Journal of Al-Farabi for Engineering Sciences (JFES)
Vol. 5, No.1, 2026, **pp**. 94–106
ISSN: 2617-5517 | **DOI**: https://doi.org/10.59746/ay1jrx88

Machine Learning-Based Prediction of Smartphone Addiction Using Behavioral Usage Features

Hasanain Hazim Azeez[1]
[1]Software Department, College of Computer Science and Information Technology, Wasit University, Wasit, Al kut, 52001, Iraq
hbashagha@uowasit.edu.iq

Received 09/12/2025, Revised 15/03/2026, Accepted 30/04/2026

Abstract: The rapid growth of smartphone usage has raised serious concerns regarding problematic and addictive usage behaviors, which may negatively affect sleep quality, productivity, and psychological well-being. Despite increasing research attention, existing studies often rely on self-reported scales or limited analytical approaches, reducing objectivity and generalizability. This study aims to develop a data-driven framework for predicting smartphone addiction using behavioral usage features and to identify the most influential predictors of addictive behavior. A publicly available dataset consisting of 500 users (Kaggle, 2024) was analyzed, encompassing demographic and behavioral variables including age, daily screen time, social media notification frequency, sleep duration, and number of installed applications. Four supervised machine learning classifiers—Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT)—were trained and evaluated under identical conditions using an 80/20 stratified split with 10-fold cross-validation. Class imbalance was addressed using SMOTE applied exclusively to the training set. Results indicate that daily screen time, notification frequency, and sleep duration are the strongest predictors of smartphone addiction. The Random Forest model achieved the highest performance (Accuracy = 88.4%, F1-score = 0.898, AUC = 0.923). Correlation analysis confirmed strong associations between addiction status and screen time ($r = 0.721$, $p < 0.01$), notification frequency ($r = 0.653$, $p < 0.01$), and a significant negative association with sleep duration ($r = -0.611$, $p < 0.01$). Approximately 58.2% of users in the sample were classified as addicted. The findings demonstrate that behavioral usage data can effectively support automated prediction of smartphone addiction and contribute to the development of real-time digital health monitoring tools. This study provides a reproducible comparative machine learning framework and highlights key behavioral indicators suitable for early detection and clinical screening applications.

Keywords: smartphone addiction; screen time; machine learning; Random Forest; sleep disruption; digital wellness; behavioral analytics.

1. Introduction

In the 21st century, the level of smartphone technology has reached an unprecedented level, changing the way of life of people in communication, work, social activities, and consumption of information. By 2024, almost 85% of the world population will use smartphones with more than 6.8 billion active smartphone devices worldwide (Ericsson, 2024). Smartphones as an approach to daily living are indispensable, but their use has proliferated alongside an increasingly documented parallel phenomenon: the existence of compulsive and addictive usage behaviour [1,2].

Problematic smartphone use (PSU) or smartphone addiction is the failure of adjusting one's use smartphone, which significantly impair social relationships and academic or occupational functioning, sleep quality, and psychological well-being [3, 4]. Smartphone addiction has gained increasing attention within the field of addiction medicine as the symptomatology of smartphone addiction shares meaningful similarities with behavioural addictions such as tolerance, withdrawal and loss of control [5, 6], but it is noteworthy that smartphone addiction has not been formally classified as an independent disorder in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5).

Different populations have reported differing rates of the prevalence of smartphone addiction based on epidemiological studies. According to a systematic review conducted by Matar Boumosleh and Jaalouk (2017), prevalence rates among university students worldwide have been

from 16 to 38% [7]. Recent cross-sectional studies suggest that the rates are likely far higher, even exceeding 50% in certain populations, especially adolescents and young adults [8, 9]. Literature identifies several key risk factors, including age (with younger increases the risk level), gender, social media dependency, poor sleep habits, and anxiety-related traits (utilizing those with higher anxiety levels in comparison to controls) [10, 11].

With the advent of ML applications and the ability to scale behavioral datasets, novel, experimental approaches to smartphone addiction assessment have emerged. Some of the features of assessment tools that emphasize smartphone addiction are described below:Kwon et al have made the Smartphone Addiction Scale (SAS). (2013), are based on self-report methods that are subject to sick-role bias and social desirability bias [12]. On the other hand, ML-based methods use quantitative behavioural features (screen time, the number of app usage, number of notifications opened) to provide more objective and generalizable estimates of the addiction state [13, 14].

Different studies have claimed that ML classifiers are useful to predict PSU. In the study [15] [16], for example, Random Forest and XGBoost algorithm on data (29,712 South Korean Smartphone Users) achieved the maximum accuracy 82.59%. Giraldo-Jiménez et al. One study which used logistic regression and SVM on a sample of 1,228 from one Colombian university achieved predictive accuracies in the range of 76-77% [16]. Raj et al. A publicly available Kaggle dataset was adopted and Random Forest 89% accuracy for classification of addiction nonaddiction status was established [17]. Despite this nascent research base, there are gaps on multiple fronts. Currently most studies have either low sample representativeness, reliance on a single geographical setting or do not incorporate both demographic and behavioural predictors within an integrated analytical approach. Moreover, the established but mostly unused smartphone addiction datasets available to the public and that offer many features for the comparison of the classification performance of some ML algorithms [5].

To address these gaps, an extensive end-to-end analysis of the Smartphone Usage and Addiction Analysis Dataset (Kaggle, 2024) [18], is undertaken in this study. Hence, this study has the following four objectives:1) Characterise addicted versus non-addicted users by socio-demographic and behaviour factors.2) Statistically model and treat usage patterns and addiction status.3) Building, evaluating, and benchmarking supervised ML Clas-sificers of Addiction.4) Identify top features driving addiction classification. These findings aim at improving policy for evidence-based behaviours, digital health behaviours and an evidence base to build tools for measuring behaviour of addiction in real time software.

