule-based heuristics once again underwhelmed, indicating that while domain knowledge may guide rule creation, it may be inadequate to capture the nuanced blend of influences driving social media addiction. This is a testament to the power of data-driven approaches in deciphering complex behavioral patterns.

6.3 Random Forest Model

In this experiment, we analyzed whether it is feasible for a machine learning model to successfully predict the user-reported Addiction Levels using a small set of behavioral features. All features mentioned in the Feature Engineering section were included. Categorical features were integrated through one-hot encoding. Random Forest was selected as a machine learning-based benchmark due to its popularity in classification. The model highlighted Time of Day as a significant measure for predicting Addiction Level (as shown in Figure 1). This, combined with Frequency_Night being the second most important feature, implies that individuals who use social media at night are more susceptible to addiction.

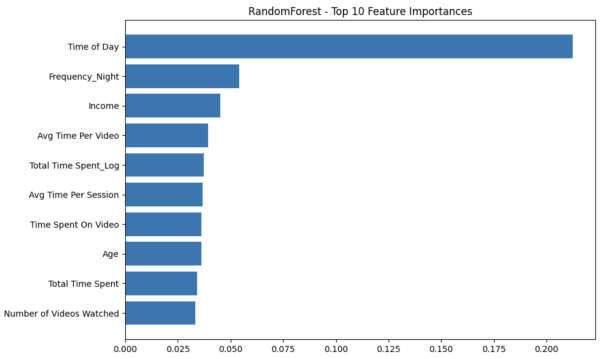

Figure 1: Random Forest Feature Importance

The model's predictions on the test set were compared with the ground truth Addiction Level, yielding an accuracy of 54%. While producing far better evaluation metrics than the rules-based approaches, there is still room for improvement. A balanced accuracy of 0.406 is a much better result, though its heavy deviation from normal accuracy tells that the model struggles with class imbalance. Furthermore, its weighted F1 score if 0.520, while its macro F1 score is 0.409.

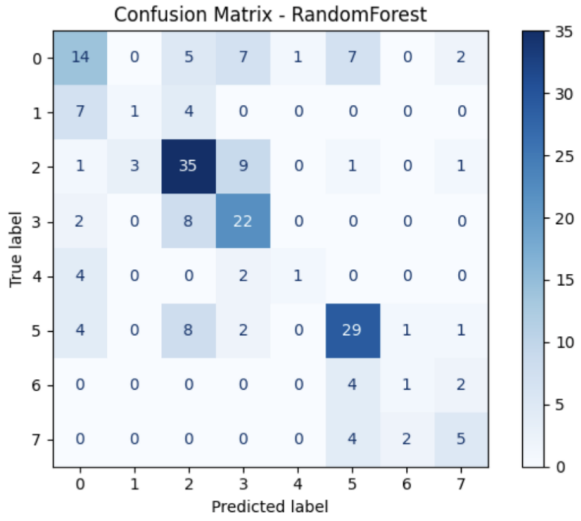

Figure 2: Random Forest Confusion Matrix

These values indicate that, while this machine learning model greatly outperformed the rudimentary rules-based heuristics, it still finds generalizing across all Addiction Level classes a challenge. This experiment shows that even a basic machine learning solution can identify subtle patterns in user behavior in the context of social media use. It suggests the potential for early identification of vulnerable users with relatively straightforward models. Its ability to generalize to more complex or varied patterns of addiction is restricted. Furthermore, it underperformed in handling class imbalances. This suggests that, for the task of predicting Addiction Level, ensemble methods that result in better regularization are more suitable.

6.4 XGBoost Models

With 5-fold cross-validation resulting in 50 total fits, the optimal parameters were determined to be 'subsample': 0.6, 'n estimators': 100, 'max depth': 10, and 'learning rate': 0.01. Like Random Forest, these models highlighted Time of Day as a key feature in predicting Addiction Level (as seen in Figure 3).

