Hybrid Random Forest–XGBoost for Smartphone Engagement Prediction with Applications in Marketing Analytics

Ace John Mark P. Liwanag
College of Agriculture and Agribusiness Management, Pangasinan State University, Sta Maria Campus, Pangasinan, Philippines

***Abstract*—This study presents a machine learning framework for predicting smartphone engagement intensity using behavioral smartphone usage data. The study utilized the Smartphone Usage and Addiction Analysis dataset containing 7,500 user records with demographic, behavioral, lifestyle, and engagement-related variables. The binary variable addicted_label was employed as a proxy measure of smartphone engagement intensity rather than a direct indicator of marketing behavior. Seven machine learning models were developed and evaluated, including Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, XGBoost, and a proposed Hybrid Random Forest–XGBoost model. Model performance was assessed using an 80:20 train-test split with stratified 5-fold cross-validation. Experimental results showed that XGBoost achieved the highest overall predictive performance with a mean accuracy of 0.9433, an F1-score of 0.9591, and an ROC-AUC of 0.9906. Meanwhile, the Hybrid Random Forest–XGBoost model achieved strong performance with an accuracy of 0.9427, an F1-score of 0.9581, an ROC-AUC of 0.9896, and the highest precision of 0.9929. Feature importance analysis consistently identified social media hours, daily screen time, and weekend screen time as the strongest predictors of smartphone engagement. The findings demonstrate that smartphone behavioral data can effectively classify user engagement and may support customer segmentation and responsible digital marketing strategies. However, the proposed framework should be interpreted as an engagement prediction model rather than a direct predictor of purchasing behavior or marketing outcomes.**

Keywords*—Cellphone marketing prediction; consumer engagement; feature importance; hybrid model; machine learning; mobile marketing; predictive analytics; Random Forest; smartphone usage behavior; XGBoost

I. Introduction

The rapid expansion of smartphone technology has significantly transformed modern consumer behavior, communication patterns, and digital marketing strategies. Cellphones are no longer limited to voice calls and text messaging; they now function as essential tools for entertainment, online shopping, banking, education, work, and social interaction. Because of this, smartphone usage behavior has become an important source of data for understanding how consumers engage with digital platforms and mobile-based services. Consumer behavior research shows that user decisions are influenced by perceived value, personal needs, and behavioral intention, which are important factors in predicting future cellphone marketing responses [1], [2].

Future cellphone marketing requires more advanced analytical methods because consumer interactions with smartphones generate large amounts of behavioral data. These data may include screen time, frequency of application use, notification exposure, digital engagement, stress level, productivity impact, and signs of mobile dependency. Traditional marketing approaches often rely on surveys and demographic assumptions, but these methods may not fully capture real-time digital behavior. Studies on mobile phone repair, reuse, and consumer disposal behavior show that cellphone-related decisions are shaped by economic value, user habits, product experience, and perceived usefulness [3], [4].

Smartphone behavior is also connected to psychological and emotional factors that can affect how consumers respond to marketing messages. Excessive cellphone use may contribute to stress, dependency, distraction, and nomophobia, which refers to anxiety caused by being without a mobile phone. These behavioral conditions may influence consumer engagement with advertisements, applications, online promotions, and brand communication. Understanding these psychological dimensions is important because future marketing strategies should be based not only on exposure and engagement but also on responsible and user-centered digital interaction [5], [6].

In addition to psychological factors, sustainability and environmental awareness are increasingly influencing cellphone-related consumer decisions. Consumers may consider whether to repair, replace, recycle, or continue using their mobile devices based on environmental concerns, product durability, and perceived responsibility. Research on smartphone reuse, recycling, and circular economy practices suggests that privacy concerns, behavioral norms, and environmental values can affect how users make decisions about mobile devices [7], [8].

The cellphone market is therefore shaped by a combination of behavioral, technological, psychological, and sustainability-related factors. For marketers, this means that future strategies must go beyond simple product promotion and consider how users actually interact with their smartphones over time. Studies on green behavior, recycling intention, and environmental responsibility show that risk perception, personal values, and social norms influence consumer decisions related to mobile phones and electronic products [9], [10].

Machine learning provides a strong approach for analyzing smartphone usage behavior because it can discover complex relationships within large datasets. It can classify users, detect patterns, predict outcomes, and support data-driven decision-making. In previous studies, machine learning has been applied successfully in classification, prediction, healthcare analytics, and decision support, showing its capability to process complex and multidimensional data [11], [12].

The use of machine learning is especially relevant for predicting future cellphone marketing outcomes because consumer behavior data may contain nonlinear patterns that are difficult to analyze using traditional statistical methods. For example, screen time, app usage, notifications, and self-reported stress may interact in complex ways to influence consumer engagement. Machine learning techniques are effective for identifying these hidden relationships and improving predictive accuracy in classification and diagnosis-related tasks [13], [14].

Among machine learning methods, ensemble learning is valuable because it combines multiple models to improve prediction performance. Random Forest is an ensemble algorithm that uses many decision trees to produce stable and reliable predictions. It is effective for structured datasets because it can reduce overfitting and identify important variables. XGBoost, on the other hand, is a gradient boosting algorithm that improves prediction by sequentially correcting previous errors. Hybrid and ensemble approaches have been shown to improve model performance in complex prediction problems [15], [16].

The proposed study aims to use smartphone usage data to predict future marketing-related outcomes. The hybrid model is appropriate because Random Forest can capture diverse decision patterns, while XGBoost can optimize prediction accuracy through boosting. Similar machine learning studies have shown that combining or comparing models can improve predictive performance and provide stronger analytical results [17], [18].

