

Machine Learning-Based Ternary Risk Stratification of Adolescent Social Media Addiction Using an Adapted Bergen Framework, PHQ-A Indicators, and Objective Screen-Time Data

Fatima Ali Mohammed^{a*}, Sarah Saadoon Jasim^b

^a Information Technology Management-Baghdad, Middle Technical University, Baghdad, Iraq. Email: fatima.alimo960@gmail.com

^b Information Technology Management-Baghdad, Middle Technical University, Baghdad, Iraq. Email: sara-sm@mtu.edu.iq

ARTICLE INFO

Article history:

Received: 04 /05/2026

Revised form: 02 /06/2026

Accepted : 03 /06/2026

Available online: 30 /09/2026

Keywords:

Social media addiction;
Adolescents; Bergen Scale; PHQ-A;
Screen time; Machine learning;
XGBoost; CatBoost; SMOTE; PSO

ABSTRACT

Social media addiction among adolescents is a complex psychological and behavioral construct, which cannot be identified simply by the screen-time duration. In this study, the development and validation of a machine learning model to identify and categorize the level of social media addiction among adolescents by combining various factors, including psychometric addiction indicators, depressive symptoms, demographic and behavioral measures, and empirical smartphone screen-time data. A local cross-sectional data set was obtained from 808 adolescents aged 13-18 years attending intermediate and preparatory schools in Baghdad, Iraq. The severity of social media addiction was quantified using an extended Bergen-based scale in Arabic language with 18 items grouped into six addiction dimensions: salience, modification, tolerance, withdrawal, conflict, and relapse. Following the Bergen Social Media Addiction Scale (BSMA) protocol, the three items per addiction dimension were averaged, and the averages were summed up to obtain a total score between 6 and 30. The classification criteria were established based on two different data sources: a national Iraqi study utilizing the same Bergen score threshold value of 24 to diagnose social media addiction and an additional item response theory study endorsing the use of Bergen scores as severity scales especially values around 19. The outcome variable was defined as a Bergen-anchored composite risk score, not as the Bergen score alone. The PHQ-A score alone. The adapted 18-item Bergen-based score was used as the primary addiction component, while the PHQ-A score was aligned to the same 6–30 range and incorporated as a supporting psychological severity indicator, not as an independent depression diagnosis. Participants were then classified into three risk categories: low, moderate, and high. Objective smartphone screen-time indicators collected over ten consecutive days were used only as dependent predictive features and were not included in target construction. Six machine-learning models were evaluated in their baseline configurations—Random Forest, XGBoost, LightGBM, CatBoost, SVM with RBF kernel, and Logistic Regression (L2)—along with optimized, hybrid, and soft-voting variants. The modeling pipeline used a stratified 20 train/test split, five-fold stratified cross-validation within the training data, and SMOTE applied only to the training set. The original ternary target distribution was imbalanced, with 313 low-risk, 474 moderate-risk, and 21 high-risk cases; after the train/test split, SMOTE balanced the training classes from 250/379/17 to 379/379/379 cases. In binary classification, the best performance was achieved by the PSO-optimized CatBoost+XGBoost hybrid model, with an accuracy of 91.77%, weighted F1-score of 0.9173, and MCC of 0.8452. This model slightly outperformed the best baseline model, Random Forest, which achieved test accuracy of 91.14%, weighted F1-score of 0.9131, and MCC of 0.8347. These findings suggest that combining adapted psychometric responses, PHQ-A-supported psychological severity information, self-reported behavioral variables, and objective screen-time indicators may support early risk stratification of adolescent social media addiction. However, the findings should be interpreted as internal, cross-sectional decision-support evidence, because objective screen-time data were available only for a subsample and the model was not designed to replace clinical assessment.

MSC..

<https://doi.org/10.29304/jqcm.2026.18.32867>

1. Introduction

Today, social media platforms are widely used for social and personal communication, entertaining content consumption, acquiring knowledge, etc., among teenagers. Despite significant advantages that social media provide in terms of socialization and education, excessive and poorly regulated usage raises serious issues associated with

*Corresponding author: Fatima Ali Mohammed

Email addresses: fatima.alimo960@gmail.com

Communicated by 'sub editor'

Problematic Social Media Use (PSMU) and viewed through the prism of behavioral addiction [1], [2]. According to components of addictive behavior model, there is a set of characteristics that are used to define addictive behaviors including salience, mood modification, tolerance, withdrawal, conflict, and relapse [2]. The latter aspects have been operationalized in Bergen Social Media Addiction Scale, which is extensively applied in evaluating addictive social media use and its correlation with mental symptoms [3]. Extensive literature evidence confirms that problematic and excessive social media use is positively correlated with depression, anxiety, psychological stress, sleep problems, and other psychological problems among adolescents [1], [4], [5].

The key challenge related to identifying risk of social media addiction among adolescents is that excessive screen time cannot always accurately identify this condition. High levels of screen time may indicate intensive usage of devices, but they do not necessarily demonstrate the absence of control over the process, experience of withdrawal-like distress, impairing functions, etc. At the same time, some adolescents might be reluctant to describe their behavior due to social desirability biases. Thus, there is a need for implementing screening strategy combining psychological tests with behavioral features. The adapted version of Bergen Framework was developed for adolescents and kept the 6–30 score system. The classification of high-risk and low-risk cases was based on two references. Firstly, a local Iraqi study applied the cutoff score of 24 in order to classify cases of social media addiction among high school students [6]. Second, there are results of Item Response Theory demonstrating that Bergen scores can indicate high probability of being at risk of social media addiction when scores equal to 14 or 19 and very high when scores are much higher [7].

