PREDICTING DIGITAL ADDICTION PATTERNS WITH MACHINE LEARNING FOR PERSONALIZED MENTAL HEALTH SUPPORT.

Abstract

The prevalence of digital addiction in the 21st century has become a mental health issue, especially among the younger generations who have witnessed a time of unrelenting Internet connectivity and ownership of a smartphone.This paper aims to present a comprehensive analysis of the possible application of machine learning techniques for predicting digital addiction patterns and providing personalized mental health support.This research aims to provide a synthesis of the results from multiple investigations that used supervised learning algorithms such as categorical boosting, random forest, support vector machines and deep learning architectures to classify and predict problematic digital behaviors, conducting a systematic review of the literature and methodological advances of recent empirical studies published.Evidence suggests that ensemble methods such as CatBoost and random forest are able to maintain good predictive power in different populations with ROC-AUC ranging from 0.78 to 0.93.Findings from various studies have shown that key predictive factors are screen time, sleep disruption, social media use rate, anxiety, depression and stress.Moreover, this paper investigates the incorporation of reinforcement learning parameters and behavioral inhibition systems as a computational measure of addiction vulnerability: it is shown that model-based decision making deficits and increased learning rates are strong predictors of internet addiction severity (Young, 1998).Clustering-based risk profiling and digital phenotyping provide a way to explore the translation of these predictive models to personalized intervention frameworks.Ethical issues around privacy, algorithmic bias and the possibility of pathologising normal behavior are discussed in detail.This paper ends by suggesting an integrative approach, integrating real-time behavioral monitoring, longitudinal analysis of data, and personalized feedback mechanisms, providing a blueprint for future research and clinical practice in the digital mental health field.

Keywords: digital addiction, machine learning, personalized mental health, predictive modelling, internet addiction, behavioral classification, digital phenotyping

1. Introduction

The increasing use of digital technologies has revolutionized the way people communicate, work, learn and access entertainment.Although such technological developments have proven themselves to be highly advantageous, they have also raised a new type of behavioral problems known as "digital addiction.It is an umbrella term that covers problematic use of the internet, social media, online gaming and use of smartphones that disrupt functioning and psychological health (Ibitoye et al., 2025).

Digital addiction is an alarmingly prevalent issue all around the world.Based on meta-analyses, the prevalence of problematic smartphone use is estimated to be around 37.1% in the general population worldwide, and the prevalence is unremittingly rising over time (Suh & Yoo, 2025).Among adolescents and young adults populations that have never known a world without high-speed internet—the prevalence is even higher.For example, a study in Bangladesh revealed that 30.1% of the adolescent population engaged in internet usage met the criteria for internet addiction, and 3.77% were classified as severely addicted to the internet (Ali et al., 2025; Thity et al., 2025).These statistics highlight the need for effective screening, prevention and intervention strategies.

Self-report questionnaires like Young's Internet Addiction Test (IAT) and Compulsive Internet Use Scale (CIUS) have been extensively used to measure digital addiction.Although these tools have shown psychometric validity, they have a number of inherent limitations.Recall bias, social desirability effects, and lack of introspective accuracy in automatic or habitual behaviors are the limitations of self-report measures (Khoziasheva, 2023).Furthermore, they give only snapshots of the cross section and not the dynamic and temporal patterns which characterize addictive behaviors.

With the arrival of machine learning, a paradigm change is sweeping across the field of understanding, prediction and action on digital addiction.Machine learning algorithms can find complex, non-linear relationships between multiple risk factors and predict risk on an individual basis, which is different from traditional statistical methods based on predefined theoretical models.These techniques can combine various types of data, such as behavioral data, physiological data, self-report data, and contextual data, to form a comprehensive profile of digital engagement patterns (Suh & Yoo, 2025).

This research is not just an academic endeavor.The capacity to capture the attention of people who are at risk for the development of problematic patterns of digital technology use at an early stage and before they cause significant harm is a key public health priority.Interventions tailored to individual data, in personalized ways, can provide scalable, cost-effective support with a level of consideration for individual differences in vulnerability, usage, and treatment preferences (Ibitoye et al., 2025).

The following are the research questions that this paper tackles.

- Which machine learning technique is best for predicting digital addiction and digital dependence?
- What are the key behavioral, psychological and demographic featuresthat should be considered as predictive features for social media and internet addiction?
- What are the ways to transform predictive models into personalized support systems to mitigate the effect of digital addiction on mental health?

2. Theoretical Framework

2.1 Conceptualizing Digital Addiction

Digital addiction is not monolithic but rather is a number of interrelated yet distinct behavioral phenomena.The most frequently investigated types are internet addiction (IA), problematic smartphone use (PSU), social media addiction (SMA) and gaming disorder.Internet Gaming Disorder is listed as a condition for further research in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR) and game disorder is officially recognized as a diagnosable condition in the International Classification of Diseases, 11th Revision (ICD-11) (American Psychiatric Association, 2013; World Health Organization, 2025).

Although no formal diagnostic criteria for the generalized phenomenon of internet addiction exist, researchers have arrived at a number of core features that define problematic internet use.The main characteristics of behavioral addictions, which also applies to digital addiction, are according to Griffiths (2005): salience (the activity becomes the primary focus of cognitive and behavioral life); mood modification (the activity is accompanied by subjective emotional changes); tolerance (larger involvement is needed for the same effect); withdrawal (aversive emotional and physical states occur when the activity is discontinued); conflict (interpersonal or intrapersonal conflicts are involved in the activity); and relapse (return to previous levels of involvement after abstaining from the activity).

These criteria can be useful in operationalizing the concept of digital addiction in machine learning contexts.Each component aligns with observable behavioral indicators: salience (approximated by the frequency of checking device); tolerance (increasing frequency of screen time over time); withdrawal (physiological or emotional reaction when there is a lack of screen time); conflict (self-report of negative consequences or observer report of impairment) (Chen et al., 2026).

2.2 The I-PACE Model

According to the Interaction of Person-Affect-Cognition-Execution (I-PACE) model developed by Brand and colleagues, this is a complete theoretical framework that explains the development and persistence of certain types of Internet use disorders (Brand et al., 2019) (Figure 1).This model posits that addictive digital behaviors result from the interplay between fundamental characteristics of the person (personality, psychopathology, and biological predispositions), affective and cognitive reactions to internet-related stimuli (craving, attentional bias, and maladaptive coping), and executive functioning (inhibitory control, decision-making).