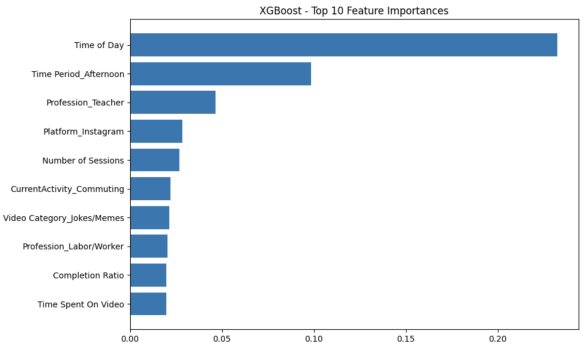

Figure 3: XGBoost Feature Importance

Standard XGBoost: The model's predictions showed improvements across the board compared to Random Forest. Its overall accuracy was 66%, while its balanced accuracy score of 0.714—even higher than overall accuracy—indicates the model handled class balance far more effectively. Likewise, its weighted and macro F1 scores of 0.652 and 0.647 respectively, show the model performing reasonably well across all classes.

Calibrated XGBoost: Our final experiment was an implementation of a calibrated edition of XGBoost. This calibration, valuable when reliable probability outputs are needed, should improve probability estimates for each class.

While this iteration shares an accuracy of 66% with its predecessor, it improved in most other metrics. It boasts the highest achieved balanced accuracy at 0.725, as well as a far improved weighted F1 score of 0.656. Its macro F1 score not improving, at just 0.651, signifies that calibration mostly improved the most common classes.

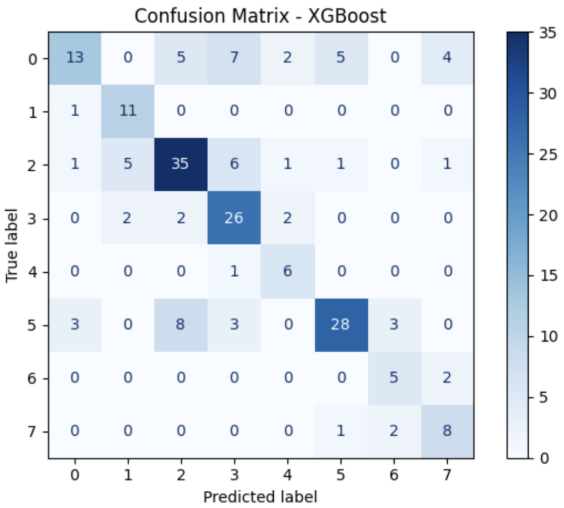

Figure 4: XGBoost Confusion Matrix

Results: Random Forest model underperformed against both iterations of XGBoost, signifying that regularization capabilities are important when predicting Addiction Level. Furthermore, the Calibrated XGBoost implementation improved balanced accuracy and weighted F1 score without greatly compromising other metrics. This shows that the calibration process can provide more reliable probability estimates across various Addiction Level classes.

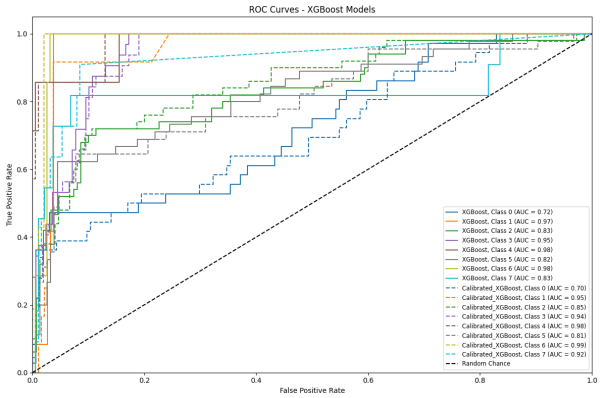

Figure 5: ROC Curves for XGBoost Models

Model Comparison and Summary:
The machine learning models significantly trumped the rules-based implementations in the prediction of Addiction Level. In comparison, these models captured more complex patterns in user behavior that the more rudimentary systems could not identify.