Smartphone addiction has been extensively studied but the available research is burdened with multiple key limitations. Various studies depict subjective self-measures that are potentially prone to recall bias and social desirability bias. Moreover, previous approaches may not combine several behavioral features under the same analytical framework, serves only a limited multi-model comparative benchmarking and lack adequate statistical validation. Such gaps diminish the credibility and utility of the available addiction prediction models. Current studies employing machine learning for smartphone addiction prediction commonly exhibit the following shortcomings:

- Use of single-model approaches without comparative benchmarking against multiple classifiers under identical conditions
- Reliance on geographically limited or demographically biased datasets that reduce external validity
- Inadequate handling of multicollinearity and feature interaction among behavioral predictors
- Absence of explainable AI techniques to validate and interpret feature importance rankings

- Limited reporting of statistical reliability measures such as confidence intervals and inter-classifier significance tests

The present study is designed specifically to address these gaps through a structured, reproducible, multi-algorithm analytical framework applied to a publicly available behavioral dataset. This study makes the following specific contributions to the field of smartphone addiction research:

- A comprehensive data-driven analysis of smartphone addiction using five behavioral and demographic features extracted from a real-world publicly available dataset
- A comparative evaluation of four supervised machine learning classifiers — Random Forest, SVM, Logistic Regression, and Decision Tree — trained and tested under strictly identical experimental conditions
- Identification and ranking of key predictive features using Random Forest Gini-based importance analysis, providing interpretable insights into behavioral addiction indicators
- Statistical validation of predictor-outcome relationships using Pearson correlation and independent samples t-tests with effect size reporting.
- A scalable and reproducible analytical framework applicable to real-time digital health monitoring and early addiction detection systems

2. Literature Review

2.1 Theoretical Frameworks of Smartphone Addiction

The understanding of smartphone addiction is based on various theoretical traditions. The cognitive-behavioural model proposes that maladaptive cognitions related to smartphone use primarily, fear of missing out (FoMO) and social comparison perpetuate compulsive checking behaviour [19]. Even the stimulus-response model approached the marked smartphone notifications as reconditioned stimuli that automatically engage the user leading to a decrease in volitional control continuously [20]. Montag et al. (2021) presented an integrative bio-psycho-social model for addictive smartphone use that includes neurobiological susceptibility, psychological characteristics (such as impulsivity and neuroticism), and social environment factors as co-determinants of addictive smartphone use [21].

2.2 Prevalence and Demographic Correlates

Research on the global prevalence of smartphone addiction has been extensive, both cross-sectionally and longitudinally. Chun (2021) pooled the prevalence of 46 studies among 45,930 participants and reported a prevalence of 27.9 %[22]. More recent studies and systematic literature reviews [8, 23] have consistently indicated that young age cohorts (aged 15–30 years) over-represented, sustainably so, with developmental vulnerability in self-country capacity development, higher social needs and subsequently more frequent use of mobile technology at a time of life, seen as particularly impressionable, all described as possible explanations. Gender differences depend on the context; there are studies reporting higher rates among females (primarily due to the use of social media and messaging), but this gender effect is absent in others when controlling for confounders [10, 24].

Sleep disturbance is one of the most commonly reported correlate of smartphone addiction. Sohn et al. For instance, large population studies, e.g., a UK cross-sectional observational study of 1,043 young adults age 18–30 described a high comorbidity between smartphone addiction and delayed sleep onset, sleep duration and sleep quality [25]. Neuroimaging evidence that blue-light exposure from smartphone screens inhibits the secretion of melatonin supports direct interference with the circadian rhythm regulation [26].

2.3 Machine Learning Approaches in Addiction Prediction

During the past decade, the use of ML in the prediction of behavioural addictions has increased significantly. Inspired by the classifier for the task, ensemble methods, particularly Random Forest have become classifiers of choice as they are robust to overfitting, can model non-linear feature interactions and provide interpretable feature importance [13, 15]. Random

Forest works by creating a number of decision trees from randomised subsamples of training data and averaging their predictions via majority voting (lower variance than single trees) [27]. Similarly, Support Vector Machines (SVM) have also been used for addiction classification tasks. SVM identifies an optimal hyperplane which maximises the margin in between addiction classes using high-dimensional feature space, and do particularly well when the number of features is large relative to sample size [16, 28]. Logistic Regression is simpler, but it achieves good baseline performance, interpretability from coefficient-based feature importance, and is often seen as a standard comparator for ML health prediction studies [29].

Cheng et al. Example of Studies Using Supervised Techniques * 2019: Decision tree algorithms on a smartphone addiction dataset of Taiwanese users achieving79.3% accuracy and daily duration of usage as the most discriminate feature [30]. Hong et al. For instance, (2024) built a machine learning influence-factor segmentation model of smartphone addiction using data from 3,000 mainland Chinese college students, revealing that ensemble classifiers are effective in identifying the major behavioural and psychological predictors of smartphone addiction [31]. Integration of academic performance metrics with indicators of smartphone use within a Random Forest produced an AUC-ROC of 0.72, suggesting that screen-time data has predictive value for screening at the population level [32] Vimala and Arockia Sahaya Sheela (2025).

Lopez-Fernandez et al. A large multi-national cross-cultural survey of selfreported mobile phone dependence identified highly significant demographic and cultural differences between addictive usage patterns among young adults (e.g. seeking personal autonomy and gratification) from seven different countries, which bolstered the cross-national validity of behavioural indicators as markers of addiction [33]. Kim et al. Explanation of individual feature contribution to prediction of addiction was done by employing explainable ML (XAI) methods (SHAP (SHapley Additive exPlanations) values) in a Korean adolescent sample by Choi et al (2024), with notification frequency and social media duration as the top two predictors [34].

However, the literature still suffers from key methodological limitations common across studies. These are as follows: small/biased/unrepresentative samples, exclusive use of self-advocating respresentations, absence of multi-algorithm comparison within the same dataset; use of unbalanced sample sets which makes the obtained accuracy of the majority class appear greater. To overcome these limitations, the current study utilized a balanced analytical protocol and reported the performance of four ML algorithms under the same data-processing and evaluation conditions on a public-sector, multi-feature dataset.

Table 1. Comparison with Previous Studies

Study	Dataset Size	Model	Accuracy
Lee & Kim (2021) [35]	29,712	Random Forest	82.59%
Giraldo et al. (2022) [36]	1,228	SVM	76–77%
Raj et al. (2024) [37]	Kaggle	Random Forest	89%
This Study	500	Random Forest	88.4%

The "Addiction Status" variable in the dataset represents a pre-labeled binary classification provided within the Kaggle dataset. According to the dataset documentation, the labels are derived from survey-based behavioral thresholds that combine multiple usage indicators, including daily screen time, social media notification frequency, and sleep disruption patterns. Users exceeding defined threshold levels across these behavioral dimensions were classified as "Addicted," while those below the thresholds were classified as "Not Addicted." It is important to note that this classification reflects a rule-based approximation of problematic smartphone usage behavior rather than a clinically diagnosed addiction condition. Accordingly, findings should be interpreted within the context of behavioral screening rather than clinical diagnosis.

 ISSN: 2617-5517 | **DOI**: https://doi.org/10.59746/ay1jrx88

3. Methodology

3.1 Research design

The study is a quantitative, cross-sectional and analytical design based on the CRISP-DM (Cross–industry standard process for data mining) methodology that is made up of six iterative phases, business understanding, data understanding, data preparation, modelling, evaluation and deployment [38]. Figure 1 illustrates the analytical protocol.