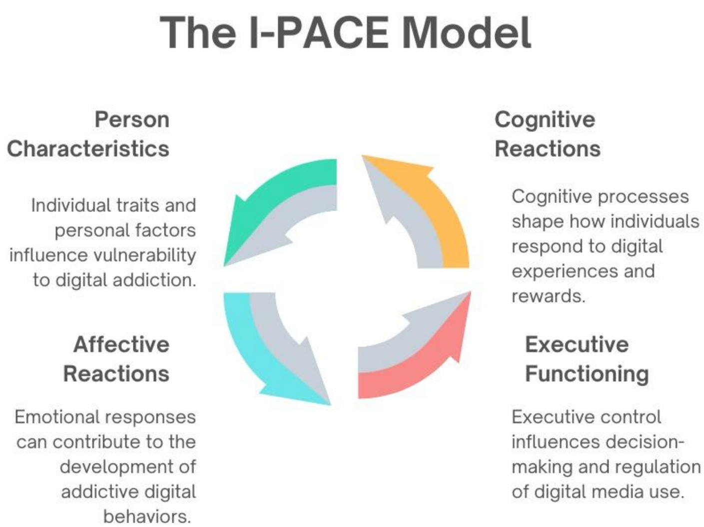

Figure 1. The I-PACE model representation

The I-PACE model is of particular value for machine learning use cases, as it defines causal pathways, which can be formalized as predictive features.For instance, the model suggests that people with high behavioral inhibition system (BIS) sensitivity, that is, high sensitivity to aversive stimuli and punishment cues, may be more prone to develop compulsive internet use as a way to avoid aversive emotional states.Likewise, the model recognizes that low executive control and high cue reactivity are also mechanisms that can sustain addictive cycles (Chen et al., 2026).

2.3 Reinforcement Learning and Addiction

Understanding the decision-making processes that underlie addictive behaviors is important and is provided by the powerful framework of computational reinforcement learning (RL) theory.RL agents learn to choose actions which maximize the cumulative reward by adapting their predictions as a function of prediction errors, which are the difference between expected and observed outcomes (Sutton & Barto, 1998; Daw et al., 2011).

Two complementary learning systems have been found:

A**model-based** (goal-directed) system that simulates the future and guides flexible decision making, and a **model-free** (habitual) system that stores cached values of actions based on the reinforcement history (Figure 2).

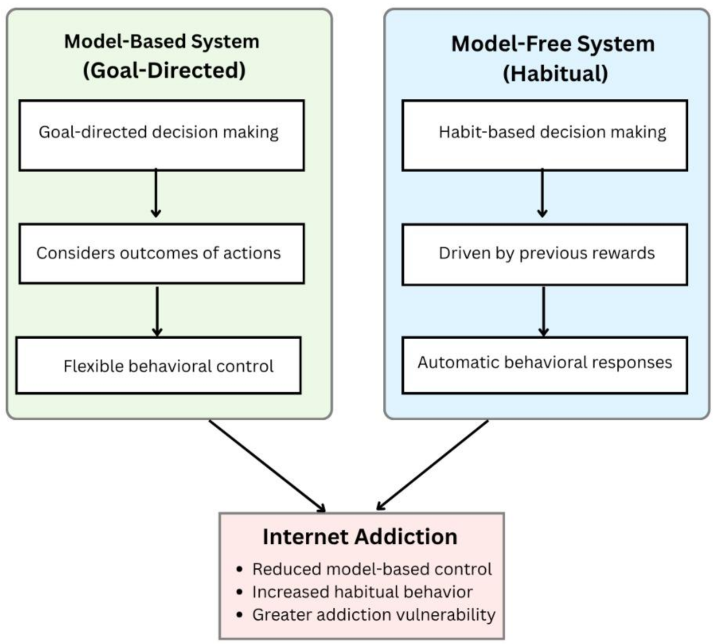

Figure 2. Complementary learning systems of reinforcement learning and addiction

Substance addiction research has shown that chronic drug use leads to inflexible, habitual responding via a switch from the model-based to the model-free system that continues despite negative outcome (Everitt & Robbins, 2016).New research shows that similar processes are involved in digital addiction.Problems associated with internet addiction have been linked to a decrease in model-based control and an increase in habitual responding in sequential decision-making tasks, as found by Liang et al. (2025).In addition, computational modelling showed that patients with internet addiction had significantly greater learning rates than healthy controls, suggesting that they are more sensitive to rewards and better at updating their representation of the value of the reward based on the outcome (Chen et al., 2026) (Figure 2).

These computational results translate to prediction.Parameters of reinforcement learning like the ratio of model-based to model-free control, and the rate of learning, could be used as objective biomarkers of addiction vulnerability, in addition to self-report measures.These computational markers, along with personality indices like BIS/BAS sensitivity, can be used with high accuracy to classify people with internet addiction from healthy controls (Chen et al., 2026).

2.4 Digital Phenotyping and Ecological Momentary Assessment

Digital phenotyping is the "moment-by-moment quantification of the human phenotype at the individual level from the data of personal digital devices" (Onnela & Rauch, 2016).This method takes advantage of the widely available sensing capabilities of smartphones to gather continuous passive information about people's behaviors, such as mobility, communication, application usage, and circadian rhythms.Digital phenotyping allows for dynamic rather than static assessment of behavior, and in real time, rather than a retrospective snapshot of behavior.

Digital phenotyping has several unique benefits for the study of digital addiction.First, it overcomes recall bias by recording objective behavior data when it happens.Second is that it allows temporal patterns and rhythms to be identified that might be diagnostically useful, such as when people use their smartphones most or how often they check their phones at night.Third, it enables the creation of just-in-time adaptive interventions (JITAIs) for providing support at the right moment, when individuals are most susceptible to excessive use or relapse (Suh & Yoo, 2025).

Digital phenotyping is also, however, accompanied by methodological and ethical considerations.The complexity and amount of information gathered passively demand complex analytical methods to extract features and recognize patterns – an area which is very well developed in the field of machine learning.Furthermore, privacy, security, and potential surveillance issues need to be taken into account when considering the implementation of digital phenotyping in mental health applications.