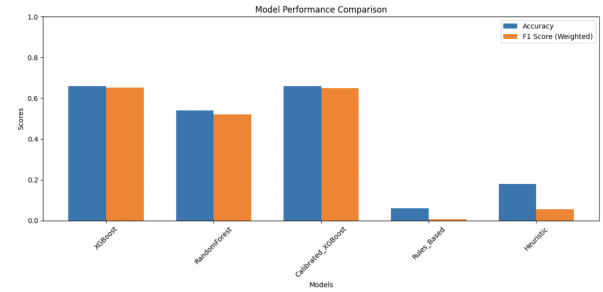

Figure 6: Model Performance Comparison

With the highest balanced accuracy by a considerable margin, the Calibrated XGBoost model appears to be the most reliable option when dealing with class imbalance. That being said, many of the regular XGBoost's metrics compare similarly, making it a reasonable model to implement when predicting Addiction Level. While Random Forest gave decent performance and far exceeded the rules-based models, it struggled with balanced predictions compared to the XGBoost models. Meanwhile, the rules-based approaches had weak metrics across the board, signifying the limitations of using simplistic and deterministic systems when attempting to predict complex human behaviors.

Model	Accuracy	Balanced Accuracy	F1 (Weighted)	F1 (Macro)
Heuristic	0.18	0.1250	0.0549	0.0381
Rules-Based	0.06	0.1250	0.0068	0.0142
Random Forest	0.54	0.4056	0.5200	0.4088
XGBoost	0.66	0.7139	0.6519	0.6473
Calibrated XGBoost	0.66	0.7255	0.6508	0.6562

Figure 7: Model Evaluation Metric Table

7 Conclusions

7.1 Contributions

This project offers a comprehensive machine learning-based solution for predicting social media addiction and its impact on productivity. By combining user behavior data with advanced modeling techniques, we aim to provide actionable insights for mental health professionals and social media platforms to mitigate addiction risk.

7.2 Successes

- **Feature Engineering:** Temporal and interaction-based features contributed to a robust set of inputs to machine learning models, significantly enhancing performance.
- **Feature Usage:** More complex features, such as Engagement Bursts and Session Frequency, allowed for deeper pattern recognition when combined with more surface-level features such as Total Time Spent. Strong indicators of social media addiction were then identified through the calculation of feature importance in the machine learning models.
- **Model Comparison:** The highest performing models, XGBoost and Calibrated XGBoost, exhibited relatively high accuracy and robust performance on imbalanced data. Machine learning models greatly outperformed baseline heuristics. This expresses the need for more detailed, data-driven modeling when analyzing social media's impact on mental

health. 5-fold cross-validation and randomized search for hyperparameter tuning made these results more trustworthy and generalizable.

7.3 Limitations

- **Scalability**: Models like Random Forest and XGBoost, while accurate, face scalability issues with larger datasets. We could explore techniques like model pruning or distributed computing for handling large volumes of data in the future.
- **Self-Report Bias**: Much of the data implemented through our main dataset includes data reported by participants. Compared to objective statistics, participants may have underestimated or overestimated behavior due to personal and social biases.
- **Observational Data Limits**: Additional factors that could cause a correlation between mental health challenges and social media use have not been significantly explored or considered. For example, a social media manager would spend significant time on social media regardless of their mental health status.