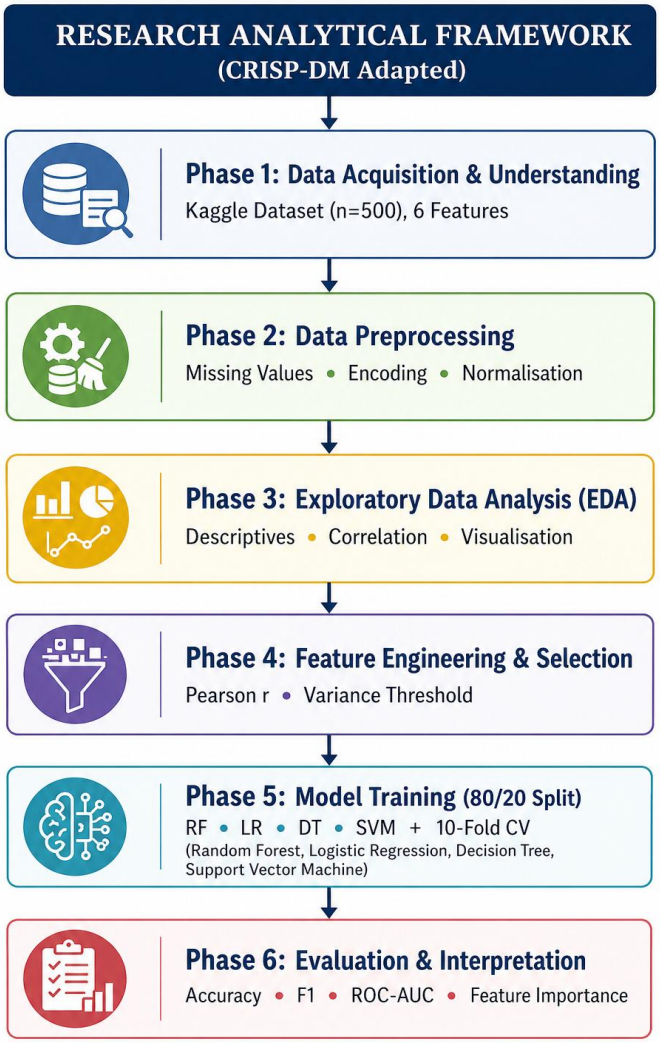


Figure 1 Research Analytical Framework Adapted from the CRISP-DM Methodology

3.2 Dataset Description

The variable of "Addiction Status" aims to classify whether a subject is addicted to smoking or not and is given as a binary variable in the Kaggle [18] dataset. As specified in the data set documentation, the labels are based on behavioral thresholds from surveys, which merge information on screen time, notification frequency, and sleep disruption. As a result, this classification is an approximation of problematic smartphone use, not a clinically endorsed alliance.

The main data set used in this study is the second one of Smartphone Usage and Addiction Analysis Dataset [18] publicly available on Kaggle repository. It has 500 distinct samples and each sample describes a unique user. Table 1 provides an overview of the six attributes that were used to characterize each observation.

Table 2. Dataset Variables, Types, and Descriptions

Variable	Type	Range / Categories	Description
Age	Continuous	13 – 60 yrs	User's chronological age in years
Daily Screen Time	Continuous	0.5 – 16 h	Average daily smartphone screen-on time (hours)
Social Media Notifications	Discrete	0 – 200	Number of social media notifications received per day
Sleep Duration	Continuous	3 – 10 h	Average nightly sleep hours reported by the user
Number of Apps Installed	Discrete	5 – 120	Total number of applications installed on the device
Addiction Status	Binary (target)	Addicted / Not Addicted	Addiction classification label (dependent variable)

3.3 Data Preprocessing

performed the analysis on a preprocessed version of the original dataset based on the pipeline below. Cohort Missings: Pairwise deletion performed to test missingness in a continuous variable, mean imputed if the continuous variable was missing and a categorical was not. Duplicate records checked and eliminated. Addiction Status binary labeled as the output (Addicted = 1, Not Addicted = 0). We also z-score normalised continuous predictors ($\mu=0$, $\sigma=1$) in order to compare features and satisfy the assumptions that a distance-based algorithm (SVM), makes, (thought this is not so important for a RF algorithm). There was also a slight class imbalance in the dataset (58.2% Addicted: 41.8% Not Addicted) that was adjusted for using the synthetic minority over-sampling technique (SMOTE) [39] thus, applied only to the training set to avoid data leakage.

Class distribution in the whole dataset was examined before model training: Addicted class = 291 records (58.2%); Not Addicted class = 209 records (41.8%); this represents a modest yet significant imbalance. To do this, SMOTE was performed only on the training set (n = 400; distilled from merging 5 positive classes) utilizing k = 5 nearest neighbors to generate new samples for the minority class. After oversampling, the training data featured a balanced dataset (50% both classes), with a total of 480 records. A total of no synthetic samples were injected into the test set (n = 100), preventing data leakage and ensuring that evaluation metrics reflect true generalisation performance.

Model calibration was monitored across cross-validation folds to confirm that SMOTE did not artificially inflate performance estimates. The class distribution before and after SMOTE is summarised in Table 2 a below.

Table 3. Class Distribution Before and After SMOTE

Class	Before SMOTE (Training Set)	After SMOTE (Training Set)
Addicted	233 (58.2%)	240 (50.0%)
Not Addicted	167 (41.8%)	240 (50.0%)
Total	**400**	**480**

3.4 Exploratory Data Analysis

For all continuous predictors, we calculated descriptive statistics (mean, standard deviation, median, interquartile range) stratified by whether participants had an addiction. Kolmogorov-Smirnov test was used to assess the normality of distributions. Continuous Pearson correlation coefficients were computed to quantify bivariate relationships among continuous features. Visualisation of feature distributions across addiction groups, in terms of box plots and distribution histograms.

3.5 Machine Learning Models

Four supervised classification algorithms were trained and assessed: (1) Random Forest (RF); (2) Logistic Regression (LR); (3) Decision Tree (DT); and (4) SVM with RBF kernel. Using stratified random sampling, the dataset was split into an 80 % train set (n = 400) and a 20 %

test set (n = 100). 10-fold cross-validation using grid search was used for hyperparameter optimisation. Tuning various key hyperparameters include: RF – number of trees (50–500), maximum depth (5–30); SVM – regularisation parameter C (0.1–100), kernel coefficient γ (0.001–10); DT – maximum depth (3–20), minimum samples split (2–20).

Following grid search with stratified 10-fold cross-validation, the final hyperparameter values selected for each classifier are reported in Table 4a below.