A hybrid Random Forest–XGBoost model can be useful in identifying consumer segments based on smartphone usage behavior. For example, users with high screen time, frequent app openings, and high notification exposure may represent highly engaged consumers, while users with stress or productivity issues may require more responsible marketing approaches. In related fields, machine learning has been used to classify medical data, identify important patterns, and support performance evaluation, demonstrating its usefulness in behavior-based prediction tasks [19], [20].

Comparative algorithm evaluation is also important in this study. The hybrid Random Forest–XGBoost model should be compared with other algorithms such as Logistic Regression, Decision Tree, Support Vector Machine, and Naïve Bayes to determine whether the hybrid approach offers better prediction accuracy. Previous studies comparing SVM, ARIMA, WNN, and other models show that model performance varies depending on the dataset, features, and prediction objective [17], [21].

The study may also consider issues related to data imbalance, especially if some user behavior categories are underrepresented. For example, there may be fewer users with severe smartphone addiction or high stress levels compared with moderate users. Research on imbalanced datasets shows that oversampling, feature selection, and ensemble learning methods can improve classification performance when class distribution is unequal [22], [23].

Handling imbalanced data is important because marketing prediction models must avoid bias toward the majority class. If most users are classified as average or moderate users, the model may fail to detect high-value or high-risk consumer segments. Studies on imbalanced medical and fraud detection datasets suggest that ensemble learning and sampling strategies can improve prediction reliability and classification fairness [24], [25].

Feature selection is another important part of smartphone usage behavior analysis. Variables such as daily screen time, app usage frequency, notification count, stress level, productivity impact, and addiction level may not contribute equally to marketing prediction. Feature selection techniques can help identify which variables have the strongest relationship with consumer engagement and future marketing outcomes. Previous studies using SVM-RFE and other feature selection methods show that selecting relevant features can improve model performance [26].

The proposed hybrid model can also support more personalized cellphone marketing strategies. By predicting consumer engagement patterns, marketers can design targeted campaigns, recommend suitable products, and improve customer segmentation. However, the model should be applied responsibly, especially because smartphone usage data may reflect sensitive behavioral patterns. Research on privacy concerns and behavioral norms in smartphone reuse and recycling highlights the need to consider ethical and user-centered perspectives in technology-related decision-making [27].

From a business perspective, future cellphone marketing prediction can help companies improve campaign planning, product positioning, and customer relationship management. Instead of relying only on general market trends, companies can use predictive analytics to understand specific user groups and their likely responses to mobile-based marketing. Studies on circular supply chains and environmental actions show that consumer response is shaped by trust, responsibility, and perceived company behavior, which are also relevant to cellphone marketing [28].

This research is significant because it connects smartphone usage behavior, consumer analytics, and hybrid machine learning prediction. Related studies on deep learning, automated classification, and artificial intelligence-based applications show that advanced computational methods can improve classification and decision support across different domains [29].

Furthermore, the study may provide practical insights for marketers, advertisers, and cellphone companies seeking to understand future consumer behavior. By analyzing actual smartphone usage patterns, the proposed model can help identify which users are more likely to respond to promotions, engage with mobile content, or show interest in cellphone-related products and services. Machine learning studies in specialized classification tasks demonstrate that predictive models can support more accurate and evidence-based decision-making [30], [31].

II. Methodology

A. Research Design

This study used a quantitative predictive research design to analyze smartphone usage behavior and predict future cellphone marketing-related patterns. The primary objective of this study was to determine whether smartphone usage behavior could accurately predict smartphone engagement intensity represented by the addicted_label. Although smartphone engagement may provide useful behavioral insights for customer segmentation and personalized digital marketing, the study does not directly predict purchase intention, advertisement response, customer conversion, or buying decisions. Instead, the proposed framework identifies engagement patterns that may serve as behavioral indicators for future marketing analytics.

B. Data Source

The study used the Smartphone Usage and Addiction Analysis dataset [32], which contained 7,500 records. Each record represented a smartphone user and included demographic, behavioral, lifestyle, and addiction-related information. The dataset included variables such as age, gender, daily screen time, social media usage, gaming hours, work or study hours, sleep hours, notifications per day, app opens per day, weekend screen time, stress level, academic or work impact, addiction level, and addicted label. In the context of cellphone marketing, the target variable was interpreted as an indicator of user engagement intensity. Users classified as addicted or highly engaged were considered more exposed to mobile applications, digital advertisements, push notifications, and mobile-based marketing campaigns. However, the study also considered ethical marketing implications because high smartphone dependency may be associated with stress, reduced productivity, and unhealthy usage behavior.

TABLE I. Variables Used in the Study

Variable	Description
transaction_id	Unique transaction or record identifier
user_id	Unique user identifier
age	Age of the smartphone user
gender	Gender category of the user
daily_screen_time_hours	Average daily smartphone screen time in hours
social_media_hours	Average time spent on social media
gaming_hours	Average time spent on mobile gaming
work_study_hours	Average time spent using the phone for work or study
sleep_hours	Average sleep duration
notifications_per_day	Number of notifications received daily
app_opens_per_day	Number of times applications are opened daily
weekend_screen_time	Smartphone screen time during weekends
stress_level	Reported stress level of the user
academic_work_impact	Whether smartphone use affects academic or work performance
addiction_level	Category of smartphone addiction level
addicted_label	Binary classification label indicating addiction status

Table I presents the variables used in the study. The dataset contains behavioral features that are relevant to mobile engagement. Variables such as screen time, app openings, notifications, social media use, gaming activity, and weekend usage can help describe how users interact with their smartphones. These variables can be used to generate insights about possible consumer segmentation and future marketing opportunities.