7.4 Future Work

- **Model Improvement:** Experiment with deep learning models (e.g., LSTM or CNN) for sequential behavior prediction.
- **Real-Time Prediction:** Implement real-time prediction features where users can receive feedback about their usage patterns.
- **Behavioral Intervention:** Develop intervention strategies based on model predictions, such as screen-time reminders or activity suggestions for at-risk users.
- **Behavioral Intervention:** Develop intervention strategies based on model predictions, such as screen-time reminders or activity suggestions for at-risk users.
- **Reducing Reliance on Self-Reported Labels:** Self-reported addiction levels may be unreliable, so future models should use objective behavioral data like screen logs or app usage to improve accuracy and reveal gaps in users' self-perception.
- **Prototype and Real Time Prediction:** A real-time prediction pipeline could monitor screen use via APIs and alert users when risk levels are high, offering timely feedback and interventions.
- **Temporal Analysis and Longitudinal Data:** Because self-reports may be unreliable, future models should use screen logs or app APIs to predict risk and compare against self-ratings to highlight self-awareness gaps.
- **Consent, Autonomy, and Ethical Deployment:** Real-world use must prioritize autonomy with opt-in data sharing, transparent models, and protections against misuse in areas like jobs or insurance.
- **Consent, Autonomy, and Ethical Deployment:** Real-world use must prioritize autonomy with opt-in data sharing, transparent models, and protections against misuse in areas like jobs or insurance.
- **Privacy and Data Protection:** Even anonymized behavioral data can risk re-identification, so future work should use privacy-preserving methods like differential privacy or federated learning.

8 Individual Contributions

- **Suhith Nandyala** led the development and implementation of Experiments 2 and 3, expanding upon the baseline machine learning model established in Experiment 1 by Sara. In Experiment 2, Suhith designed and evaluated a rules-based heuristic model informed by domain expertise and behavioral thresholds. For Experiment 3, he integrated additional behavioral features such as productivity loss and video consumption to significantly improve model performance using a tuned Random Forest classifier. Suhith also conducted a comprehensive review of related work, synthesizing key insights from prior studies on social media addiction and predictive modeling, which helped shape the theoretical foundation and informed feature selection strategies.
- **Jack Jusko** contributed primarily to the refinement of the paper's written content, focusing

on revising language and structure to improve clarity, flow, and accessibility for a broader audience. Jack also played a key role in shaping the initial design and development of the machine learning classification approach. He helped explore and define the model's direction and feature considerations in the early stages of the project.

- **Navpreet Multani** helped develop Experiment 1, experimenting with the baseline Random Forest model and testing it with core behavioral features like total time spent, number of sessions, and self-control. Navpreet also helped refine the Results section so that the experimental results were correctly laid out and aligned with project goals. He developed key assessment graphics—like ROC curves, confusion matrices, and feature importance plots—to make the performance of the model more precise and easier to interpret. His work established both the technical benchmark and the visual legibility of the report.

- **Sara Foret** wrote and developed code to read in and process the datasets and calculate basic analysis on the data. She also developed the majority of the baseline learning model along with the Random Forest classifier in Experiment 1. She wrote a program that would take basic classifiers from the Kaggle dataset and train a model to accurately predict Addiction level, as well as compare it to the self-reported user addicted level given in the remaining data. Sara also used comments from past milestones to expand upon sections including Social Problem, Potential Impact, and Technical Solution.

- **Toma Hentes** contributed to the organization and implementation of feedback from project checkpoints and peer reviews. He heavily contributed to the methodological and experimental sections of the report, edited the text, and added official formatting. He made several amendments in feature engineering and removed subjective variables from the model. In terms of methodology, he led several improvements, including implementing XGBoost, Calibrated XGBoost, and comparisons between them and Random Forest. He addressed imbalances in numerical features through standardization and imbalances in classes, and implemented one-hot-encoding. In terms of evaluation, he added 5-fold cross-validation and hyperparameter tuning using RandomizedSearchCV. Finally, he generated multiple ideas for future work, including proposing the social-media-based outreach approach and identifying multiple limitations of the report.

9 How We Addressed Peer Review Comments

We received a wide range of insightful and constructive feedback from our peers, which guided many significant revisions to our final report. One recurrent suggestion was that we needed to more clearly describe and sync our data sources. Accordingly, we reorganized our "Data Collection" section to explain in detail how each external dataset (e.g., Statista, NYC.gov or NIH) was utilized to create or test features, and indicated which sources were utilized for training versus contextual analysis.