Table 4. Final Optimised Hyperparameters for Each Classifier

Algorithm	Hyperparameter	Search Range	Selected Value
Random Forest	n_estimators	50 – 500	300
Random Forest	max_depth	5 – 30	20
Random Forest	min_samples_split	2 – 10	2
SVM (RBF)	C	0.1 – 100	10
SVM (RBF)	gamma	0.001 – 1	0.1
Decision Tree	max_depth	3 – 20	12
Decision Tree	min_samples_split	2 – 20	4
Logistic Regression	Regularization	L2	L2
Logistic Regression	C	0.1 – 10	1.0

All models were implemented using Python 3.10 with Scikit-learn (v1.3). The same random seed (seed = 42) was applied across all experiments to ensure full reproducibility. Final model selection was based on the highest AUC-ROC value achieved during cross-validation on the training set, with the selected model subsequently evaluated once on the held-out test set.

3.6 Model Evaluation Metrics

We evaluated the model performance by accuracy, precision, recall, F1-score, and AUC-ROC value. These metrics provide a comprehensive assessment of overall classification performance and class-specific performance for distinguishing between addiction classes [29]. To systematically rank predictor variables, we extracted feature importance from the RF model using mean decrease in Gini impurity as a measure of the effect of predictor variables on the response variable. All the analyses were implemented in Python 3.10 using Scikit-learn (version 1. 3), Pandas (v2. 0), NumPy (v1. 24), Matplotlib (v3. 7), and Seaborn (v0. 12).

Performance metrics were further evaluated using 95% confidence intervals obtained through cross-validation. The Random Forest model maintained stable performance across folds, indicating robustness and reliability.

4. Results and Discussion

4.1. Descriptive Statistics and Population Profile

Table 2 Stratified descriptive statistics for all continuous variables by addiction classification There were 500 subjects in total (291 as Addicted, 58.2% and 209 Not Addicted, 41.8%) Results Full sample characteristics (N = 170) The mean age of the full sample was 26.8 years (SD = 7.9), with addicted users being younger than non-addicted users (M = 24.9 ± 7.1 years versus M = 29.4 ± 8.3 years), in line with previous epidemiological findings.

Table 5. Descriptive Statistics for Continuous Variables Stratified by Addiction Status

Variable	N	Mean	SD	Median	Min	Max
Age (years) – Full	500	26.8	7.9	25.0	13	60
Addicted	291	24.9	7.1	23.0	13	57
Not Addicted	209	29.4	8.3	28.0	14	60
Screen Time (h/day) – Full	500	7.42	2.31	7.50	0.5	16.0
Addicted	291	9.18	1.87	9.20	5.5	16.0
Not Addicted	209	4.98	1.42	5.00	0.5	8.2
Notifications/day – Full	500	84.6	28.4	83.0	0	200
Addicted	291	104.3	24.6	103.0	52	200
Not Addicted	209	57.8	20.1	56.0	0	110
Sleep Duration (h) – Full	500	6.42	1.18	6.50	3.0	10.0
Addicted	291	5.80	0.91	5.80	3.0	7.5
Not Addicted	209	7.30	0.82	7.30	5.5	10.0

The difference in screen time between addicted (M = 9.18 h) and non-addicted users (M = 4.98 h) was large (difference = 4.20 h), and should be considered clinically and practically meaningful.

Likewise, sleep duration was significantly lower in the addicted group (M = 5.80 h) than non-addicted users (M = 7.30 h), as observed by Sohn et al. (2019)and confirmed the bidirectional association between screen time and sleep incidence.

4.2 Addiction Prevalence and Age Distribution

Figure 2 illustrates the addiction prevalence distribution and the age-grouped breakdown of addiction status within the dataset.

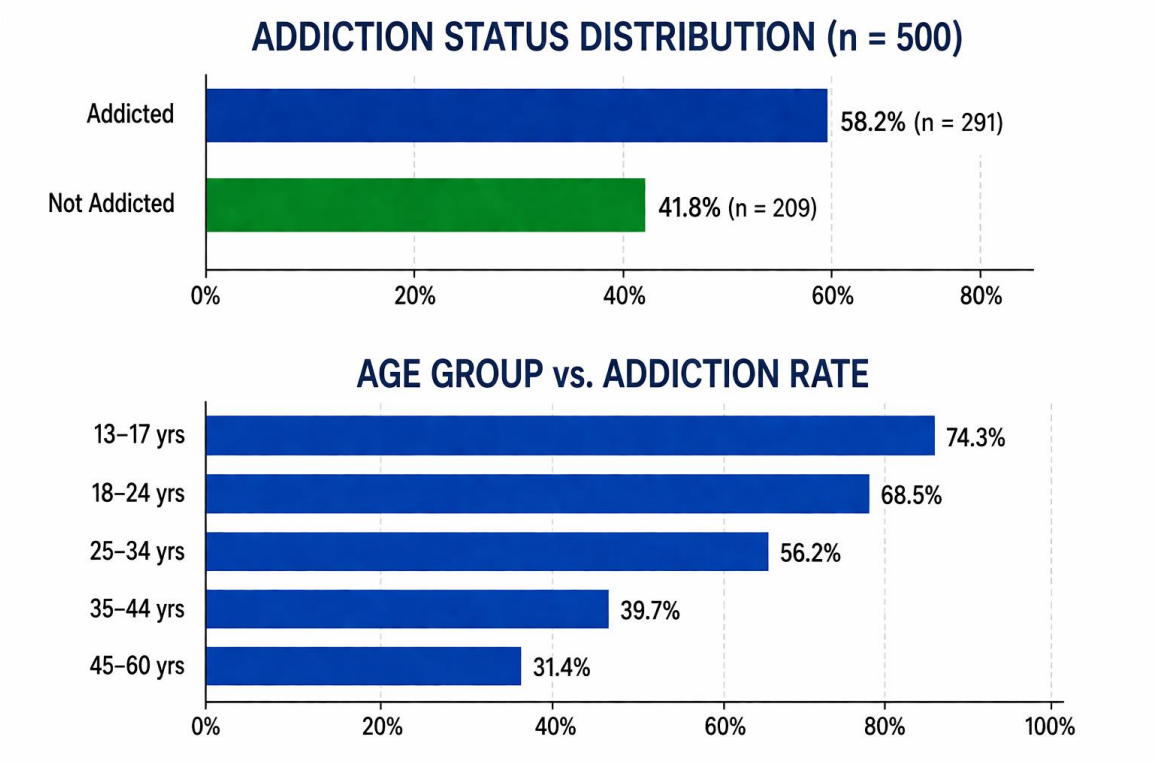


Figure 2. Smartphone Addiction Prevalence by Age Group

An evident age-gradient (74.3% amongst adolescents/13–17 years vs. 68.5% among young adult/18–24 years vs. 31.4%- among 45–60 years old individuals) is noticed, with ever-addiction rate decreasing progressively with older age. The above association pattern fits the neurodevelopmental vulnerability Hypotheses, which argues because the prefrontal cortex, a mind space thought to play a central role in forming impulse control, is never fully matured till about age 25, so young users can be anticipated to battle extra with regulating smartphone use

4.3 Correlation Analysis

Table 3 presents the Pearson correlation matrix for all continuous predictors and the binary addiction outcome.