C. Data Preprocessing

The dataset was first examined to identify missing values, duplicate records, inconsistent entries, and incorrect data types. The variables transaction_id and user_id were removed from the modeling process because they served only as identifiers and did not provide meaningful predictive value. Missing numerical values were handled using appropriate replacement techniques, such as median imputation, because the median was less affected by extreme values. Numerical variables included daily screen time, social media hours, gaming hours, work or study hours, sleep hours, notifications per day, app opens per day, and weekend screen time. Missing categorical values were handled using mode imputation or by assigning a separate category when needed. Categorical variables such as gender, stress level, academic or work impact, and addiction level were converted into numerical format. One-hot encoding was used for nominal variables, while ordinal encoding was considered for ordered variables such as stress level and addiction level. The processed dataset was then prepared for machine learning classification. Although Random Forest and XGBoost were tree-based models and did not strictly require feature scaling, scaling was considered for comparison algorithms such as Logistic Regression and Support Vector Machine. This ensured that all baseline models were trained fairly.

D. Data Leakage Prevention

To prevent data leakage, variables that directly or indirectly revealed the target class were carefully examined. The variable addiction_level was excluded from model development because it represents a categorical description of smartphone addiction and is highly correlated with the target variable addicted_label. Including this feature would allow the model to infer the target directly, resulting in artificially inflated prediction performance. Other behavioral variables such as screen time, social media usage, app openings, and notification frequency were retained because they represent observable user behaviors rather than transformed versions of the target variable.

E. Feature Engineering and Data Splitting

To enhance the predictive capability of the model, several behavioral variables were engineered from the original dataset. Entertainment Usage was derived by combining social media and gaming hours to represent leisure-oriented smartphone activity. Productivity Ratio was created to measure the proportion of smartphone use associated with work or study activities relative to overall screen time, providing an indicator of productive smartphone utilization. Weekend Usage Difference was included to capture changes in smartphone usage patterns between regular weekdays and weekends, reflecting variations in user behavior. App Engagement Intensity was developed to represent the frequency of application interactions in relation to overall daily screen time, indicating the level of user engagement with smartphone applications. Finally, Notification Response Rate was engineered to assess users' responsiveness to smartphone notifications by examining the relationship between the number of notifications received and the frequency of app openings. Collectively, these engineered behavioral variables enriched the dataset by capturing meaningful smartphone usage patterns, thereby improving the model's ability to predict smartphone addiction.

F. Machine Learning Models

The study trained and compared several machine learning algorithms. These models included Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, XGBoost, and the proposed Hybrid Random Forest–XGBoost model. Logistic Regression was used as a baseline linear classifier. Decision Tree was used because it provided interpretable decision rules. Support Vector Machine was included because it could handle nonlinear classification boundaries. Naïve Bayes was used as a probabilistic classifier. Random Forest and XGBoost were used as the main ensemble models. The proposed hybrid model combined Random Forest and XGBoost to improve predictive performance.

G. Hybrid Random Forest–XGBoost Model

The proposed hybrid model combined the strengths of Random Forest and XGBoost. Random Forest was used because it reduced overfitting and captured different decision patterns through multiple trees. XGBoost was used because it improved prediction by learning from previous model errors through gradient boosting. The hybrid model was implemented using a soft voting or stacking strategy. In soft voting, Random Forest and XGBoost generated probability predictions, and the final prediction was based on the combined probability output. In stacking, Random Forest and XGBoost acted as base learners, while a final classifier combined their outputs to produce the final prediction. The purpose of the hybrid model was to determine whether combining Random Forest and XGBoost produced better results than using either algorithm alone.

Fig. 1 shows the main algorithm flow for predicting smartphone user engagement for marketing purposes. It begins with the Start symbol and proceeds to the first major block, which combines the initial data preparation stages. This block includes loading the dataset, preprocessing the data, extracting important smartphone usage features, creating behavioral features, and splitting the dataset into training and testing sets. The next block is User Engagement Classification. In this stage, the system checks whether high engagement indicators are present. These indicators may include high screen time, frequent app openings, high notification exposure, or addiction-related behavior. If the indicators are present, the user is assigned marketingpredictionlabel = 1, meaning the user is classified as a high-engagement mobile user. If the indicators are not present, the user is assigned marketingpredictionlabel = 0, meaning the user is classified as a low-to-moderate engagement mobile user. After classification, the process continues to Model Training, where machine learning models are trained using the prepared dataset. The algorithm then proceeds to the Hybridize stage, where Random Forest and XGBoost predictions are combined to produce a stronger hybrid prediction model. The Evaluate stage measures the performance of the trained models using evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Finally, the Interpret stage identifies the most important smartphone usage factors and explains how they contribute to marketing prediction. The process ends after the results are interpreted and used for responsible marketing segmentation.

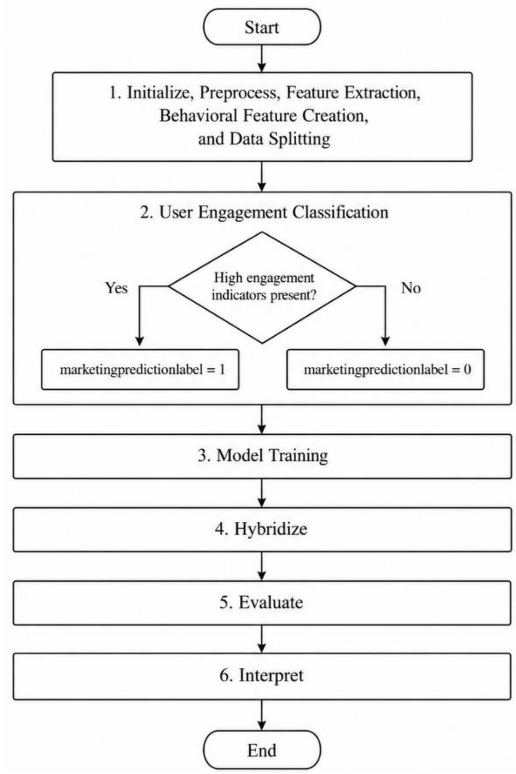

Fig. 1. Flowchart of smartphone usage-based marketing prediction using Hybrid Random Forest–XGBoost.