3. Machine Learning Approaches to Digital Addiction Prediction

3.1 Overview of Predictive Modelling Paradigms

Machine learning models for digital addiction can be broadly divided into: supervised, unsupervised, and reinforcement learning models each of which plays a different role in the process of predictive modelling (Figure 3).The most widely used paradigm in this field has been supervised learning, where models are trained using labelled data that is used to predict known results.Classification algorithms classify individuals into risk categories (such as addicted vs. non-addicted, mild vs. severe), whereas regression algorithms take predictions for continuous measures of addiction severity (Thity et al., 2025; Ali et al., 2025).

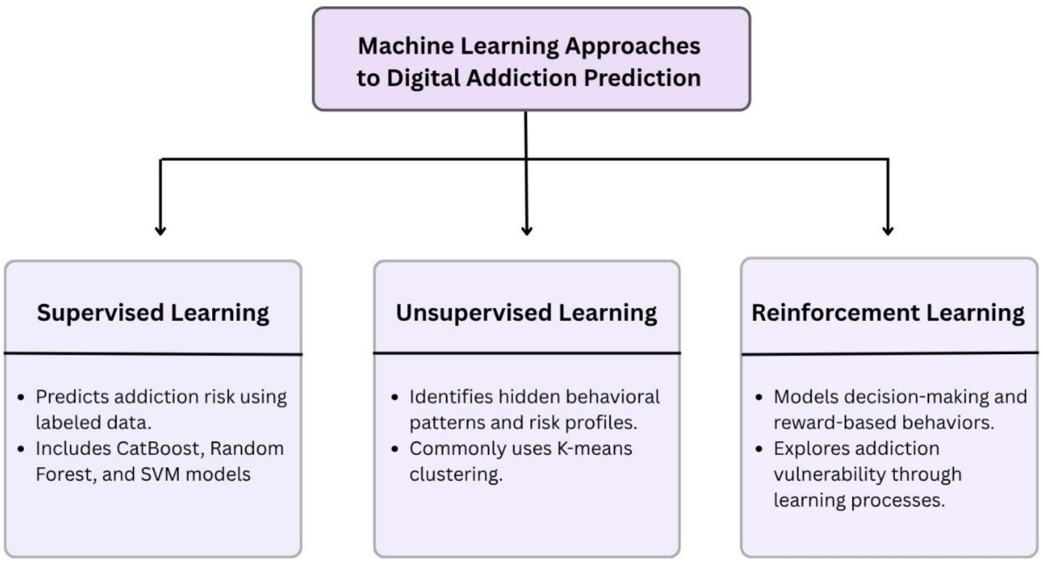

Figure 3.Machine learning approaches to digital addiction prediction

In addition, unsupervised learning methods, specifically clustering methods, can be used to find latent patterns and subgroups in unlabeled data.Personalized intervention is possible through behavioral risk profiles and it has been derived using K-means clustering, hierarchical clustering and density-based methods.For instance, Ibitoye et al. (2025) used K-means clustering to determine different digital addiction risk profiles based on screen time, social media use, and sleep habits, which could then be used to make personalized recommendations according to the specific characteristics of each profile.

Reinforcement learning, although not as widely used directly in the prediction of addiction, is a potentially fruitful approach for creating adaptive policies for intervention.In this paradigm, the agent learns to choose intervention actions (push notifications, content blocking, therapeutic suggestions, etc.) by interacting with an environment (the user) and getting feedback as changes in behavioral outcomes.

3.2 Supervised Learning Algorithms

Various supervised learning algorithms have been tested in predicting the digital addiction and ensemble model has always shown better performance.Table 1 provides an overview of the main results of the latest empirical research.

Table 1. Performance of Machine Learning Models for Digital Addiction Prediction

Study	**Sample Size**	**Target Condition**	**Best Model**	**Performance Metrics**
Ibitoye et al. (2025)	3,200	Digital addiction	CatBoost	Precision: 85.4%, ROC-AUC: 0.93
Ali et al. (2025)	385	Internet addiction	SVM (linear kernel)	Accuracy: 81.9%, AUC: 0.89
Ali et al. (2025)	424	Internet addiction (4-class)	Random Forest	Accuracy: 53.1%, Micro-AUC: 0.78
Likhith et al. (2024)	Not reported	Smartphone addiction	Random Forest	Accuracy: 89.0%
Chen et al. (2026)	122	Internet addiction	SVM	Accuracy: 83.3%

Among them, Random Forest has proven to be a very strong algorithm for the classification of digital addiction.In the study by Thity et al. (2025), random forest (RF) demonstrated the best performance for multi-class classification of the severity of internet addiction with a micro-average AUC of 0.78 compared with decision tree (DT), support vector machine (SVM) and logistic regression (LR).The ensemble aspect of the algorithm, which is based on combining the predictions of multiple decision trees trained on bootstrap samples of the data, naturally generates robustness against overfitting, and allows for modeling high-dimensional feature spaces with complex interactions.

A newer development in gradient boosting theory is Categorical Boosting (CatBoost), which supports categorical features without having to do significant amounts of feature preprocessing.According to Ibitoye et al. (2025), the best performing model among the evaluated models was CatBoost with a precision of 85.4%, and ROC-AUC of 0.93.These findings indicate that CatBoost could be an ideal choice for predicting digital addiction, especially since most datasets tend to have categorical dimensions, such as the platform users prefer, the context where they use it, and the various demographic categories.

Support Vector Machines (SVM) have proven to be highly successful, especially with relatively clear and moderately sized classes.Ali et al. (2025) concluded that SVM using linear kernel significantly outperform logistic regression, decision trees, and random forest in the classification of adolescent internet addiction with an accuracy of 81.9% and AUC of 0.89.By leveraging the SVM's capability to build the optimal separating hyperplane in the high dimensional feature space, it can capture complex pattern in the behavioral data (Ali et al., 2025).

In some comparisons with CatBoost, Gradient Boosting Machines (XGBoost, LightGBM) have also demonstrated good performance but generally did not surpass CatBoost.According to Ibitoye et al. (2025), the performance of XGBoost and graph neural networks was less satisfactory compared to CatBoost but still showed good accuracy performance.The result shows that the choice of the algorithm matters and that boosting methods which take into account categorical features can be beneficial.