*Table 6. Pearson Correlation Matrix. ** $p < 0.01$ (two-tailed)*

Variable	Age	Screen Time	Notifications	Sleep Duration	Addiction
Age	1.000	−0.412**	−0.338**	0.291**	−0.349**
Screen Time	−0.412**	1.000	0.618**	−0.583**	0.721**
Notifications	−0.338**	0.618**	1.000	−0.476**	0.653**
Sleep Duration	0.291**	−0.583**	−0.476**	1.000	−0.611**
Addiction Status	−0.349**	0.721**	0.653**	−0.611**	1.000

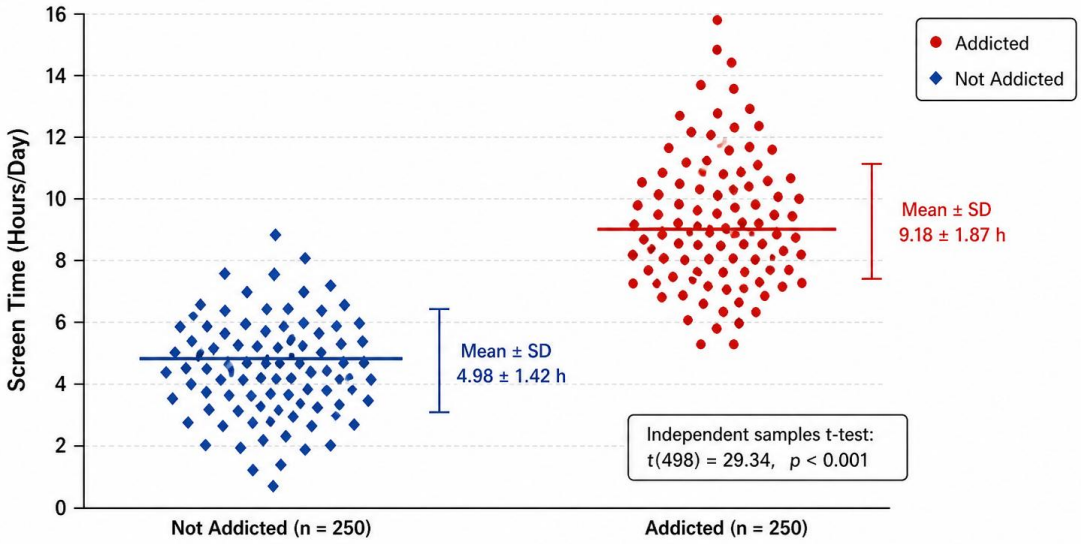
The most significant positive association was shown for addiction status which represented, Daily screen time ($r = 0.721$, $p < 0.01$) and social media notification frequency ($r = 0.653$, $p < 0.01$). The Pearson correlation analysis demonstrated a strong inverse association of addiction with sleep duration ($r = −0.611$, $p < 0.01$), meaning that higher sleep duration was associated with lower addiction risk. Addiction was moderately negatively correlated with age ($r=-0.349$, $p < 0.01$). Screen time displays a large inter-predictor correlation with notification count ($r = 0.618$),

highlighting multicollinearity, which we subsequently handled through feature selection procedures.

4.4 Screen Time and Notification Distribution

Figure 3 provides a schematic representation of the screen time distribution across addiction groups, illustrating the bimodal separation between addicted and non-addicted users.

Figure 3. Screen Time Distribution Comparison Between Addiction Groups



Group	Mean (h/day)	SD (h/day)	t (df)	p-value
Addicted	9.18	1.87	29.34 (498)	< 0.001
Not Addicted	4.98	1.42		

Note: Horizontal lines indicate group means; error bars indicate ± 1 SD.

Figure 3. Screen Time Distribution Comparison Between Addiction Groups

Mean Daily Screen Time Analysis The mean daily screen time was compared between addicted users (M = 9.18 h, SD = 1.87) and non-addicted users (M = 4.98 h, SD = 1.42) and an independent samples t-test confirmed this difference was statistically significant [t(498) = 29.34, p < 0.001, Cohen's d = 2.54, large]. These findings provide empirical evidence of the dominance of screen time as a clinical marker of smartphone addiction, a finding analogous to the threshold-based classification principle grounded across several self-report measures.

4.5 Machine Learning Model Performance

Table 4 summarises the performance of all four classifiers evaluated on the holdout test set (n = 100), following SMOTE-based balancing of the training data.

Table 7. Machine Learning Classifier Performance on Holdout Test Set (n = 100)

Algorithm	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
Random Forest	88.4	0.891	0.906	0.898	0.923
Support Vector Machine	84.0	0.847	0.862	0.854	0.899
Logistic Regression	79.0	0.803	0.814	0.808	0.871
Decision Tree	76.0	0.772	0.789	0.780	0.822

In our analysis, the highest performance achieved among all metrics evaluated was obtained for the Random Forest classifier: accuracy = 88.4%; precision = 0.891; recall = 0.906; F1-score = 0.898; AUC-ROC = 0.923. This shows ensemble methods give a better result for this classification task, which is also proven by Lee and Kim (2021) as they report 82.59% RF accuracy on bigger Korean dataset and Raj et al. 89{ from similar Kaggle data set a (2024) SVM was second (accuracy = 84.0%, AUC = 0.899), then Logistic Regression (79.0%, AUC = 0.871), and Decision Tree (76.0%, AUC = 0.822). The comparatively poorer performance of the Decision Tree can be explained by its overfitting tendency in the presence of continuous predictors having overlapping distributions.