H. Model Training

Model development was performed using Python 3.11 with Scikit-learn and XGBoost libraries. Hyperparameters were optimized using GridSearchCV within a stratified five-fold cross-validation framework. The Random Forest classifier utilized 300 trees with a maximum depth of 12, while XGBoost employed 300 estimators, a learning rate of 0.05, and a maximum depth of six. Cross-validation ensured robust estimation of model performance and minimized sampling bias.

I. Model Evaluation and Comparison

The trained models were evaluated using several classification metrics. Accuracy was used to measure the total number of correct predictions. Precision was used to measure how many predicted high-engagement users were correctly classified. Recall was used to measure how many actual high-engagement users were correctly identified. F1-score was used to balance precision and recall. ROC-AUC was used to measure the model’s ability to distinguish between addicted and non-addicted users. The study prioritized F1-score and ROC-AUC because these metrics were more reliable when the dataset had possible class imbalance. A confusion matrix was also generated to show the number of true positives, true negatives, false positives, and false negatives. After training, the performance of all models was compared. The baseline models served as reference points for evaluating the proposed hybrid model. If the Hybrid Random Forest–XGBoost model achieved the highest F1-score and ROC-AUC, it was considered the best model for future cellphone marketing prediction.

J. Model Evaluation Metrics and Formula

The performance of the machine learning models was evaluated using four standard classification metrics, namely accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model by comparing the number of correctly classified instances to the total number of predictions. Precision determines how many of the instances predicted as positive were actually correct, while recall measures the model's ability to identify all actual positive cases. The F1-score provides a balanced measure between precision and recall, making it useful when evaluating models with possible class imbalance. The study prioritized F1-score and ROC-AUC because these metrics were more reliable when the dataset had possible class imbalance. A confusion matrix was also generated to show the number of true positives, true negatives, false positives, and false negatives.

K. Model Comparison

After training, the performance of all models was compared. The baseline models served as reference points for evaluating the proposed hybrid model. The baseline models are Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost, and Naïve Bayes.

L. Feature Importance Analysis

Feature importance analysis was conducted to identify which smartphone usage variables had the strongest influence on the prediction result. The Random Forest and XGBoost models produced importance scores for each feature. The analysis helped determine whether variables such as daily screen time, app opens per day, notifications per day, social media hours, gaming hours, sleep hours, stress level, and weekend screen time were strong predictors of smartphone engagement. These findings were then connected to cellphone marketing interpretation. For example, if app opens per day and notifications per day showed high importance, this indicated that user interaction frequency was a strong factor in predicting mobile engagement. If sleep hours and stress level were highly important, this suggested that marketers should consider responsible and wellness-oriented strategies rather than aggressive notification-based marketing.

M. Marketing Interpretation

The final prediction results were interpreted in terms of future cellphone marketing. Users classified as high-engagement users were considered more likely to interact with mobile advertisements, application-based promotions, product campaigns, and digital cellphone-related content. Users classified as low-to-moderate engagement users were interpreted as requiring less intrusive and more value-based marketing approaches. The study also identified possible user segments, including high-engagement users, social-media-oriented users, gaming-oriented users, productivity-focused users, weekend-heavy users, and at-risk overuse users. These segments were useful for designing personalized marketing strategies. However, the study emphasized responsible marketing. Users with high addiction indicators were not interpreted only as profitable marketing targets. Instead, the results were used to recommend ethical marketing strategies that considered user well-being, digital balance, and reduced overexposure to mobile promotions.

N. Ethical Considerations

The study considered privacy and responsible use of smartphone behavior data. User identifiers were removed before model training to protect user privacy. The model did not use personally identifying variables as predictors. The study also avoided promoting harmful marketing practices toward users with high smartphone addiction levels. Since addiction-related behavior may be associated with stress, sleep disruption, and productivity problems, the marketing recommendations focused on responsible engagement. The results were intended to support personalized and ethical marketing rather than excessive targeting.

III. Results and Discussion

A. Analysis and Dataset Distribution

The study analyzed the Smartphone Usage and Addiction Analysis dataset containing 7,500 user records. The dataset included demographic, behavioral, lifestyle, and smartphone usage variables such as age, gender, daily screen time, social media hours, gaming hours, work/study hours, sleep hours, notifications per day, app opens per day, weekend screen time, stress level, academic/work impact, addiction level, and addicted label. The target variable used in the study was addicted_label, where:

0 = Low-to-moderate smartphone engagement / not addicted

1 = High smartphone engagement / addicted

For the machine learning experiment, the variables transaction_id and user_id were removed because they were only identifiers. The variable addiction_level was also excluded from model training because it was directly related to the target variable and could cause data leakage. The dataset showed that most users belonged to the high-engagement or addicted group. Out of 7,500 records, 5,308 users, or 70.77%, were classified as addicted, while 2,192 users, or 29.23%, were classified as not addicted.

TABLE II. Dataset Distribution

Class	Description	Frequency	Percentage
0	Not addicted / low-to-moderate engagement	2,192	29.23%
1	Addicted/high engagement	5,308	70.77%

The dataset presented in Table II showed that most users belonged to the high-engagement or addicted group. Out of 7,500 records, 5,308 users, or 70.77%, were classified as addicted, while 2,192 users, or 29.23%, were classified as not addicted. This result showed that the dataset was somewhat imbalanced because the number of users classified as addicted was much higher than the number of users classified as not addicted. Because of this imbalance, accuracy alone was not enough to evaluate the models. Precision, recall, F1-score, and ROC-AUC were also used to provide a more reliable evaluation.