3.3 Feature Engineering and Selection

Input features are crucial in determining the predictive performance of machine learning models.Research in this area has found a number of classes of features which have reliably been shown to predict digital addiction.

Behavioral Features: Screen time, defined as the amount of time spent per day or each session, is the most consistent factor found to be a strong predictor of digital addiction (Ibitoye et al., 2025; Suh & Yoo, 2025).The frequency of use of SMM applications, the number of uses per day, and the length of the longest session of social media used are also very predictive.Sleep duration and sleep quality measures, with a focus on decreased sleep due to the use of devices at night, are identified as important secondary predictors (Ibitoye et al., 2025).

Psychological Features: Depression, anxiety and stress scales are highly correlated with digital addiction severity.Among the features selected by the Boruta algorithm for predicting internet addiction among university students, scores from depression, anxiety and stress were among the most important ones identified by Thity et al. (2025).The connection is two-way: mental health issues can lead to a rise in digital use as a way to provide relief and/or coping, while too much digital time can intensify mental health issues.

Personality and Trait Features: Reinforcement sensitivity (BIS/BAS scales) has shown predictive power when it comes to internet addiction.The study by Chen et al. (2026) revealed that BIS scores were significant in predicting the severity of internet addiction, meaning that higher scores on BIS were linked to a greater severity of internet addiction.This is consistent with theoretical models which suggest that those with high BIS may seek digital media as a way to avoid aversive emotional states (Ali et al, 2025).

Demographic and Contextual Features: Age, gender, educational level, father's education and socioeconomic status have been found to have varying predictive significance in different studies.According to Ali et al. (2025), the most significant characteristics to be considered for distinguishing adolescent internet addiction in Bangladesh are father's education status and preferred activities on the internet.In some studies, COVID-19 infection history proved to be an important predictor, accounting for the effects of the pandemic on digital engagement patterns (Thity et al., 2025).

Feature selection is an important technique in improving the performance and interpretability of models.In this space, Boruta is a wrapper method which sequentially checks the importance of a feature by comparing the original with the shadow features which are randomly generated, and it is widely used.Boruta has been shown to accurately select the true set of predictive variables and remove irrelevant or redundant ones, resulting in a more parsimonious and generalizable model (Thity et al., 2025).

3.4 Handling Imbalanced Data and Multi-Class Classification

Digital addiction data sets often have an imbalance of classes, where the number of moderate or severe cases is much higher than the number of normal or mild cases.The prevalence of severe internet addiction among participants in the study conducted by Thity et al. (2025) was low at 3.77%, resulting in a significant class imbalance that could lead to a bias in predicting the majority class.Likewise, Ali et al. (2025) found that 30.1% of adolescents were internet addicted (a lesser imbalance, yet still needing methodological attention).

Class imbalance has been tackled with several approaches in this field.Several resampling methods have been employed such as oversampling of minority classes (e.g., SMOTE – Synthetic Minority Over-sampling Technique) and undersampling of the majority classes, with varying degrees of success.Additionally, cost-sensitive learning, which penalizes misclassification of errors of minority classes more heavily, is an alternative approach that does not have the potential problem of information loss due to resampling (Suh & Yoo, 2025).

Other than binary prediction, there are also challenges for multi-class classification.Thity et al. (2025) used one-vs-rest and softmax output layers for four class classification (normal, mild, moderate, severe).Performance for severe class classification was significantly lower than for binary classification, which was partly due to the small number of severe cases, and partly because of the intrinsic difficulty of discriminating between nearby class categories.This discovery is a reminder of the need to have bigger and more balanced datasets for multi-class applications.

3.5 Model Interpretability and Explainable AI

In addition to predictive accuracy, interpretability is essential for adopting machine learning models in clinical and applied settings.Black-box models could be more accurate but could be deemed unacceptable by clinicians and users for lack of understanding of how predictions are being made.To overcome this tension, a new field of explainable AI (XAI) has emerged with techniques.

Basic level of interpretability is achieved by feature importance analysis that quantifies the contribution of each input feature towards the model prediction.Ibitoye et al. (2025) found that excessive screen time, checking social media too often, and getting too little sleep were the most important factors to consider, based on permutation importance and SHAP (SHapley Additive exPlanations) values for each factor.Tree-based models such as random forest or CatBoost have built-in feature importance metrics that help capture the importance of each feature based on impurity reduction or gain.

Coefficient interpretation gives direct insight into the direction and the magnitude of the effects of features for linear models.Ibitoye et al., (2025) used linear regression and other more complex models to provide interpretable links between the behavioral variables and the risk of addiction.A hybrid model that integrates high-accuracy prediction models with interpretable models is a practical approach to achieving competing goals (Ibitoye et al., 2025).

Structural equation modelling (SEM) can be used as an alternative method to understanding the causal pathways of digital addiction.Anxiety and depression were identified as mediator variables between digital behaviors and addiction risk in Ibitoye et al. (2025) found using SEM.Although SEM itself is not a machine learning method, the incorporation of SEM into predictive modelling is a significant trend towards hybrid approaches to analytical techniques.

4. Key Predictors and Risk Factors

4.1 Behavioral Predictors

Screen time duration has been the strongest and most consistent predictor of digital addiction among studies (Figure 4).The results of Ibitoye et al. (2025) showed that excessive screen time is the most important feature in their CatBoost model, significantly higher than all the other features.This is in line with the part of addiction that involves tolerance, or enhanced dependence on something to attain the same pleasure or mood-altering effect.