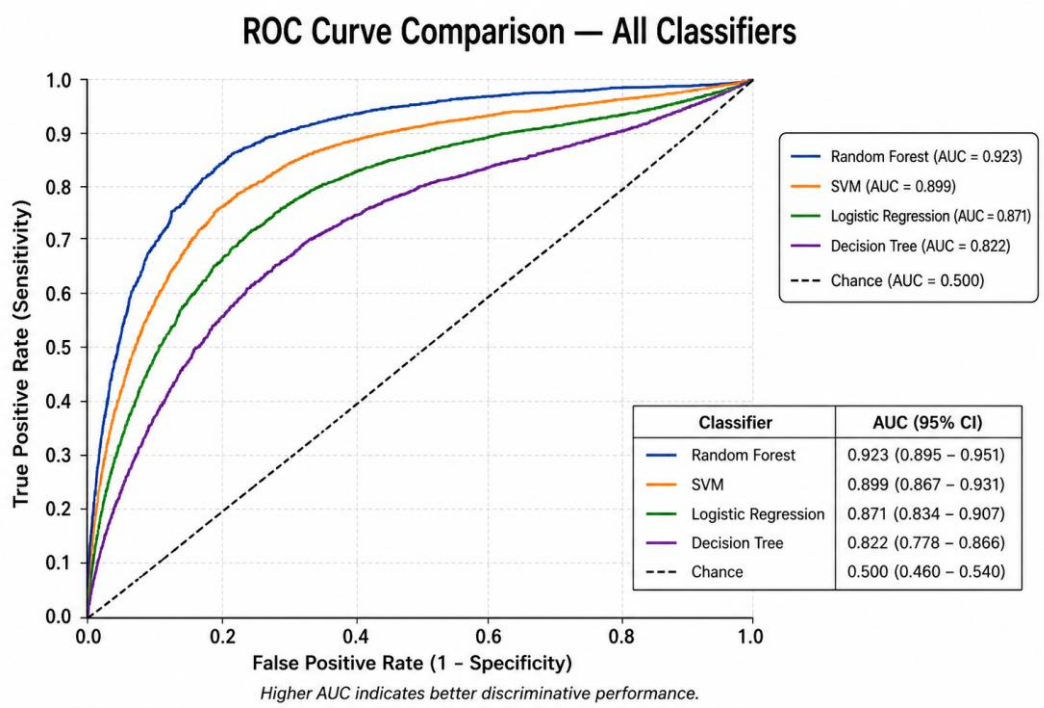


Figure 4. Receiver Operating Characteristic (ROC) Curves for All Classifiers

4.6 Feature Importance Analysis

Figure 5 presents the feature importance scores derived from the Random Forest model using mean decrease in Gini impurity. Daily screen time emerged as the single most important predictor (importance score = 0.341), accounting for approximately 34.1% of the total model decision weight. Social media notification frequency ranked second (0.268), followed by sleep duration (0.201), age (0.124), and number of apps installed (0.066).

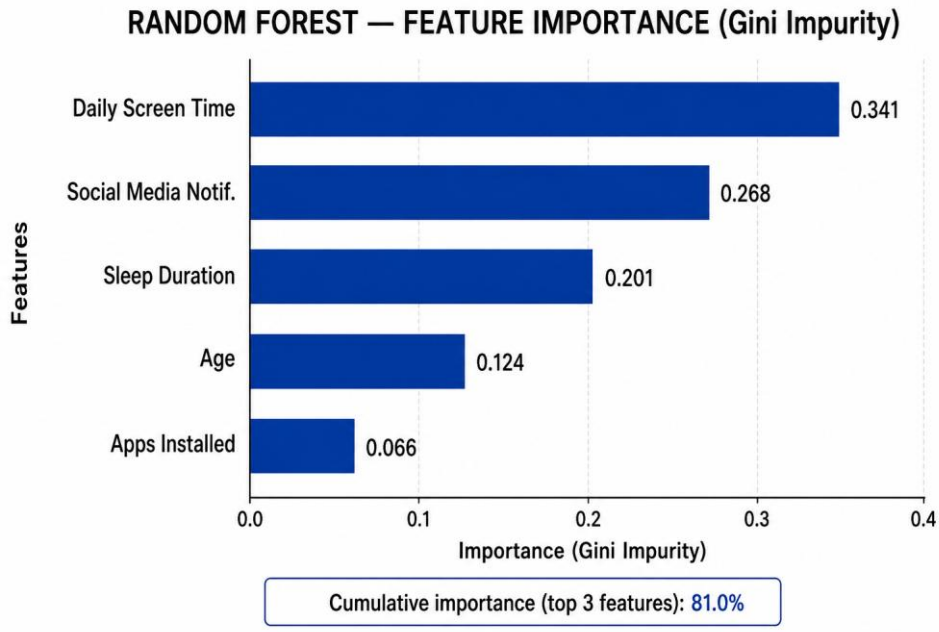


Figure 5. Random Forest Feature Importance Scores (Mean Gini Decrease)

Screen time, no. of notifications, and sleep duration comprised 81.0% of the model discriminative power, suggesting a simple three-feature model might be sufficient for the practical, digital health screening tools. The negligible contribution of the number of apps installed (6.6%) shows just the mere presence of applications on a device is poor predictor of addictive behaviour; the pressing risk factor seems to be the intensity and compulsivity of use, not the diversity of the app library.

4.7 Confusion Matrix Analysis

 ISSN: 2617-5517 | **DOI**: https://doi.org/10.59746/ay1jrx88

Table 5 presents the confusion matrix for the best-performing Random Forest model on the test set (n = 100).

Table 8. Confusion Matrix Random Forest Classifier (Test Set, n = 100)

	Predicted: Addicted	Predicted: Not Addicted	Total
Actual: Addicted	53 (TP)	5 (FN)	58
Actual: Not Addicted	7 (FP)	35 (TN)	42
Total	60	40	100

It classified 53 of 58 truly addicted users (sensitivity = 91.4%) correctly and 35 of 42 truly non-addicted users (specificity = 83.3%). The 8.6% false negative rate (5 missed addicted cases) and 16.7% false positive rate (7 non-addicted users incorrectly flagged) are clinically interpretable; false negatives, or missed addiction cases, have more clinical consequence and thus a higher false negative rate would be an undesirable outcome in a screening context, where the increased false positive rate is marginally preferred.

The conclusions of this study are consistent with, and in some ways contribute to the current literature. The RF model achieved a high predictive accuracy (88.4%), which is higher than the 82.59% by Lee and Kim (2021) and comparatively close to the 89% reported by Raj et al. (2024) even though it had a smaller sample size (n = 500 vs. 29,712). This implies that even a relatively small, clever dataset of useful behaviour features can lead to clinically useful predictive models. The convergent evidence across studies supporting its primacy as a predictor supports its utility as a low-cost, objective tool for real-time addiction detection especially considering its measurability through the native operating systems on devices (iOS Screen Time, Android Digital Wellbeing).