It is well-known that problems related to social media addiction are often accompanied by mental and physical health problems, specifically depression. Thus, Patient Health Questionnaire for Adolescents (PHQ-A) was added to the model, but it was considered only as additional information for psychological symptom intensity. Screen time was selected as an objective behavioral feature, but it was used only as predictor and not combined with Bergen score to construct the target variable.

Machine learning presents a promising paradigm for incorporating diverse types of datasets, such as psychometric measures, psychological symptoms, demographic data, self-reported behavior, and objective screen-time measurements. As recommended by the existing guidelines on developing prediction models using the regression analysis or machine learning methods, this research applies a supervised learning pipeline with several classifiers, including Random Forest, XGBoost, LightGBM, CatBoost, Support Vector Machines, and Logistic Regression with L2 regularization. As an approach to balancing the classes, SMOTE will be implemented exclusively on the training set to prevent majority class bias without contamination of the test set. Moreover, metaheuristics, such as Particle Swarm Optimization, Grey Wolf Optimizer, and Harris Hawks Optimization will be employed for optimizing hyperparameters.

Recent machine-learning studies have increasingly examined problematic smartphone use, Internet addiction, and social media addiction using psychological, behavioral, demographic, and digital-use indicators. Marengo et al. [21] used Random Forest to examine the contribution of smartphone application use to social media addiction tendency among adolescents, while Xiao et al. [22] applied regularized and tree-based models to identify psychological predictors of problematic smartphone use. Kim et al. [23] compared several machine-learning and deep-learning models, including Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, MLP, and CNN, for predicting problematic smartphone use among children and adolescents. Similarly, Jović et al. [24], Mim et al. [25], Ali et al. [28], Nurjoko et al. [29], and Tegar et al. [30] demonstrated the usefulness of machine-learning models for predicting or classifying Internet addiction, social media addiction, or related digital-addiction risks. In addition, Ehsan and Basit [26] classified social media addiction into low, medium, and high levels, while Hylkilä et al. [27] identified graded low-, moderate-, and high-risk profiles using clustering and regression-based analysis. These studies show that machine-learning models are relevant for digital-addiction research and can capture complex relationships among psychological symptoms, usage patterns, and behavioral variables. However, the present study differs from prior work by focusing on Iraqi adolescents and by combining a Bergen-anchored composite target, PHQ-A-supported psychological severity information, item-level psychometric responses, self-reported behavioral variables, and objective smartphone screen-time indicators within a leakage-controlled ternary modeling pipeline. To the best of our knowledge, no previous study has evaluated a PSO-optimized CatBoost+XGBoost hybrid model for ternary social media addiction risk stratification in an Iraqi adolescent sample.

Although previous studies have examined problematic social media use using psychometric indicators, psychological variables, behavioral measures, and machine-learning techniques, the evidence remains limited in the Iraqi adolescent context. More specifically, prior work has not sufficiently examined a Bergen-anchored composite risk framework supported by PHQ-A severity information, item-level psychometric responses, self-reported behavioral variables, and objective smartphone screen-time indicators for ternary risk stratification among Iraqi adolescents. From a modeling perspective, the present study addresses this gap by evaluating a leakage-controlled ternary machine-learning pipeline and by testing whether a PSO-optimized CatBoost+XGBoost hybrid model can improve risk classification compared with individual baseline classifiers.

Therefore, this study proposes a machine-learning-based framework for classifying adolescent social media addiction risk into three clinically interpretable risk levels. The novelty of the study lies in integrating a Bergen-anchored composite target with PHQ-A severity information, item-level psychometric responses, demographic and self-reported behavioral variables, and objective screen-time indicators within a leakage-controlled evaluation pipeline. The proposed model is intended as an early risk-screening and decision-support tool for researchers and practitioners, not as an autonomous clinical diagnostic system.

Based on this objective, the study addresses the following research questions:

RQ1: Can a hybrid CatBoost+XGBoost model optimized using Particle Swarm Optimization (PSO) outperform individual baseline classifiers in ternary social media addiction risk classification?

RQ2: Do objective smartphone screen-time features provide incremental predictive value when combined with psychometric and self-reported behavioral features?

RQ3: Does a ternary risk scheme provide a more informative early-risk stratification structure than a binary addicted/non-addicted classification approach?

1.1. Structure of the Paper

The rest of the paper is organized as follows. Section 2 describes the experimental design, subjects, instruments used for data collection, target development, pre-processing techniques, and the proposed machine learning workflow. Section 3 presents the results of the three-way classification task and evaluates the effectiveness of baseline, improved, hybrid, and ensemble classifiers. Section 4 offers an analysis of the results, discusses the significance of the methodology, its limitations, and practical application of the proposed approach to screening. Section 5 concludes the paper and offers suggestions for future studies.