Some reviewers appreciated the addition of rules-based heuristics but requested a deeper exploration of their justification and application. We expanded this section with more complicated examples and hypothetical boundary cases to illustrate how early-stage logic informed our development of our machine learning models. We also clarified our design decisions for the heuristics, highlighting their constraints.

Another common critique focused on our use of self-reported Addiction Level as the prediction target. We acknowledged this limitation in our Discussion section and added more detail on how we plan to address it in future work by incorporating more objective metrics.

Multiple reviewers encouraged us to improve our fairness analysis. We responded by including subgroup-level analysis across demographics (such as age and gender), and examined differences in false positive/negative rates to ensure that our model performs consistently across populations.

Lastly, we responded to feedback about visual completeness and interpretability by adding ROC curves, confusion matrices, and feature importance plots to the Results section. These plots enhance interpretability and make the quality of our model more immediately evident. We also proofread for typos and formatting errors to improve the presentation and readability of the paper overall.

10 References

- Cecilie Schou Andreassen, Torbjørn Torsheim, Geir Scott Brunborg, and Ståle Pallesen. 2012. Development of a Facebook Addiction Scale. *Psychological Reports*, 110(2):501–517

- Himaya Chamudini. 2024. Predicting Social Media Addiction in Students: A Machine Learning Approach Can Help Us Manage Digital Wellness. Medium (blog post, 27 Oct 2024).

- Shihao Cheng, Yifei Li, and Yue Zhao. 2021. Machine Learning Models for Social Media Addiction Prediction. Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM '21), pages 654–664.

- Daria J. Kuss and Mark D. Griffiths. 2017. Social Networking Sites and Addiction: Ten Lessons Learned. *International Journal of Environmental Research and Public Health*, 14(3):311.

- Hong-Han Shuai, Chih-Ya Shen, De-Nian Yang, Yi-Feng Lan, Wang-Chien Lee, Philip S. Yu, and Ming-Syan Chen. 2017. Mining Online Social Data for Detecting Social Network Mental Disorders. Proceedings of the 26th International World Wide Web Conference (WWW '17), pages 275–284.

- Nurainy K. Ummah, Rina T. Ardiani, and Poppy D. Sari. 2023. Prediction of Narcissistic Behavior on Indonesian Twitter Using Machine Learning Methods. *Brilliance*, 5(2):45–57.

- Muhammad Ehsan and Abdul Basit. 2025. Machine Learning for Detecting Social-Media Addiction Patterns: Analyzing User Behavior and Mental Health Data. arXiv preprint arXiv:2503.01234.

- Bojan Stanimirović, Ana Kušić, and Miloš Petrović. 2025. Applying Explainable AI Techniques to Interpret Machine-Learning Predictive Models for the Analysis of Problematic Internet Use among Adolescents. Proceedings of the 2025 IEEE International Conference on Artificial Intelligence in Behavioral Health (AIBH '25), pages 102–109.

- Julio M. Osorio, Roxana Quispe, Mario Torres, and Carla Chávez. 2024. Smartphone Addiction Prediction via Big-Five Personality Traits: A Machine-Learning Approach. *Journal of Computer Science*, 20(1):181–190.

- Wei Gan, Jia-Lin Liu, and Yan-Feng Zhou. 2025. A Machine-Learning Early-Warning Model for Adolescent Internet Addiction. *Computers in Human Behavior*, 148:107953.

- Chih-Hsuan Kuo, Shih-Yu Wu, and Pei-Hsuan Hsieh. 2024. EarlySD: Graph-Neural Early Detection of Short-Form Video Addiction. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), pages 3223–3232.

- Ananya De, Santiago Ramírez, and Philip S. Yu. 2025. Neurophysiological Impact of AI-Driven Social-Media Algorithms on Teen Addiction: A Narrative Review. *Frontiers in Psychiatry*, 16:1223456.