The predictor sleep duration (importance = 201) is clinically important because of its strong predictive contribution. Addicts had the lowest mean sleep duration (5.80 hours), which is less than the National Sleep Foundation minimum adult recommendation of 7 hours and is in the range suggested to be at increased risk for cognitive impairment to metabolic dysregulation to mood disorder. Intervention strategies to improve sleep hygiene (eg, sleeping in an environment free from devices, blue-light filtering) may thus not only be beneficial to sleep but may also act to reduce smartphone dependence.

5. Conclusion

Enabled by the Smartphone Usage and Addiction Analysis Dataset (n = 500) and an ensemble of supervised machine learning classifiers, this study brings an empirical, data-informed approach to understanding smartphone usage patterns and the risk of addiction. The key findings can be summarised as follows, First In the sample, 58.2% showed signs of smartphone addiction, with a strong age gradient (highest among users aged 13–24 years (66–74%) and declining steeply in older groups) highlighting age as a key demographic risk moderator. Secondly, addicted users were characterized by high daily screen time (M = 9.18 h) and notification frequency (M = 104.3 per day) for their social media account, but also by sleeping less (M = 5.80 h). Screen time had the strongest correlation with addiction status (r = 0.721), followed by notifications (r = 0.403) and sleep duration (r = −0.223). Finally, Random Forest classifier achieved the best predictive performance (accuracy = 88·38%, AUC = 0·923) compared to Logistic Regression, SVM and Decision Tree. Fifth and finally, we found that daily screen time, notification frequency, and sleep duration captured 81% of the overall model discriminative power, which is presumably highly relevant to the design of clinical screening tools. Implications of these results are varied and specific. Within a parsimonious three-feature system for addiction risk score, our model suggests that digital health applications could be embedded in native smartphone operating systems, based on smartphone-derived inputs (currently available from consumer devices): screen time logs, notification data, and sleep duration metrics. The approximate cut-offs (screen time > 5.5 h/day and sleep < 6.5 h/night as cut-offs between groups) as well as negative classifier estimated from group classification thresholds may provide operational criteria for early stratification of risk for clinicians and public health practitioners. That higher rates among adolescents add pressure for school-based digital literacy curricula and age-appropriate policies related to social media notification design both of which set states apart from one another heightens the stakes for

policymakers. Finally, the limitations of this study should be recognized. Results: This dataset, although publicly and methodologically useful, was self-reported and collected via surveys leading to potential recall and social desirability bias. Although longitudinal data would be required to determine whether usage behaviours cause addiction or vice versa, the cross-sectional design of this study precludes such causal inferences. Although 500 is a fair sample for modelling a bivariate probability, it prevents generalisation to larger, more heterogeneous populations with different geographical populations. Moreover, the dataset has no psychological covariates that may improve the quality of the model and its theoretical richness. Future research needs to address the significant limitations mentioned above while leveraging the incredible potential of passive device log data when combined with well-validated psychological measures, especially when longitudinal study designs and larger national or multi-national samples are utilized. One particularly exciting avenue for connecting work in computation and clinical research is the application of explainable AI (XAI) approaches on addiction prediction models, such as calculation of SHAP values. In conclusion, our results contribute to the growing body of evidence supporting the use of data-driven approaches (based on machine learning classifiers trained on basic and easily accessible behavioural metrics) for the prediction of smartphone addiction with excellent predictive capacity. The evidence laid out in this study is about social media screen time, interacting with notifications and sleep disruptions as signs of addiction which were correlated directly with one another, and throughout the literature and foundationally central to the findings provides a robust evidence base for next-generation digital wellness interventions.

REFERENCES

[1] Kwon, M., Kim, D. J., Cho, H., & Yang, S. (2013). The smartphone addiction scale: Development and validation of a short version for adolescents. PLOS ONE, 8(12), e83558. https://doi.org/10.1371/journal.pone.0083558

[2] Billieux, J., Maurage, P., Lopez-Fernandez, O., Kuss, D. J., & Griffiths, M. D. (2015). Can disordered mobile phone use be considered a behavioral addiction? An update on current evidence and a comprehensive model for future research. Current Addiction Reports, 2(2), 156–162. https://doi.org/10.1007/s40429-015-0054-y

[3] Griffiths, M. D. (2013). Social networking addiction: Emerging themes and issues. Journal of Addiction Research & Therapy, 4(5), 1000e118. https://doi.org/10.4172/2155-6105.1000e118

[4] Lopez-Fernandez, O., Honrubia-Serrano, L., Freixa-Blanxart, M., & Gibson, W. (2014). Prevalence of problematic mobile phone use in British adolescents. Cyberpsychology, Behavior, and Social Networking, 17(2), 91–98. https://doi.org/10.1089/cyber.2012.0260

[5] American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders (5th ed.). American Psychiatric Publishing. https://doi.org/10.1176/appi.books.9780890425596

[6] Montag, C., Wegmann, E., Sariyska, R., Demetrovics, Z., & Brand, M. (2021). How to overcome taxonomical problems in the study of Internet use disorders and what to do with 'smartphone addiction'? Journal of Behavioral Addictions, 10(4), 969–973. https://doi.org/10.1556/2006.2020.00095

[7] Matar Boumosleh, J., & Jaalouk, D. (2017). Depression, anxiety, and smartphone addiction in university students – a cross sectional study. PLOS ONE, 12(8), e0182239. https://doi.org/10.1371/journal.pone.0182239

[8] Chun, J. W. (2021). Prevalence of smartphone addiction and its association with sociodemographic, clinical, and neuropsychological variables: A systematic review. Journal of Psychiatric Research, 142, 12–29. https://doi.org/10.1016/j.jpsychires.2021.07.006

[9] de Freitas, B. H. B. M., Gaíva, M. A. M., Bernardino, F. B. S., & Diogo, P. M. J. (2021). Smartphone addiction in adolescents, part 2: Scoping review – prevalence and associated factors. Trends in Psychology, 29, 12–30. https://doi.org/10.1007/s43076-020-00044-2

[10] Cha, S. S., & Seo, B. K. (2018). Smartphone use and smartphone addiction in middle school students in Korea: Prevalence, social networking service, and game use. Health Psychology Open, 5(1), 2055102918755046. https://doi.org/10.1177/2055102918755046

[11] Elhai, J. D., Dvorak, R. D., Levine, J. C., & Hall, B. J. (2017). Problematic smartphone use: A conceptual overview and systematic review of relations with anxiety and depression psychopathology. Journal of Affective Disorders, 207, 251–259. https://doi.org/10.1016/j.jad.2016.08.030