2. Methodology

2.1. Design and Study Participants

This research utilized a cross-sectional study design based on the school context in order to develop and assess the internal validity of a ternary machine-learning model to predict the likelihood of social media addiction in adolescents. This analysis used a dataset obtained locally, specifically from students enrolled in intermediate and preparatory schools in Baghdad, Iraq. After data validation and organization, the dataset contained a total of 808 participants ranging between 13 to 18 years old. Each participant was assigned an anonymous identification number in order to match their survey results with the available smartphone screen-time records without violating confidentiality requirements. The sample consisted of both males and females within the two educational levels of interest. As illustrated in Table 1, males comprised 56.3% and females made up 43.7% of the sample size. The age distribution was evenly divided, whereby 51.4% of the sample were aged between 13 to 15 years, whereas 48.6% were between 16 to 18 years. Educational stage-wise, 44.2% were intermediate-level school students and 55.8% were preparatory level students. The smartphone screen-time data was only available for a sub-sample of participants and matched with the survey data using the same anonymous participant identification number.

Table 1. Participant Demographics

Variable	Category	Frequency	Percentage (%)
Gender	Male	455	56.3
Gender	Female	353	43.7
Age group	13–15 years	415	51.4
Age group	16–18 years	393	48.6
School stage	Intermediate	357	44.2
School stage	Preparatory	451	55.8

Table 1 shows the main demographic features of the study sample. The dataset for analysis consisted of 808 adolescents who were enrolled in intermediate and preparatory schools in Baghdad. With respect to gender, the ratio between males (56.3%) and females (43.7%) was almost equal. There was also an even split in terms of age, as 51.4% of the subjects belonged to the age range of 13-15 years, whereas 48.6% were between 16-18 years old, ensuring that both early and late adolescent groups were adequately included. In relation to education, preparatory school participants constituted 55.8%, while intermediate school participants made up 44.2%.

2.2. Data Sources and Self-Reported Behavioral Variables

Two data sources were used. First, the main data source included the structured questionnaire which measured demographic variables, self-reported behavioral variables, Bergen-based social media addiction items, and PHQ-A items measuring symptoms of depression. Self-reported behavioral variables reflected the adolescents' reports of their behaviors associated with social media use including the number of hours used daily, the most used social media platform, and the duration of account ownership. As can be seen from Table 2 below, many adolescents report spending 3 to 5 hours daily on social media and even 5+ hours daily. For example, the number of adolescents spending 3 to 5 hours was 287 (35.5%), and those spending more than 5 hours was 216 (26.7%), giving the total of 503 adolescents (62.2%) who used social media for no less than 3 hours daily. As far as the most popular platform is concerned, the leading social media platform was Instagram with 367 adolescents (45.4%). The second most popular was TikTok, with 361 adolescents (44.7%). Finally, as regards the ownership of accounts, the greatest part of adolescents (41.0%) reported owning their social media account(s) for 1 to 3 years, whereas for 4 to 6 years, there were 32.4% adolescents.

Table 2. Adolescents' Reported Behavioral Features Associated with Social Media Use

Variable	Category	Frequency	Percentage (%)
Daily social media use duration	Less than 1 hour/day	63	7.8
Daily social media use duration	1–2 hours/day	242	30.0
Daily social media use duration	3–5 hours/day	287	35.5
Daily social media use duration	More than 5 hours/day	216	26.7
Most-used social media platform	TikTok	361	44.7
Most-used social media platform	Instagram	367	45.4
Most-used social media platform	Facebook	39	4.8

Most-used social media platform	Snapchat	41	5.1
Duration of social media account ownership	1-3 years	331	41.0
Duration of social media account ownership	4-6 years	262	32.4
Duration of social media account ownership	6-9 years	114	14.1
Duration of social media account ownership	10 years or more	89	11.0

Table 2 below highlights the behavioral factors associated with social media usage. It was found that the daily usage pattern was mainly concentrated in the medium-high category, with 62.2% of individuals using social media for three or more hours daily. There was a clear preference for two main social media platforms, namely Instagram and TikTok, which made up more than 90% of the most commonly used platforms. Over half of the subjects had their social media accounts for 1 to 6 years.

2.3. Psychometric Measures

Two psychometric measures were used in order to estimate both addiction and psychological factors that characterize the social media addiction risk among adolescents. One psychometric measure is the adapted and extended Bergen Social Media Addiction Scale developed on the basis of the six-factor model of behavioral addiction. Adaptation included translating the scale into Arabic and expanding its conceptual scope beyond the Facebook-related addiction. The adaptation process involved the examination of the psychometric characteristics of the scale by eight psychologists in order to increase the relevance for adolescents.

The extended Bergen Social Media Addiction Scale includes 18 questions spread across six addiction components: salience, mood modification, tolerance, withdrawal, conflict, and relapse. Three items were used to assess each addiction component, instead of one item used in the original Bergen Social Media Addiction Scale. The use of three items for each addiction component helps to obtain a more detailed estimation of the addictive behavior of the adolescent. At the same time, because of the necessity to maintain the comparability of the score with the Bergen Social Media Addiction Scale, it was not possible to add up 18 items. Therefore, three items related to each addiction component were added up, and their mean values were used to calculate the final Bergen score in the range between 6 and 30.