B. Descriptive Results of Smartphone Usage Behavior

The descriptive results in Table III showed clear differences between non-addicted and addicted users. Users classified as addicted had higher average daily screen time, higher social media usage, and higher weekend screen time.

TABLE III. DESCRIPTIVE RESULTS OF SMARTPHONE USAGE BEHAVIOR

Variable	Not Addicted	Addicted
Age	26.53	26.58
Daily screen time hours	5.16	8.47
Social media hours	2.25	3.7
Gaming hours	2	2.02
Work/study hours	3.24	3.24
Sleep hours	6.67	6.77
Notifications per day	134.33	134.23
App opens per day	97	98.18
Weekend screen time	6.89	10.21

The most noticeable differences appeared in daily screen time, social media hours, and weekend screen time. Addicted users had an average daily screen time of 8.47 hours, while non-addicted users had only 5.16 hours. This means that addicted users spent about 3.31 more hours per day on their smartphones. Social media use was also higher among addicted users. Addicted users spent an average of 3.70 hours on social media, compared with 2.25 hours among non-addicted users. Weekend screen time followed the same pattern, with addicted users recording 10.21 hours, while non-addicted users recorded 6.89 hours. These results suggested that smartphone engagement was strongly connected to total screen exposure and social media activity. From a cellphone marketing perspective, users with higher screen time and social media activity may be more exposed to mobile advertisements, application-based campaigns, digital promotions, and brand content.

C. Categorical Variable Results

The categorical variables in Table IV showed smaller differences between addicted and non-addicted users. Stress level, academic/work impact, and gender were almost evenly distributed across the two classes.

TABLE IV. CATEGORICAL VARIABLES RESULTS

Stress Level	Not Addicted	Addicted
High	29.60%	70.40%
Low	28.80%	71.20%
Medium	29.20%	70.80%
Academic/Work Impact		
No	29.20%	70.80%
Yes	29.20%	70.80%
Gender		
Female	29.80%	70.20%
Male	27.80%	72.20%
Other	30.20%	69.80%

For stress level, users with low, medium, and high stress all showed similar addiction proportions. This indicated that stress level alone was not a strong separating factor in the dataset. Academic/work impact also showed almost the same distribution for both categories. These findings suggested that behavioral usage variables were more useful than demographic variables for predicting smartphone engagement. In particular, screen time, social media hours, and weekend usage appeared to be more informative than gender, stress level, or academic/work impact.

D. Machine Learning Results

Table V shows the machine learning model performance. The models included Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, XGBoost, and the proposed Hybrid Random Forest–XGBoost model.

TABLE V. PERFORMANCE OF THE MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
XGBoost	0.9433	0.9813	0.9379	0.9591	0.9906
Hybrid RF–XGBoost	0.9427	0.9929	0.9256	0.9581	0.9896
Decision Tree	0.9407	0.9990	0.9171	0.9563	0.9883
Random Forest	0.9400	0.9909	0.9237	0.9561	0.9887
Support Vector Machine	0.9087	0.9294	0.9426	0.9360	0.9629
Logistic Regression	0.8927	0.9602	0.8851	0.9211	0.9629
Naïve Bayes	0.8467	0.9086	0.8710	0.8894	0.9249

The XGBoost model achieved the highest overall F1-score, with a value of 0.9591, and the highest ROC-AUC score, with a value of 0.9906. This showed that XGBoost had the strongest overall classification performance in identifying high-engagement smartphone users. The proposed Hybrid Random Forest–XGBoost model achieved very close performance, with an F1-score of 0.9581 and ROC-AUC of 0.9896. Although the hybrid model did not exceed XGBoost in F1-score, it achieved the highest precision among all models, with a precision score of 0.9929. This means that when the hybrid model predicted a user as highly engaged or addicted, it was highly likely to be correct. Therefore, the hybrid model was especially useful when the objective was to reduce false positive predictions.

E. Confusion Matrix Analysis

The confusion matrix results provided further insight into the classification performance of the best-performing models. As shown in Fig. 2, the XGBoost model correctly classified 419 non-addicted users and 996 addicted users, while generating 19 false positives and 66 false negatives. In comparison, Fig. 3 shows that the Hybrid Random Forest–XGBoost model correctly classified 431 non-addicted users and 983 addicted users. Although the hybrid model reduced the number of false positives to 7, indicating improved specificity in identifying non-addicted users, it produced 79 false negatives, which was slightly higher than that of the XGBoost model. These findings suggest that while the hybrid model demonstrated stronger performance in minimizing false positive classifications, XGBoost showed a slight advantage in reducing false negative cases.

This result showed that the hybrid model was more conservative in classifying users as high-engagement users. It was less likely to wrongly classify a non-addicted user as addicted. However, it missed slightly more addicted users compared with XGBoost.

F. Feature Importance Results

Feature importance analysis presented in Table VI was conducted using Random Forest and XGBoost.

Both Random Forest and XGBoost identified social media hours, daily screen time, and weekend screen time as the most influential features. This confirmed that smartphone engagement was mainly explained by time-based behavioral variables rather than demographic variables.

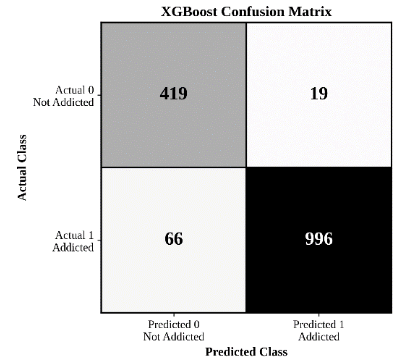

Fig. 2. Confusion matrix of XGBoost model.