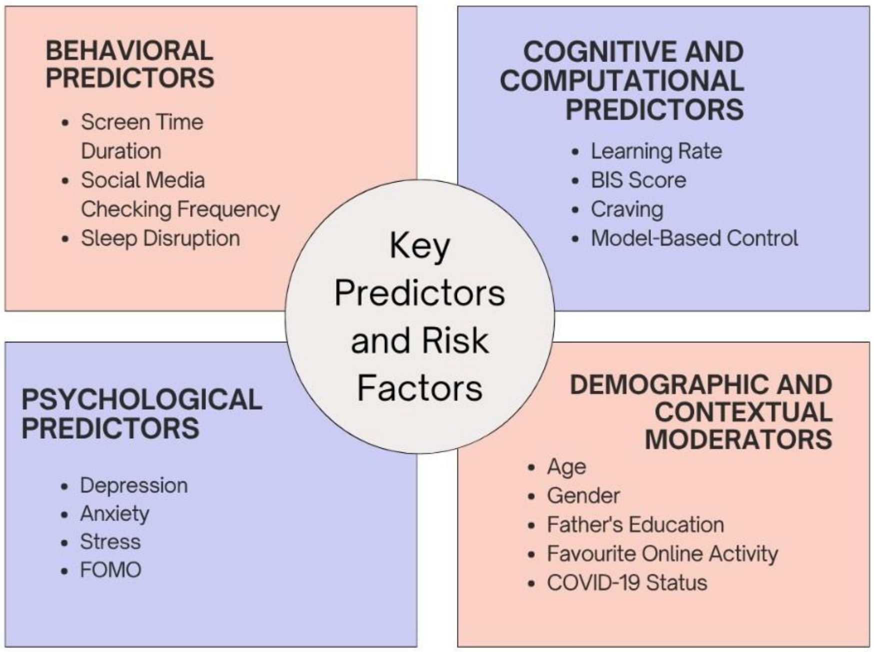

Figure 4. Key predicators and risk factors

But watching the TV isn't enough on its own.Patterns of use (how often they are checked, where they are used, and multi-tasking activities) have added predictive value.Checking social media frequently, operationalized as number of applications opened per hour or percentage of waking hours with the device within reach, was the second significant feature in the model by Ibitoye et al. (2025).It mirrors the intermittent reinforcement schedules that are typical of most social media platforms, leading to strong habit formation (Ibitoye et al., 2025).

Sleep disruption is a third significant behavioral predictor.Across studies, there is a consistent prediction of the severity of digital addiction by reduced sleep duration, difficulty falling asleep after using the device, and nocturnal awakenings to check the device's notifications.This relationship is bidirectional and can be self-reinforcing: digital engagement could be displacing sleep and sleep deprivation could also contribute to compulsive checking behaviors (Ali et al., 2025; Suh & Yoo, 2025).

4.2 Psychological Predictors

A wide range of psychological correlates of digital addiction have been examined and the most important mental health factors are depression, anxiety and stress (Figure 4).The results from the study conducted by Thity et al. (2025) indicated that moderate to severe depression, anxiety, and stress (DASS-21) were the significant features for predicting internet addiction among the university students.These psychological factors were consistently selected as relevant, using the Boruta feature selection algorithm, indicating the close relationship between digital addiction and general mental health (Thity et al., 2025).

How psychological distress is connected to digital addiction is complicated.Some people may find that too much Internet time is a maladaptive coping mechanism, or a way to avoid negative emotional feelings, or to avoid problems in the real world (Arpaci, 2023).Social media interactions, gaming milestones or new content can offer immediate gratification that can temporarily help to reduce dysphoria, encouraging the behavior through negative reinforcement.This coping mechanism can eventually become ingrained and continue to be used even though it can have harmful effects in other areas of life (Chen et al., 2026).

A more specific psychological predictor, relevant to social media addiction is fear of missing out (FOMO).Those who suffer from FOMO are constantly checking out, and they may spend more time on the platform because they feel like they are missing out on something rewarding.Although it is not often evaluated in machine learning research because it is not included in traditional instruments, FOMO is highly correlated with problematic SMC in traditional research.

4.3 Cognitive and Computational Predictors

In recent years, there has been progress in the field of computational psychiatry toward identifying reinforcement learning parameters as potential biomarkers to vulnerability to addiction.Chen et al. (2026) reported that the model-based (goal-directed) control was significantly reduced and the learning rate was significantly increased in individuals with internet addiction (Figure 4).Multiple linear regression showed that learning rate, BIS score, and craving were significant predictors of the severity of internet addiction that accounted for 46.9% of the variance in IAT scores.

These findings have implications for prediction.Reinforcement learning parameters are objective and more likely to reflect the underlying cognitive processing related to addiction, as they are obtained during behavioral task performance rather than by self-report and are retrospective.The two-step Markov decision task employed by Chen et al. (2026) is about 30-40 minutes long and could be administered in the clinic but not in large-scale screening applications.

Computational markers in combination with the traditional markers have potential to increase prediction accuracy.Chen et al. (2026) discovered that a blend of reinforcement learning parameters, BIS scores, and craving ratings yielded an 83.3% classification accuracy with support vector machines, highlighting the complementary nature of these distinct types of predictors.

4.4 Demographic and Contextual Moderators

The significance of demographic variables is different depending on the study and population (Figure 4).Age is always identified as a moderator variable, with the adolescents and young adults being more vulnerable to digital addiction than older adults.Generation Z (those born in the mid-1990s to early 2010s) is especially susceptible because they have grown up with digital exposure and social media was adopted during key times of identity development and peer relationships (Ibitoye et al., 2025).

Gender impacts are less consistent.Some studies have shown that internet addictions are more prevalent in males (especially with regard to gaming), and others that social media addictions are more prevalent among females.These differences may be due to gender differences in online activities, not necessarily in vulnerability to online addiction.These nuanced patterns can be captured by machine learning models that use features specific to activities, not just summary ones.

There are varying degrees of socio-economic and family factors that predict.Father's education was found as an important feature in the classification of adolescent internet addiction which may be related to parental rule-setting and supervision.Another major finding was that favorite online activity (gaming, social media, video streaming or information seeking) was also a major predictor, suggesting that activity-specific assessment is required and that not all digital activity is equal (Ali et al., 2025).

Digital engagement behavior has seen significant changes due to contextual factors, such as the Covid-19 pandemic.Self-reported COVID-19 positive status was also a significant predictor of internet addiction among university students, which may be due to increased use of the internet during lockdown for social interaction, education, or entertainment (Thity et al., 2025).The results highlight the need for time-aware prediction models.