Another psychometric measure used in the study is the Patient Health Questionnaire for Adolescents (PHQ-A). The PHQ-A contains nine items that are evaluated in terms of the 0-3 scale. The obtained score is calculated in the range between 0 and 27. In the present study, the PHQ-A was considered as an additional psychological factor that correlates with social media addiction risk. This approach allowed incorporating depression as a secondary variable for estimating the social media addiction risk.

The internal consistency reliability was analyzed using Cronbach's alpha coefficient for two psychometric measures used in the study. According to Table 3, both scales had a high level of internal consistency (Cronbach's alpha >0.80).

Table 3. Internal Consistency Reliability of the Psychometric Scales

Instrument	Number of Items	Cronbach's Alpha	Reliability Interpretation
Adapted Bergen-based Social Media Addiction Scale	18	0.809	Good
Patient Health Questionnaire for Adolescents (PHQ-A)	9	0.813	Good

Table 3 shows the reliability of the psychometric tools used in the study. The value of Cronbach's Alpha was above 0.80 in both the modified Bergen instrument and PHQ-A. This suggests high internal consistency and justifies their inclusion in constructing the overall risk assessment score using ternary classification.

2.4. Objective Smartphone Screen-Time Data

Objective smartphone screen-time data were utilized to provide behavioral evidence about adolescents' real-world digital engagement habits. In comparison to subjective estimates of daily phone use, which can be biased or even intentionally understated, screen-time records provide more direct evidence of participants' exposure to smartphones and social media. Screen-time data were collected from a subset of participants who voluntarily agreed to share screenshots demonstrating smartphone use on ten consecutive days. Every record of screen-time was matched with the response to the questionnaire through an identical anonymous participant ID code, which allowed combining psychometric, demographic, subjective behavioral, and objective behavioral data at the participant level while protecting confidentiality.

Screengrab images covered all types of smartphones (iOS and Android devices). Specifically, for iOS devices, the amount of phone usage was extracted from the Screen Time interface, whereas for Android, usage data were retrieved from either the Digital Wellbeing interface or App Activity Details interface. Considering differences in visual presentation of usage indicators and categories among operating systems, a common extraction protocol was developed. All time indicators were transformed into minutes, either from hours-and-minutes format or from only minutes. Indicators extracted from the screens included overall screen time, overall social media usage, and time spent on apps, including Instagram, TikTok, Facebook, and Snapchat.

If there was no social media category for the given phone, overall social media usage was estimated by adding times when particular social media applications were used. Extracted screen-time variables were considered as objective behavioral predictors in ternary classification models. Importantly, none of these variables were used for constructing the target variable and only were included as additional behavioral features. Such approach was required to avoid circularity between predictor creation and outcome formation and verify the usefulness of objective behavior variables in predicting addiction risk.

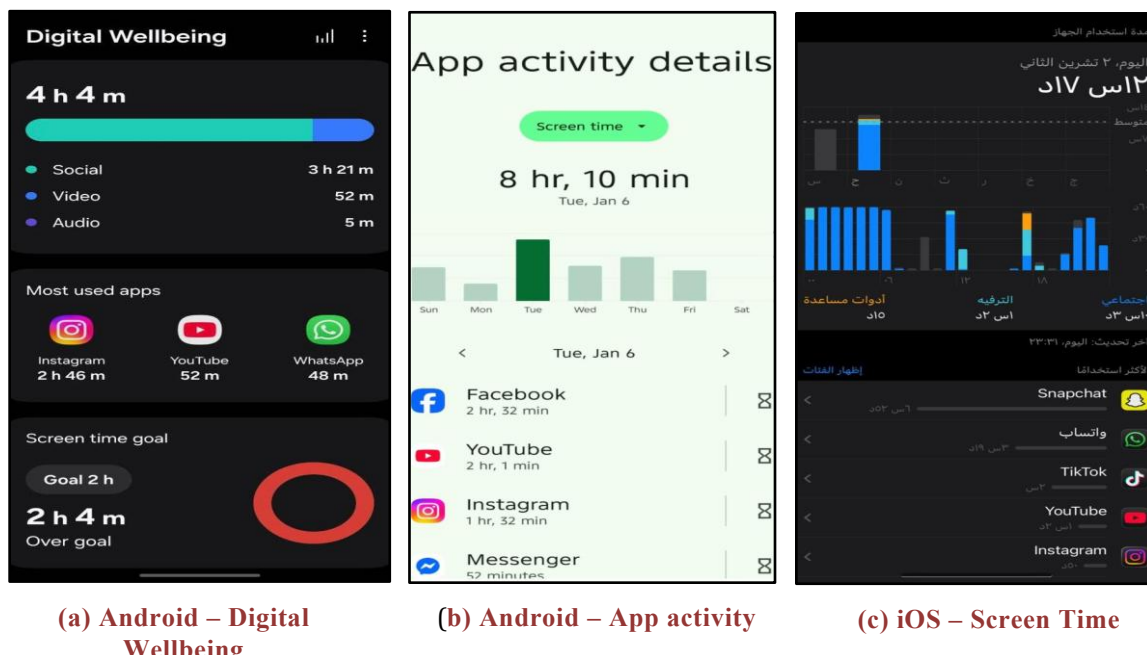


Figure 1. Sample interfaces used for collecting data regarding smartphone screen-time: (a) Android Digital Wellbeing, (b) Android App Activity Details, and (c) iOS Screen Time.