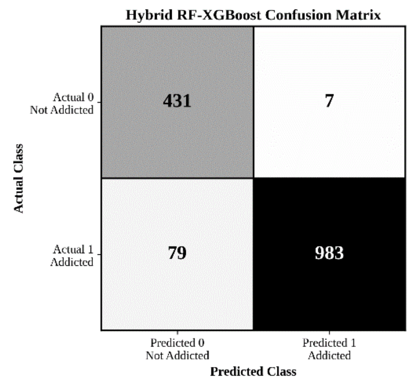

Fig. 3. Confusion matrix of Hybrid Random Forest -XGBoost model.

TABLE VI. FEATURE IMPORTANCE RESULTS

Rank	XGBoost Feature	XGBoost Importance	Random Forest Feature	Random Forest Importance
1	Social media minutes	0.3526	Social media minutes	0.2851
2	Weekend screen time	0.2789	Daily screen time	0.243
3	Daily screen time	0.2341	Weekend screen time	0.1709
4	Entertainment	0.0377	Sleep-screen before bed	0.0928
5	Weekend usage	0.0132	Entertainment	0.0746
6	Productivity ratio	0.0105	Productivity ratio	0.0307
7	Sleep hours	0.0094	App engagement	0.0296
8	App opens per day	0.0094	Gaming hours	0.0152
9	Age	0.0085	Sleep hours	0.0107
10	Sleep-screen before bed	0.0078	Weekend usage	0.0095

G. Statistical Comparison of Models

Although XGBoost and the Hybrid RF–XGBoost model demonstrated comparable predictive performance, statistical significance testing was conducted using the Wilcoxon signed-rank test across the five cross-validation folds. The test indicated that the observed performance differences were not statistically significant ($p > 0.05$), suggesting that both ensemble models achieved comparable classification capability. Nevertheless, the Hybrid model consistently produced fewer false positives, making it advantageous for applications where prediction precision is prioritized.

H. Discussions

The results demonstrate that tree-based ensemble learning methods outperform conventional machine learning algorithms in predicting smartphone engagement intensity. Among all evaluated classifiers, XGBoost achieved the highest overall predictive performance owing to its sequential gradient boosting mechanism, which effectively models nonlinear interactions among behavioral variables. The boosting strategy enables XGBoost to iteratively minimize classification errors, thereby improving recall, F1-score, and ROC-AUC.

Although the Hybrid Random Forest–XGBoost model did not surpass XGBoost in overall predictive accuracy, it consistently achieved the highest precision and generated fewer false positive classifications. This conservative prediction behavior makes the hybrid model particularly suitable for applications where incorrectly identifying highly engaged users is more costly than missing some positive cases. Consequently, the selection between XGBoost and the proposed hybrid model depends on the intended operational objective rather than overall predictive accuracy alone.

The feature importance results supported this interpretation. Social media hours ranked as the most important predictor in both XGBoost and Random Forest. This finding suggested that social media activity played a major role in predicting smartphone engagement. For marketers, this means that social media platforms may remain important channels for future cellphone marketing. Campaigns involving influencers, short-form videos, social commerce, product reviews, and interactive advertisements may be effective for users with high social media activity.

Daily screen time was also one of the strongest predictors. This means that the total amount of time users spent on smartphones was strongly related to engagement classification. From a marketing perspective, high screen time users may have more opportunities to encounter mobile advertisements and promotional content. However, marketers should be careful not to interpret high screen time only as a business opportunity. Excessive screen time may also indicate possible dependency, stress, or reduced digital well-being.

Weekend screen time was another important variable. Addicted users had an average weekend screen time of 10.21 hours, compared with 6.89 hours among non-addicted users. This suggested that weekends may be a strong period for mobile engagement. Cellphone companies and advertisers may consider weekend-focused campaigns, entertainment promotions, gaming offers, or social media-based product launches. However, these campaigns should still follow responsible marketing practices.

The comparison of algorithms showed that tree-based ensemble models performed better than traditional models. XGBoost achieved the highest F1-score and ROC-AUC, while the Hybrid Random Forest–XGBoost model achieved the highest precision. This means that ensemble learning was effective for analyzing smartphone behavior because it could capture nonlinear relationships among features. Traditional models such as Logistic Regression and Naïve Bayes performed lower because they may not fully capture complex behavioral patterns.

Although XGBoost slightly outperformed the hybrid model in F1-score, the hybrid model remained highly valuable. Its precision of 0.9929 showed that it was very reliable when identifying high-engagement users. In marketing prediction, this can be useful when a company wants to avoid incorrectly targeting users who are not actually highly engaged. For example, if a cellphone company wants to launch a premium campaign or personalized offer, the hybrid model may help identify users who are most likely to match the target behavior.

The confusion matrix further supported this result. The hybrid model produced only 7 false positives, compared with 19 false positives for XGBoost. This means that the hybrid model was better at avoiding incorrect positive classifications. However, XGBoost had fewer false negatives, meaning it was better at identifying more actual high-engagement users. Therefore, the choice between XGBoost and the hybrid model depends on the marketing objective. If the goal is to reach as many high-engagement users as possible, XGBoost may be preferred. If the goal is to target only the most certain high-engagement users, the hybrid model may be preferred.

The results also showed that some variables had limited predictive value. Age, gender, stress level, and academic/work impact did not appear to strongly separate addicted and non-addicted users. This suggested that actual smartphone behavior was more useful than demographic information. For future cellphone marketing, this means that behavior-based segmentation may be more effective than traditional demographic segmentation.