5. Personalized Mental Health Support Systems

5.1 From Prediction to Intervention

The final aim of digital addiction pattern prediction is not only classification but also to enable personalized and effective interventions (Figure 5).There are some intermediate steps necessary for moving from prediction to intervention: risk stratification, identification of modifiable mechanisms, and tailoring of intervention strategies to individual profiles.Machine learning models can be used to stratify risk, which can be used to screen the population and identify those who might benefit from additional evaluation or intervention.Stratified approaches are more efficient in the use of resources than universal approaches, which apply the same strategy to everyone, by identifying and assigning higher level interventions to those most likely to benefit from them.In the case of digital addiction, risk stratification may be used with a brief screening tool and then followed up by algorithmic risk scoring, where the cut-off points are adjusted to maximize sensitivity and specificity as per the specific context of application (Ibitoye et al., 2025).

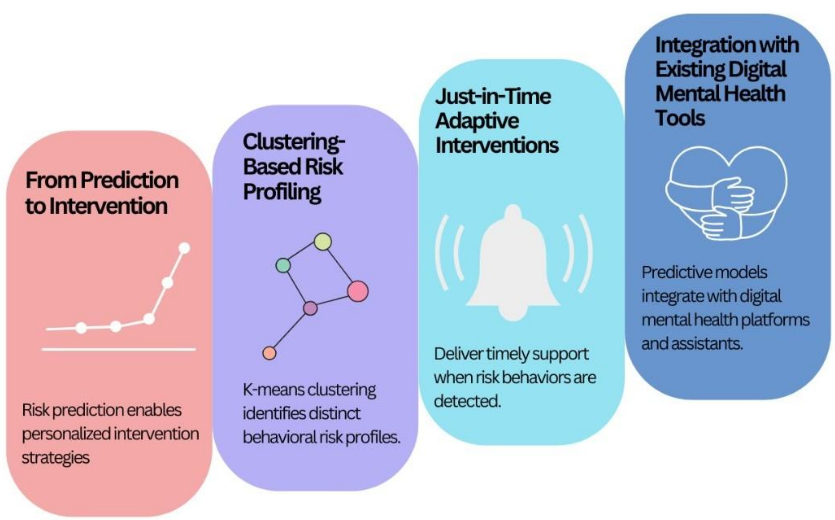

Figure 5. Key points of personalized mental health support systems

For effective interventions, modifiable mechanisms need to be identified.Interpretable feature importance in machine learning models can help users and clinicians interpret the behavioral patterns that are most important to individual risk.Too much screen time might be a key factor for one person; for another, disrupted sleep or anxiety could be more significant.Individual interventions can be specific to the mechanisms that are relevant to each person.

Personalization of intervention strategies is the last step in the personalization pipeline.Ibitoye et al. (2025) adopted the data value K-means clustering for determining their behavioral risk profiles, thereby allowing unique sets of recommendations for each profile.For instance, those in clusters with high levels of screen time and normal sleep were advised on time management and digital boundaries, whereas those in clusters with disrupted sleep were also given extra guidance on their use of screens in the evening and on sleep hygiene.

5.2 Clustering-Based Risk Profiling

There is a need for a systematic approach that can be provided by unsupervised learning techniques, especially clustering algorithms, to identify distinct digital addiction risk subtypes (Figure 5).Ibitoye et al. (2025) used K-means clustering to categorize behavioral and psychological factors and identified multiple distinct cluster profiles with varying behaviors related to social media usage, sleep hygiene, and mental health outcomes, and screen time.

The significance of clustering goes beyond description of typology.There is a potential for different approaches to intervention with different risk profiles.For example, people with a primary motive for addiction related to anxiety and escape might benefit most from alternative coping skills and anxiety management strategies, while people with a primary motive related to habit and automaticity might benefit most from environmental restructuring and implementation intentions.Profile characteristics may be paired with personalized interventions to help increase engagement and improve outcomes.

Cluster analysis can also be used for audience segmentation for digital health interventions.Instead of developing one intervention for all user types, developers can develop multiple intervention versions for different types of users.Automatic classification of users according to screening data can be used to assign them to the best matching variant, thus avoiding the need for intensive individual assessment for scalable personalisation.

5.3 Just-in-Time Adaptive Interventions

Just-in-Time adaptive interventions (JITAIs) are a more sophisticated version of personalized support that provide components of intervention just when users are at the highest risk of relapse/excessive use.JITAIs use data from smartphone sensors and user habits, in real time, to identify risk states and implement the necessary supports (Figure 5).

In digital addiction, the JITAIs could identify behavioral patterns that could signal an increased risk (e.g., longer periods of continuous use, use late into the night, or use after negative emotional events) and provide short-term interventions, such as reminders to pay attention to your current behavior, alternative activities, or limits on use.For example, machine learning models can be trained to identify risk signatures from passive sensor data, allowing for automated risk detection without explicit user input.

However, there are certain technical and behavioral challenges that need to be tackled in developing effective JITAIs.Models that predict or guess the outcome should be designed to be sensitive while not being too specific, so there are no too-numerous false alarms, which defeat the purpose of user engagement.The time of intervention must take into account user context and preferences: a well-timed prompt could be welcomed, an ill-timed prompt could be ignored or disrespected.Reinforcement learning methods, which adaptively update intervention policies according to the users' feedback and results, show potential in addressing these challenges.

5.4 Integration with Existing Digital Mental Health Tools

The predictive framework as presented in this paper is complementary to the current set of tools that exist in the digital mental health space.Examples of emotional support and cognitive interventions include the use of conversational agents, large language models (LLMs), and virtual mental health assistants like Woebot, Wysa, and Replika, which have shown effectiveness in providing emotional support and cognitive interventions through natural language interaction (Ibitoye et al., 2025).These tend to be very good in terms of real-time, engaging support, but generally fail to include structured behavioral data for risk prediction.

Predictive analytics and conversational agents could be made even better by combining them.Predictive models could detect those who are at a higher risk and could cause a proactive approach by the conversational agent, instead of waiting for the user to take the first step.On the other hand, data generated in a conversation, such as linguistic indicators of distress, concerns voiced, and engagement levels, could be used as input features in predictive models, making them more accurate and timely (Figure 5).

There are also other integration opportunities such as integration with electronic health records, wearable devices, and digital phenotyping platforms.Future research and implementation directions include the development of interoperable systems that integrate different sources of data and allow for user control and privacy protection.