Figure 1 illustrates the sample interfaces from which the objective behavioral indicators have been collected. The screen shots show that screen-time data may be presented in various ways depending on whether the mobile device uses the Android operating system or the Apple iOS. As a result, all the time values in minutes have been used consistently without regard to the type of mobile device. The extracted behavioral indicators include the total screen time, social media usage, and application usage for Instagram, TikTok, Facebook, and Snapchat.

2.5. Construction of the Ternary Target

The dependent variable was constructed as a **Bergen-anchored composite risk score**. First, the adapted Bergen score was calculated by averaging the three items within each of the six addiction dimensions and summing the six dimension means, preserving the Bergen-compatible 6–30 range. The PHQ-A total score was then linearly rescaled to the same 6–30 range and incorporated as a supporting psychological severity component, rather than as an independent depression diagnosis. The final composite score was calculated as follows:

$$\text{Adjusted PHQ-A} = 6 + (\text{PHQ-A} / 27) \times 24$$

$$\text{Composite risk score} = 0.80 \times \text{Bergen score} + 0.20 \times \text{Adjusted PHQ-A}$$

This 80:20 weighting was adopted to keep the outcome primarily anchored in social media addiction symptoms while allowing depressive symptom severity to support risk stratification. Participants were classified as low risk (<14), moderate risk (≥ 14 and <24), or high risk (≥ 24). Objective smartphone screen-time indicators were not used to construct the target and were retained only as independent predictive features.

Table 4. Ternary classification thresholds for the Bergen-anchored composite risk score

Risk Level	Score Range	Interpretation
Low risk	<14	Below the early severity boundary.
Moderate risk	≥ 14 and <24	Intermediate risk band below the locally supported high-risk cutoff.
High risk	≥ 24	Meets or exceeds the locally supported high-risk cutoff.

Table 4 summarizes how the continuous Bergen-anchored composite risk score was converted into three interpretable risk categories. Because the adapted Bergen component was calculated by averaging the three items within each addiction dimension and then summing the six dimension means, the resulting score may include decimal values, the low-risk group was defined as scores below 14, while the high-risk group was defined as scores of 24 or higher. The moderate-risk group was retained as a single intermediate band because the present study focuses on ternary risk stratification rather than a finer four-level severity classification. Subdividing the 14–23 interval into 14–18 and 19–23 would be more suitable for a quaternary framework, whereas the purpose of this model is to distinguish low-risk, transitional/moderate-risk, and high-risk adolescents. Therefore, the adopted thresholds provide a practical and interpretable three-class structure for early risk detection, while preserving the locally supported high-risk cutoff of 24.

2.6. Preprocessing and Feature Engineering

Data preprocessing was done to prepare questionnaire, psychometric, and screen-time variables for classification using machine learning models with ternary outputs. First, the dataset was inspected to find any missing data, inconsistent categorical entries, and incorrect participant codes. Important psychometric responses needed to compute the composite risk score were checked, and only valid observations were selected. Then, the categorical variables – including gender, school grade, self-reported duration of use, most frequently used social media platform, and account age – were standardised and encoded into numerical format.

Numerical conversion of the Bergen items and the PHQ-A items was performed according to the response scale associated with them. The Bergen score was calculated as the average of the individual components, preserving the range of 6 to 30. On the other hand, the PHQ-A score was calculated based on its nine items, and subsequently transformed to correspond with the Bergen score range. Lastly, smartphone screen-time entries were standardized by transforming them to minutes since some screenshots may have shown time in hours and minutes and others in minutes only, thus ensuring that both iOS and Android smartphones were handled equally.

Feature engineering was applied after target construction using a leakage-controlled predictor design. The final target label, composite risk score, Bergen total score, Bergen dimension-level means used to compute the total score, PHQ-A total score, and adjusted PHQ-A score were excluded from the predictor matrix before model fitting. This exclusion was applied to prevent direct score-to-label or label-to-feature leakage. Item-level Bergen and PHQ-A responses were retained only as observed psychometric response features, not as derived total scores or class labels. Therefore, the model should be interpreted as learning from item-level symptom response patterns, demographic characteristics, self-reported behavioral variables, and objective screen-time indicators, rather than from the constructed outcome itself. All data-dependent preprocessing and modeling operations, including scaling, SMOTE-based class balancing, and hyperparameter tuning, were performed using only the training data, while the independent test set was kept unchanged for final evaluation. This design was intended to reduce the risk of direct target leakage while preserving the study aim of integrating item-level psychometric responses, psychological symptom information, self-reported behavioral variables, and objective behavioral indicators for risk stratification.

Before the application of machine learning algorithms, a consistent preprocessing process for the generation of the feature matrix was employed for fair comparisons. Numerical features were scaled depending on the requirements of the particular algorithm, especially when scaling mattered for models such as the SVM with RBF kernel and Logistic Regression. The same feature matrix, after applying the preprocessing steps, was utilized for training both baseline, optimized, hybrid, and ensemble models.

2.7. Machine-Learning Pipeline

The machine-learning pipeline was developed for evaluation of the ability of several supervised learning models to classify adolescents based on their social media addiction risks in three classes: low, medium, and high risks. As the final dataset consists of three target classes, it is required to maintain their proportional representation in training and test subsets during the train/test splitting procedure. Therefore, to balance the training dataset, stratified splitting of 80/20 was used; moreover, SMOTE was only applied to the training subset to prevent the synthetic minority samples from contaminating the test subset.