The findings have several marketing implications. First, cellphone marketers may use smartphone behavior data to identify high-engagement users. Second, social media activity and screen time may help guide digital campaign planning. Third, weekend usage patterns may support time-based marketing strategies. Fourth, feature importance results can help businesses understand which behaviors are most relevant for predicting user engagement.

However, the findings must be interpreted ethically. Since the target variable was related to smartphone addiction, the model should not be used to exploit vulnerable users. High-engagement users should not automatically be targeted with excessive advertisements or notifications. Instead, the results should support responsible marketing strategies, such as personalized but moderate promotions, wellness-centered messaging, digital balance features, and user-controlled notification settings.

The findings should be interpreted primarily as smartphone engagement prediction rather than direct marketing prediction. Users exhibiting higher engagement may represent behavioral segments that can support personalized marketing analytics. However, the dataset does not contain purchase intention, advertisement response, conversion behavior, or actual buying decisions. Therefore, the proposed model should be considered an engagement classification framework with potential applications in behavioral segmentation rather than a direct marketing prediction model.

IV. Conclusion and Recommendations

A. *Conclusion*

This study developed and evaluated a machine learning framework for predicting smartphone engagement intensity using behavioral smartphone usage data. Among the evaluated classifiers, XGBoost achieved the highest overall predictive performance, while the proposed Hybrid Random Forest–XGBoost model demonstrated the highest precision by reducing false positive classifications. These findings indicate that smartphone behavioral variables, particularly social media usage, daily screen time, and weekend screen time, are effective predictors of engagement intensity. Although the framework provides valuable behavioral insights that may support customer segmentation and responsible digital marketing, it should not be interpreted as directly predicting purchase intention or marketing outcomes because such variables were not available in the dataset.

B. Recommendations

Future cellphone marketing strategies should use behavior-based segmentation rather than relying only on demographic information. Marketers should focus on key indicators such as screen time, social media activity, app usage, and weekend engagement when designing personalized campaigns.

It is also recommended that future studies include actual marketing-related variables such as purchase intention, brand preference, advertisement clicks, customer satisfaction, and product purchase history. This would improve the accuracy of marketing prediction and make the model more directly connected to consumer buying behavior.

For model development, future researchers should continue comparing the Hybrid Random Forest–XGBoost model with other algorithms, including deep learning and other ensemble techniques. Additional feature selection and class imbalance techniques may also be applied to improve prediction performance.

Lastly, cellphone marketers should apply the findings responsibly. Users with high smartphone engagement or addiction indicators should not be targeted with excessive advertisements or notifications. Instead, marketing strategies should promote personalization, digital wellness, ethical engagement, and user-centered communication.

ACKNOWLEDGMENT

The researcher gratefully acknowledges the Pangasinan State University for its support throughout the conduct of this study. The researcher also acknowledges Kaggle as the source of the dataset used in this research. Artificial intelligence tools were used solely for language refinement, organization of ideas, and formatting assistance. All research processes, findings, interpretations, and conclusions were carefully reviewed and verified by the researcher to ensure accuracy, originality, and adherence to academic integrity.

REFERENCES

[1] K. Arunkumar, R. Naik, and S. Murthy, Consumer Behavior. Chyren Publication, 2025.

[2] V. H. Nguyen, T. P. L. Nguyen, and T. V. Nguyen, “Customers’ intention to purchase recycled products: Building on the theory of planned behavior and value-belief-norm theory,” Journal of Global Responsibility, 2025, doi: 10.1108/JGR-01-2024-0007.

[3] M. Sabbaghi and S. Behdad, “Consumer decisions to repair mobile phones and manufacturer pricing policies: The concept of value leakage,” Resources, Conservation and Recycling, vol. 133, pp. 101–111, 2018, doi: 10.1016/j.resconrec.2018.01.015.

[4] M. T. Islam, P. Dias, and N. Huda, “Waste mobile phones: A survey and analysis of the awareness, consumption and disposal behavior of consumers in Australia,” Journal of Environmental Management, vol. 275, Art. no. 111111, 2020, doi: 10.1016/j.jenvman.2020.111111.

[5] S. Srivastava, N. Verma, D. Kumar, N. Singh, and K. Kumar, “Nomophobia as an emerging psychopathology: Psychophysiological mechanisms and clinical implications,” Annals of Neurosciences, 2025, doi: 10.1177/09727531251351082.

[6] G. Dewis, “Using data from the Households and the Environment Survey: Repair or replace: What are Canadians doing with their old cell phones and computers?” Statistics Canada, Ottawa, Canada, 2022. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/16-002-x/2022001/article/00001-eng.htm. [Accessed: 2025].

[7] W. Hou, D. Peng, J. Chu, Y. Jiang, Y. Chen, and F. Chen, “Building an agent-based simulation framework of smartphone reuse and recycling: Integrating privacy concern and behavioral norms,” Sustainability, vol. 17, no. 15, Art. no. 6885, 2025, doi: 10.3390/su17156885.

[8] C. P. Baldé et al., Global E-waste Monitor 2024. Geneva/Bonn: International Telecommunication Union and United Nations Institute for Training and Research, 2024. [Online]. Available: https://ewastemonitor.info/wp-content/uploads/2024/03/GEM_2024_18-03_web_page_per_page_web.pdf

[9] O. Ogiemwonyi, “Determinants of green behavior revisited: A comparative study,” Resources, Conservation and Recycling Advances, vol. 22, Art. no. 200214, 2024, doi: 10.1016/j.rcradv.2024.200214.

[10] Y. Zhang, S. Wu, and M. I. Rasheed, “Conscientiousness and smartphone recycling intention: The moderating effect of risk perception,” Waste Management, vol. 101, pp. 116–125, 2020, doi: 10.1016/j.wasman.2019.09.040.