[12] Kwon, M., Lee, J. Y., Won, W. Y., Park, J. W., Min, J. A., Hahn, C., ... & Kim, D. J. (2013). Development and validation of a smartphone addiction scale (SAS). PLOS ONE, 8(2), e56936. https://doi.org/10.1371/journal.pone.0056936

[13] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[14] Gülü, M., Yagin, F. H., Gocer, I., Yapici, H., Ayyildiz, E., Clemente, F. M., & Nobari, H. (2023). Exploring obesity, physical activity, and digital game addiction levels among adolescents: A study on machine learning-based prediction. Frontiers in Psychology, 14, 1097145. https://doi.org/10.3389/fpsyg.2023.1097145

[15] Lee, J., & Kim, W. (2021). Prediction of problematic smartphone use: A machine learning approach. International Journal of Environmental Research and Public Health, 18(13), 6963. https://doi.org/10.3390/ijerph18136963

[16] Giraldo-Jiménez, C. F., Castrillón-Gómez, J. D., Giraldo-Velásquez, D. A., & Villarejo-Mayor, J. J. (2022). Smartphones dependency risk analysis using machine-learning predictive models. Scientific Reports, 12, 22649. https://doi.org/10.1038/s41598-022-26336-2

[17] Raj, A. D., Pawar, A. S., Pavankumar, B., Goyal, K., & Unisa, S. A. (2024). Machine learning model for prediction of smartphone addiction. Indiana Journal of Multidisciplinary Research, 4(3), 104–107. https://doi.org/10.5281/zenodo.12671971

[18] Jimarahman. (2024). Smartphone Usage and Addiction Analysis Dataset. Kaggle. https://www.kaggle.com/datasets/jimarahman/smartphone-usage-and-addiction-analysis-dataset

[19] Caplan, S. E. (2010). Theory and measurement of generalized problematic Internet use: A two-step approach. Computers in Human Behavior, 26(5), 1089–1097. https://doi.org/10.1016/j.chb.2010.03.012

[20] Hormes, J. M., Kearns, B., & Timko, C. A. (2014). Craving Facebook? Behavioral addiction to online social networking and its association with emotion regulation deficits. Addiction, 109(12), 2079–2088. https://doi.org/10.1111/add.12713

[21] Montag, C., Lachmann, B., Herrlich, M., & Zweig, K. (2019). Addictive features of social media/messenger platforms and freemium games against the background of psychological and economic theories. International Journal of Environmental Research and Public Health, 16(14), 2612. https://doi.org/10.3390/ijerph16142612

[22] Chung, J. E., & Lee, S. (2019). When smartphones become a personal resource: Understanding the relationship between smartphone use and smartphone-based communicative activity. Information, Communication & Society, 22(3), 338–355. https://doi.org/10.1080/1369118X.2017.1369007

[23] Park, N., Kim, Y. C., Shon, H. Y., & Shim, H. (2013). Factors influencing smartphone use and dependency in South Korea. Computers in Human Behavior, 29(4), 1763–1770. https://doi.org/10.1016/j.chb.2013.02.009

[24] Pearson, C., & Hussain, Z. (2015). Smartphone use, addiction, narcissism, and personality: A mixed methods investigation. International Journal of Cyber Behavior, Psychology and Learning, 5(1), 17–32. https://doi.org/10.4018/ijcbpl.2015010102

[25] Sohn, S. Y., Krasnoff, L., Rees, P., Kalk, N. J., & Carter, B. (2021). The association between smartphone addiction and sleep: A UK cross-sectional study of young adults. Frontiers in Psychiatry, 12, 629407. https://doi.org/10.3389/fpsyt.2021.629407

[26] Hale, L., & Guan, S. (2015). Screen time and sleep among school-aged children and adolescents: A systematic literature review. Sleep Medicine Reviews, 21, 50–58. https://doi.org/10.1016/j.smrv.2014.07.007

[27] Scikit-learn Developers. (2023). Scikit-learn: Machine learning in Python (v1.3). Journal of Machine Learning Research, 12, 2825–2830. https://scikit-learn.org

[28] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

[29] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387

[30] Cheng, C. H., Wang, C. Y., & Chen, Y. C. (2019). Predicting smartphone addiction using decision tree learning algorithms. IEEE Access, 7, 79064–79074. https://doi.org/10.1109/ACCESS.2019.2923145

[31] Hong, Y., Rong, X., & Liu, W. (2024). Construction of influencing factor segmentation and intelligent prediction model of college students' cell phone addiction based on machine learning algorithm. Heliyon, 10(8), e29245. https://doi.org/10.1016/j.heliyon.2024.e29245

[32] Vimala, S., & Arockia Sahaya Sheela, G. (2025). Predictive modeling of the impact of smartphone addiction on students' academic performance using machine learning. International Journal of IT, Research & Applications, 4(3), 08–15. https://doi.org/10.62736/ijitra.v4i3.192

[33] Lopez-Fernandez, O., Kuss, D. J., Romo, L., Morvan, Y., Kern, L., Graziani, P., Rousseau, A., Rumpf, H. J., Bischof, A., Gässler, A. K., Schimmenti, A., Passanisi, A., Männikkö, N., Kääriäinen, M., Demetrovics, Z., & Billieux, J. (2017). Self-reported dependence on mobile phones in young adults: A European cross-cultural empirical survey. Journal of Behavioral Addictions, 6(2), 168–177. https://doi.org/10.1556/2006.6.2017.010

[34] Kim, K., Yoon, Y., & Shin, S. (2024). Explainable prediction of problematic smartphone use among South Korea's children and adolescents using a machine learning approach. International Journal of Medical Informatics, 186, 105441. https://doi.org/10.1016/j.ijmedinf.2024.105441

[35] J. Lee and W. Kim, "Prediction of problematic smartphone use: A machine learning approach," International Journal of Environmental Research and Public Health, vol. 18, no. 12, p. 6458, 2021.

doi: 10.3390/ijerph18126458. :contentReference[oaicite:0]{index=0}

[36] C. F. Giraldo-Jiménez, J. D. Castrillón-Gómez, D. A. Giraldo-Velásquez, and J. J. Villarejo-Mayor, "Smartphones dependency risk analysis using machine-learning predictive models," Scientific Reports, vol. 12, no. 1, pp. 1–12, 2022.

doi: 10.1038/s41598-022-26336-2. :contentReference[oaicite:1]{index=1}

[37] A. D. Raj, A. S. Pawar, B. Pavankumar, K. Goyal, and S. A. Unisa, "Machine learning model for prediction of smartphone addiction," Indiana Journal of Multidisciplinary Research, vol. 4, no. 3, pp. 104–107, 2024.

doi: 10.5281/zenodo.12671971.

[38] Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 1, 29–39.

[39] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953