The six baseline algorithms were selected among the most popular machine learning models: namely, Random Forest, XGBoost, LightGBM, CatBoost, support vector machine with RBF kernel, and logistic regression with L2 regularization. Tree-based ensemble learning methods (Random Forest, LightGBM, and CatBoost) and gradient boosting (XGBoost) can be grouped together as both belong to iterative learning methods but use different decision tree training methods. Moreover, support vector machines and regularized logistic regression are examples of discriminative binary classifiers with different loss functions. Therefore, the same preprocessing and feature extraction were used for all models to make sure that performance differences are due to learning behavior.

In addition to baseline models, advanced steps of modeling process were applied to improve the predictive power of the algorithm. In particular, hyperparameter optimization with PSO, GWO, and HHO was carried out. Hybrid modeling was also considered when applying the probabilistic outputs from one model to be used as new features to another classifier. Soft-voting ensemble was then constructed from several models with high generalization capabilities. Thus, baseline models, hyperparameter tuning, hybrid algorithms, and ensemble were compared using the same ternary classification problem.

To improve methodological transparency and reproducibility, the main experimental settings were fixed across all model comparisons. The same stratified train/test split, random seed, preprocessing strategy, SMOTE procedure, and evaluation metrics were used for baseline, optimized, hybrid, and ensemble models. Data-dependent

preprocessing steps were fitted only on the training data, and the test set was kept unchanged throughout model evaluation. Table 5 summarizes the main experimental configuration used in the ternary classification pipeline.

Table 5. Experimental configuration of the ternary machine-learning pipeline

Setting	Adopted configuration
Classification task	Ternary risk classification: low, moderate, and high risk
Train/test split	Stratified 80/20 split
Random seed	42
Cross-validation	Five-fold stratified cross-validation within the training data
Class imbalance handling	SMOTE applied only to the training set
Test set	Kept unchanged and not used in SMOTE, tuning, or model fitting
Baseline classifiers	Random Forest, XGBoost, LightGBM, CatBoost, SVM-RBF, Logistic Regression-L2
Main optimized ternary model	CatBoost+XGBoost hybrid optimized using PSO
CatBoost stage	200 iterations, learning rate = 0.10, depth = 6
XGBoost stage	200 estimators, learning rate = 0.10, max depth = 6
PSO setting	15 particles and 20 iterations
PSO search space	n_estimators: 50–300; max_depth: 2–15; learning_rate: 0.01–0.30
Final meta-model	Logistic Regression
Main evaluation metrics	Accuracy, weighted F1-score, balanced accuracy, and MCC

The fixed settings summarized in Table 5 were adopted to ensure that performance differences among models were attributable to the learning algorithms and model architectures rather than to changes in data splitting, class balancing, or evaluation protocol. The PSO search was also conducted only within the training process, using cross-validation to select candidate hyperparameters, while the independent test set was used only once for final model evaluation.

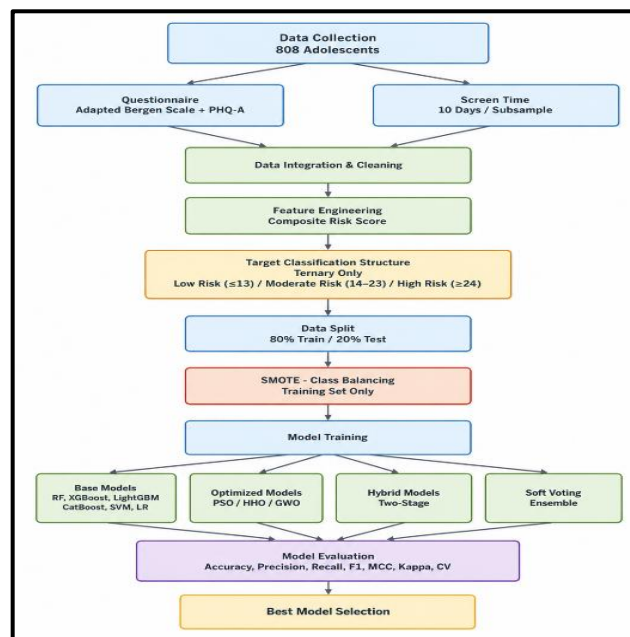


Figure 2 illustrates a schematic diagram of a ternary machine-learning pipeline used in forecasting adolescent social media addiction risk.

As shown in Figure 2, the suggested process starts with multidomain datasets, then moves through preprocessing, ternary target creation, feature matrix development, stratified train/test split, training the model with SMOTE on the training set alone, and ends with evaluation.

2.8. Model Evaluation

The models were evaluated using both global and class-specific metrics for a more complete assessment of the proposed ternary classification approach. As the outcome variable contains three levels (low, medium, and high), and there is no equal representation of classes within the dataset, accuracy alone was considered inadequate for model selection. For this reason, the models were evaluated by accuracy, precision, recall, F1 score, balanced accuracy, Matthews correlation coefficient (MCC), and Cohen's kappa. The F1 score is especially valuable in assessing how well precision and recall are balanced. Balanced accuracy, MCC, and Cohen's kappa are especially useful in estimating performance when there is class imbalance.