6. Ethical Considerations and Challenges

6.1 Privacy and Data Protection

There are significant privacy concerns involved in collecting and analyzing behavioral data for the prediction of digital addiction.Smartphone usage logs, screen time records, application access patterns and contextual data can give intimate details about the lives of people, such as their social links, daily habits, mood states, and potentially sensitive behaviors.This combination of data streams forms detailed digital dossiers that can cause significant damage if mishandled.

This area should have some guiding principles for data use responsibly.Data minimization involves only collecting data that is relevant to the particular predictive purpose, as opposed to collecting all possible behavioral signals.Clear communication to users is essential to transparency, letting them know what data are gathered, how they are used and who has access to them.The user must be able to access, correct, export and delete personal data.To ensure security, it is necessary to employ encryption, access control, and regular security audits (Suh & Yoo, 2025).

Further concerns exist when predictive models are applied in contexts like schools, workplaces, insurance contexts and so on.Even with high error rates, predictions of addiction risk could be used to discriminate against individuals or to thwart opportunities.Regulatory frameworks should include acceptable use cases and prohibited applications to prevent damaging applications and preserve beneficial applications.

6.2 Algorithmic Bias and Fairness

Machine learning models are representative only as much as the data they are trained with.When training data is systematically underrepresented for some populations, based on age, gender, socioeconomic factors, cultural background or geographical region, then the performance of the models can be biased when applied to such populations.These biases could result in systematic overestimation or underestimation of risk for some groups, resulting in inequitable distribution of interventions and/or unnecessary distress.

There is already a bias literature in the field of digital addiction.However, most research has taken place in a particular country (Bangladesh, China, Nigeria and the United Kingdom) and few studies have been cross-validated with different populations.Predictive aspects found in one cultural context may not translate to other contexts where digital engagement patterns, social norms, and psychological frameworks may be different.

To address algorithmic bias, proactive action needs to be taken at various points along the machine learning lifecycle.Diverse data collection is important and should include underrepresented populations, so that models can learn patterns that will generalize across the board.Bias auditing should assess model performance across demographic sub-populations to help uncover disparities.Introduce fairness constraints during model training to ensure fairness of the model's performance.Affected populations should be included in decisions concerning model development, validation and deployment in a community engagement process.

6.3 The Pathologisation of Normal Behavior

One of the key issues within the field of digital addiction is delineating the boundaries between normal and pathological digital behaviors.Digital engagement can become so frequent that it is no longer indicative of disorder, but rather a form of norm and adaptation to the use of digital technologies in the workplace, in education, in social life and in leisure activities.There is a significant risk of over-pathologisation, pathological interpretation of normal behavior.

This risk can be exacerbated by machine learning models that are trained with data that lumps high use with addiction.Models can only learn to predict high use, and not necessarily problematic use, if core features of addiction (continued use despite negative consequences, loss of control, withdrawal symptoms) are not carefully operationalized.An additional worry is that the three variables that are most correlated to risky behavior in existing models time spent on screens, checking frequency, and sleep disruption are also associated with higher, but non-problematic, use in many other contexts.

It is important for researchers and developers to make a distinction between risk prediction and diagnosis.Predictive models can help to identify individuals who might benefit from additional evaluation or low intensity preventive interventions without suggesting that the individual has a disorder.When communicating about model outputs, careful, non-stigmatizing language should be used, normalizing the spectrum of digital engagement and recognizing that there are indeed concerns around problematic digital interactions.

6.4 Clinical Validation and Implementation

Machine learning models need to be well validated beyond their predictive capabilities in order to be transferred from their research context to clinical and applied practice.Clinical utility of this use of the model (does this use of the model enhance patient outcome versus other approaches) should be established in prospective studies.Feasibility of implementation—ability of the models to be implemented in real-world situations with costs, user burden and technical requirements that are acceptable must be determined.

The majority of existing evidence of prediction of digital addiction comes from convenience samples of retrospective studies.Although these studies offer important proof of concept, they do not prove that predictive models lead to better clinical outcomes.Future research is required to determine clinical utility that will involve the identification of at-risk individuals and the evaluation of the effects of model-guided interventions over time (Ibitoye et al., 2025).

Human factors are important to consider when integrating predictive models into current mental health processes.Clinicians need to be aware of the limitations and the appropriate use of models.Clear actionable information to users regarding their risk level, but without causing unnecessary anxiety or fatalism.Systems need to provide a means of human oversight and override of algorithmic choices, especially when they have substantial consequences.

7. Future Directions and Recommendations

7.1 Methodological Advancements

Future research on digital addiction prediction could benefit from some methodological improvements to increase the rigor and impact of research.Future research using longitudinal designs with multiple assessments over long periods of time would allow for the study of within-person changes in digital engagement and risk for addiction rather than cross-sectional associations and potentially uncover causal dynamics.Triangulation across modalities of measurement and richer feature sets for prediction could be achieved with multi-modal data integration, involving self-report, passive sensing, physiological measures, and neurocognitive assessments.

Standardization of outcome measures would make it easier to compare and meta analyze between study outcomes.The Internet Addiction Test (IAT) is broadly utilized, but there is significant variation in cut-offs, scoring and administration protocols.An agreement on core outcome measures in the field of digital addiction research will help to speed up the accumulation of knowledge and facilitate the more rigorous testing of predictive models.

To evaluate generalizability and prevent overfitting, predictive models should be tested on independent external data sets.Models are often internally cross-validated, and not evaluated with data from other populations, time periods, or data collection techniques.Independent replication and validation efforts would be facilitated by sharing of models, for example in open-source repositories.

7.2 Integration with Broader Digital Mental Health Ecosystems

Developing digital addiction prediction should not be done in isolation, but should be an integrated part of the overall digital mental health efforts.The models could be integrated into current digital applications (such as social media, operating systems, and healthcare applications) to deliver real-time risk evaluation and intervention provision.EHR integration could allow for risk-informed care coordination and population health management.

Open standards and interoperable data formats would help to integrate systems and platforms.Application programming interfaces (APIs) that facilitate secure, privacy-aware data sharing would enable predictive models to make use of applicable and relevant behavioral data without demanding every application construct its own information sensing system.