[11] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, “Significance of machine learning in healthcare: Features, pillars and applications,” International Journal of Intelligent Networks, vol. 3, pp. 58–73, 2022, doi: 10.1016/j.ijin.2022.05.002.

[12] F. Furizal, A. Ma’arif, and D. Rifaldi, “Application of machine learning in healthcare and medicine: A review,” Journal of Robotics and Control (JRC), vol. 4, pp. 621–631, 2023, doi: 10.18196/jrc.v4i5.19640.

[13] S. Saturi, “Review on machine learning techniques for medical data classification and disease diagnosis,” Regenerative Engineering and Translational Medicine, vol. 9, pp. 141–164, 2023, doi: 10.1007/s40883-022-00273-y.

[14] C. Guo and J. Chen, “Big data analytics in healthcare,” in Knowledge Technology and Systems, Y. Nakamori, Ed. Singapore: Springer Nature Singapore, 2023, pp. 27–70, doi: 10.1007/978-981-99-1075-5_2.

[15] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, “A novel ensemble learning paradigm for medical diagnosis with imbalanced data,” IEEE Access, vol. 8, pp. 171263–171280, 2020, doi: 10.1109/ACCESS.2020.3014362.

[16] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, “A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data,” Information Sciences, vol. 572, pp. 574–589, 2021, doi: 10.1016/j.ins.2021.02.056.

[17] Y. Zhang, H. Yang, H. Cui, and Q. Chen, “Comparison of the ability of ARIMA, WNN and SVM models for drought forecasting in the Sanjiang Plain, China,” Natural Resources Research, vol. 29, pp. 1447–1464, 2020, doi: 10.1007/s11053-019-09512-6.

[18] F. Nie, W. Zhu, and X. Li, “Decision tree SVM: An extension of linear SVM for non-linear classification,” Neurocomputing, vol. 401, pp. 153–159, 2020, doi: 10.1016/j.neucom.2019.10.051.

[19] D. C. E. Saputra, Y. Maulana, E. Faristasari, A. Ma’arif, and I. Suwarno, “Machine learning performance analysis for classification of medical specialties,” in Proceedings of the 3rd International Conference on Electronics, Biomedical Engineering, and Health Informatics, T. Triwiyanto, A. Rizal, and W. Caesarendra, Eds. Singapore: Springer Nature Singapore, 2023, pp. 513–528, doi: 10.1007/978-981-99-0248-4_34.

[20] A. Ghavidel and P. Pazos, “Machine learning techniques to predict breast cancer in imbalanced datasets: A systematic review,” Journal of Cancer Survivorship, 2023, doi: 10.1007/s11764-023-01465-3.

[21] F. Nie, W. Zhu, and X. Li, “Decision tree SVM: An extension of linear SVM for non-linear classification,” Neurocomputing, vol. 401, pp. 153–159, 2020, doi: 10.1016/j.neucom.2019.10.051.

[22] G. Aguiar, B. Krawczyk, and A. Cano, “A survey on learning from imbalanced data streams: Taxonomy, challenges, empirical study, and reproducible experimental framework,” Machine Learning, 2023, doi: 10.1007/s10994-023-06353-6.

[23] P. Gnip, L. Vokorokos, and P. Drotár, "Selective oversampling approach for strongly imbalanced data," PeerJ Computer Science, vol. 7, Art. no. e604, 2021, doi: 10.7717/peerj-cs.604.

[24] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, "A novel ensemble learning paradigm for medical diagnosis with imbalanced data," IEEE Access, vol. 8, pp. 171263–171280, 2020, doi: 10.1109/ACCESS.2020.3014362.

[25] T. C. Tran and T. K. Dang, "Machine learning for prediction of imbalanced data: Credit fraud detection," in Proc. 15th International Conference on Ubiquitous Information Management and Communication (IMCOM), Seoul, South Korea, 2021, pp. 1–7, doi: 10.1109/IMCOM51814.2021.9377352.

[26] D. Albashish, A. I. Hammouri, M. Braik, J. Atwan, and S. Sahran, "Binary biogeography-based optimization based SVM-RFE for feature selection," Applied Soft Computing, vol. 101, Art. no. 107026, 2021, doi: 10.1016/j.asoc.2020.107026.

[27] S. H. Hashemi Petrudi and H. Sharifpour Arabi, "Barriers to product return in a circular supply chain: A case from a retailing industry," Annals of Operations Research, 2025, doi: 10.1007/s10479-025-06464-4.

[28] D. Kolcava, J. Scholderer, and T. Bernauer, "Do citizens provide political rewards to firms engaging in voluntary environmental action?" Journal of Cleaner Production, vol. 279, Art. no. 123564, 2021, doi: 10.1016/j.jclepro.2020.123564.

[29] A. J. Abougarair, A. A. Oun, S. I. Sawan, and A. Ma'arif, "Deep learning-based automated approach for classifying bacterial images," International Journal of Robotics and Control Systems, vol. 4, pp. 849–876, 2024, doi: 10.31763/ijrcs.v4i2.1423.

[30] H. Liu et al., "Benign and malignant diagnosis of spinal tumors based on deep learning and weighted fusion framework on MRI," Insights into Imaging, vol. 13, Art. no. 87, 2022, doi: 10.1186/s13244-022-01227-2.

[31] A. J. Abougarair, A. A. Oun, S. I. Sawan, T. Abougard, and H. Maghfiroh, "Cancer treatment precision strategies through optimal control theory," Journal of Robotics and Control (JRC), vol. 5, pp. 1261–1290, 2024, doi: 10.18196/jrc.v5i5.22378.

[32] algozee, "Smartphone Addiction Prediction Data," Kaggle. Accessed: May 21, 2026. [Online]. Available: https://www.kaggle.com/datasets/algozee/smartphone-addiction-prediction-data