The test set was used for evaluating the performance of the proposed models after training, balancing, optimizing, and constructing the model. The test set was not included in SMOTE oversampling, hyperparameter tuning, and model fitting to prevent any bias. Furthermore, five-fold stratified cross-validation was used during the training phase to evaluate the internal consistency of the model and reduce its dependency on a specific split of the dataset. The best-performing model was chosen based on its performance on the test set, balanced class metrics, and cross-validation results.

3. Results

3.1. Target Class Distribution and SMOTE Application

Before evaluating the ternary classification models, the distribution of the target classes was examined to clarify the degree of class imbalance in the dataset. The original dataset was first divided using a stratified 80/20 train/test split to preserve the relative distribution of the low-risk, moderate-risk, and high-risk classes. SMOTE was then applied only to the training set, while the independent test set was kept unchanged for final evaluation. Table 5 reports the class distribution in the full dataset before splitting, in the training set before SMOTE, and in the training set after SMOTE.

Table 5. Ternary target class distribution before and after SMOTE

Risk class	Full dataset before split	Training set before SMOTE	Training set after SMOTE
Low risk	313	250	379
Moderate risk	474	379	379
High risk	21	17	379
Total	808	646	1137

As shown in Table 5, the original ternary target distribution was imbalanced, particularly for the high-risk class. After the stratified split, the same imbalance pattern was retained within the training data. SMOTE was therefore applied exclusively to the training set to increase the representation of minority classes during model learning. The test set was not oversampled or modified, ensuring that final model evaluation was conducted on original unseen data.

3.2. Ternary Classification Performance Analysis

The task of the present study involves ternary classification of adolescents into low-risk, moderate-risk, and high-risk groups of social media addiction based on specific criteria. In comparison with the binary classification approach, the ternary model will allow the identification of transitional cases where an individual shows symptoms of social media addiction but does not yet reach the high-risk threshold. Therefore, due to possible overlaps among the boundaries of the three categories, in addition to accuracy, evaluation considered weighted F1-score and Matthews correlation coefficient (MCC).

According to Table 5, all tested models showed good performance in the problem of ternary classification. One of the most accurate models involved combining CatBoost and XGBoost with the use of Particle Swarm Optimization (PSO). It reached the highest test accuracy of 91.77%, a weighted F1-score of 0.9173, and an MCC of 0.8452. Hence, it can be concluded that the hybrid and optimized algorithm was highly effective in predicting the three different risk categories based on the data collected.

The effectiveness of the proposed model is explained by its two-stage learning structure that includes the use of CatBoost for the analysis of heterogeneous data and XGBoost for capturing complex non-linear relationships between predictors. Moreover, the hybridization of algorithms and optimization through PSO allowed enhancing the model's performance even further. An almost equally high level of performance could be attained through the combination of XGBoost and Grey Wolf Optimizer (GWO). Namely, this model also achieved the highest test accuracy of 91.77% and weighted F1-score of 0.9168.

In addition, CatBoost, XGBoost, and Random Forest, which were used in baseline models, also showed rather good performance in the current study. However, as a rule, they were inferior to the proposed models in terms of all three evaluation criteria used in research. This fact confirms the potential benefit from the integration of hybrid learning and optimization techniques in predicting adolescent risk of social media addiction.

Consequently, it can be stated that the developed machine learning algorithm enables stable and reliable classification among the three groups (low, moderate, high risk). The obtained results confirm the potential usefulness of the approach, which combines psychometric assessments adapted for adolescents, PHQ-A scores, and behavioral variables provided by adolescents themselves along with objective data on their screen time.

Table 6. Performance Comparison of the Main Models in Ternary Classification

Model	Accuracy	Weighted F1-score	Precision	Recall	Balanced Accuracy	MCC
CatBoost + XGBoost Hybrid + PSO	0.9177	0.9173	0.9210	0.9177	0.8123	0.8452
XGBoost + GWO	0.9177	0.9168	0.9185	0.9177	0.8096	0.8429
XGBoost baseline	0.9114	0.9108	0.9135	0.9114	0.8061	0.8318
Random Forest baseline	0.9114	0.9131	0.9177	0.9114	0.8061	0.8347
Soft Voting Ensemble	0.9051	0.9048	0.9087	0.9051	0.8026	0.8210
CatBoost baseline	0.8987	0.9004	0.9027	0.8987	0.8246	0.8077

In Table 6, a comparison evaluation of the major models used for the three-class problem is provided. It can be seen from the results obtained that the best tradeoff is achieved by the model based on the combination of CatBoost with XGBoost and PSO technique. This method exhibits excellent performance concerning the weighted F1-score and MCC indicators. However, even though the XGBoost combined with the GWO technique demonstrates the same test accuracy as the proposed approach, the use of the PSO hybridization gives slightly better results regarding the F1-score and MCC indicators. This suggests that hybridization along with the optimization algorithm may improve model stability and discriminatory ability.