7.3 Personalized Intervention Development

An important priority is the transfer of predictive models to effective interventions.The systematic evaluation of which intervention components are most effective for which user profiles – component-wise intervention testing – would advance beyond one-size-fits-all intervention to personalized treatment.The adaptive intervention design that varies treatment components according to the response that is observed would maximize outcome while also minimizing burden (Griffiths, 2005).

More than just being available to use when needed, digital phenotyping needs to be integrated into the process of delivering the intervention when needed, so that continuous monitoring and adaptive tailoring can occur.Just-in-time interventions could be triggered whenever there are detected early warning signs such as increased screen time, disrupted sleep or emotional distress without awaiting the onset of full relapse.Predictive models for relapse based on passive sensor data could be used to inform adaptive intervention policies.

7.4 Policy and Regulatory Frameworks

Predictive models of digital addiction have policy implications beyond individual clinical interactions.Predictive models should be subject to regulatory oversight to guarantee that they are safe, effective and fair before being marketed and/or deployed in high-stakes environments.Digital health tool regulation is rapidly changing and models like the FDA's Digital Health Innovation Action Plan offer guidance for a risk-based approach.

The privacy-preserving analytics could be used for public health surveillance to track population-level trends in the risk of digital addictions, discover new risk factors, and assess the effectiveness of public health interventions.Aggregate (non-individualized) data for analytics that do not identify individuals can be used to support policy-making in a privacy-respecting way.

Education and digital literacy programmes are key supplements to technological solutions.Educating people about their own digital engagement style, understanding what early indicators of problematic use are and how to be able to help them manage their digital life themselves without outsourcing the monitoring or restriction, is a way of helping them take control of their digital life.

8. Conclusion

Digital addiction is a mental health issue that is increasing, especially among the younger population and the seriousness and complexity of the issue has been a challenge to tackle using traditional self-report based strategies.The supervised learning algorithms, particularly those based on ensemble models like CatBoost and random forest, consistently demonstrated high predictive accuracy in addiction risk classification, highlighting important features such as time spent on screens, frequency of checking social media, sleep disturbances, and psychological distress.Combining computational markers derived from reinforcement learning tasks with self-reports further enhances the predictive accuracy and provide insights into underlying mechanisms.The transformation of these models into personalized mental health support is progressing with clustering-based risk profiling, just-in-time adaptive interventions, and integration with current digital tools, allowing for scalable and personalized interventions that take into account the heterogeneity of digital addiction and decrease the need for resource-heavy clinical services.Ethical concerns, such as privacy, algorithmic bias and over-pathologisation, however, must be addressed proactively by rigorous clinical validation, diverse data collection, transparent algorithmic model development, and engagement with the community.To conclude, the incorporation of predictive analytics with digital mental health support is a conceptual leap towards dynamic, personalized, and preventative solutions, and as digital technologies become increasingly ingrained in everyday life, well-designed, thoughtfully implemented, and carefully validated machine learning will play a critical role in comprehending and encouraging healthy digital interactions.

References

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders, 5th edn Arlington. *VA: American Psychiatric Publishing*, *10*.

Ali, A., Hosain, M. S., Siddik, M. A. B., Hasan, M., Habib, M. A., Kabir, M. A.,& Mahmud, A. (2025). Classifying internet addiction using machine learning approach: A study among adolescents in Bangladesh. *Public Health Challenges*, *4*(4), e70165.

Arpaci, I. (2023). Predicting problematic smartphone use based on early maladaptive schemas by using machine learning classification algorithms. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *41*(3), 634-643.

Brand, M., Wegmann, E., Stark, R., Müller, A., Wölfling, K., Robbins, T. W., & Potenza, M. N. (2019). The interaction of person-affect-cognition-execution (I-PACE) model for addictive behaviors: Update, generalization to addictive behaviors beyond internet-use disorders, and specification of the process character of addictive behaviors. *Neuroscience & Biobehavioral Reviews*, *104*, 1-10.

Chen, J., Xie, Z. Y., Chen, T. Z., Lu, J. X., Zeng, N. N., & Zheng, H. (2026). Internet addiction from a computational model: the predictive value of reinforcement learning parameters and behavioral inhibition systems. *Current Psychology*, *45*(9), 936.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215.

Everitt, B. J., & Robbins, T. W. (2016). Drug addiction: updating actions to habits to compulsions ten years on. *Annual review of psychology*, *67*(1), 23-50.

Griffiths, M. (2005). A ‘components’ model of addiction within a biopsychosocial framework. *Journal of Substance use*, *10*(4), 191-197.

Ibitoye, A. O., Ravindran, S., Afe, O. F., & Abiodun, A. O. (2025). Predicting digital addiction patterns with machine learning for personalised mental health support. *Journal of Social Computing*, *6*(4), 359-377.

Khoziasheva, A. (2023, April). Building the neural network-based system for identifying the gaming addiction level in children and adolescents. In *Eurasia Business and Economics Society Conference* (pp. 21-38). Cham: Springer Nature Switzerland.

Liang, W. X., Ye, W. Y., Ng, K. X., Dou, K., & Ning, Z. J. (2025). The longitudinal association between internet addiction and prosocial behavior among Chinese adolescents: testing a moderated mediation model. *Behavioral Sciences*, *15*(3), 322.

Likhith, S., Chitteti, C., Dharani, M., Nivedhitha, V., Geethika, N. G., & Godwin, V. (2024, April). Machine learning model for prediction of smartphone addiction. In *2024 International Conference on Expert Clouds and Applications (ICOECA)* (pp. 924-929). IEEE.

Onnela, J. P., & Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*, *41*(7), 1691-1696.

Suh, Y., & Yoo, J. (2025). Risk level prediction for problematic internet use: A digital health perspective. *Internet Interventions*, *41*, 100863.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1, No. 1, pp. 9-11). Cambridge: MIT press.

Thity, N. T., Rahman, A., Dulmini, A., Yasmin, M. N., & Rois, R. (2025). An illustration of multi-class roc analysis for predicting internet addiction among university students. *Plos one*, *20*(7), e0325855.

World Health Organization. (2025). International classification of diseases (11th ed.). World Health Organization.

Young, K. S. (1998). Internet addiction: The emergence of a new clinical disorder. *Cyberpsychology & behavior*, *1*(3), 237-244.