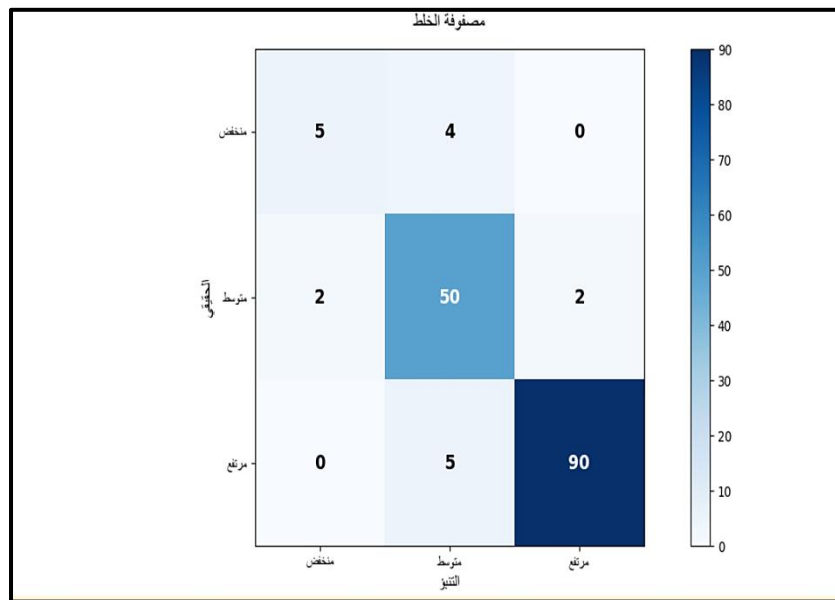


Figure 3. Confusion Matrix of the Optimal Ternary Classification Model

As illustrated in Figure 3, the confusion matrix corresponds to the best ternary classification model. This confusion matrix suggests that 145 correct classifications have been made out of 158 test records, which correlates with the test accuracy. In addition, the classification is primarily on the main diagonal.

3.3. INCREMENTAL VALUE OF OBJECTIVE SCREEN-TIME FEATURES

To address the second research question, a feature-ablation analysis was conducted using the Random Forest baseline model to evaluate whether objective smartphone screen-time indicators provide incremental predictive value beyond questionnaire-based psychometric and self-reported behavioral variables. The analysis was restricted to the complete-case screen-time subset, because objective screen-time records were available only for participants who provided valid ten-day smartphone usage data. Both experimental settings were evaluated using the same target definition, the same complete-case subset, the same train/test split, the same preprocessing protocol, the same SMOTE procedure applied only to the training set, and the same Random Forest baseline configuration. The only difference between the two settings was the exclusion or inclusion of objective screen-time features. This design was adopted to isolate the contribution of real smartphone usage indicators without changing the target construction or the evaluation protocol.

Table 7. Incremental predictive value of objective screen-time features using Random Forest baseline

Feature setting	N used	Test size	Model	Accuracy	Weighted F1-score	Balanced Accuracy	MCC
Without objective screen-time features	257	52	Random Forest baseline	0.8408	0.8335	0.5573	0.6637
With objective screen-time features	257	52	Random Forest baseline	0.9114	0.9075	0.7699	0.8278
Difference	—	—	—	+0.0706	+0.0740	+0.2126	+0.1641

As shown in Table 7, incorporating objective screen-time indicators improved the Random Forest baseline performance across all main evaluation metrics. Test accuracy increased from 0.8408 to 0.9114, while the weighted F1-score increased from 0.8335 to 0.9075. More importantly, balanced accuracy increased from 0.5573 to 0.7699, and MCC increased from 0.6637 to 0.8278. The largest gains were observed in balanced accuracy and MCC, suggesting that real smartphone usage indicators improved discrimination across the three risk categories rather than merely increasing overall accuracy. These findings support the incremental value of objective behavioral data for early risk stratification. Nevertheless, screen-time variables should be interpreted as supporting behavioral predictors, not as standalone diagnostic indicators or as components of the target variable.

4. Discussion

This study examines the development and evaluation of a machine-learning-based model for predicting the severity level of social media addiction risk of adolescents. Empirical evidence suggests that the proposed machine-learning model achieves high test accuracy in categorizing participants into low, moderate, and high-risk classes. Specifically, the best-performing model, namely the CatBoost + XGBoost Hybrid + PSO, scored 91.77%, 0.9173, and 0.8452 in terms of test accuracy, weighted F1-score, and MCC, respectively, indicating the overall soundness of the model concerning its predictive power, class-wise effectiveness, and the agreement between actual and predicted classes.

The obtained results should be interpreted in relation to the study's screening objective rather than as evidence of clinical diagnosis. The ternary structure was adopted to identify low-risk, moderate-risk, and high-risk adolescents, with particular attention to the moderate-risk group as a transitional category that may be missed in a binary addicted/non-addicted framework. From a modeling perspective, the improvement of the PSO-optimized CatBoost+XGBoost hybrid model over the best baseline model was modest but meaningful, because it was observed not only in accuracy but also in weighted F1-score and MCC. This indicates that the hybrid model did not merely increase the number of correct predictions, but also improved the agreement between predicted and true classes in an imbalanced multiclass setting. Therefore, the contribution of the proposed model lies in providing a more stable and interpretable early risk-stratification framework rather than replacing psychometric assessment or clinical evaluation.

First, it was reasonable to choose the three-class problem setting because the issue of social media addiction is better represented in such a framework. While a binary classification might be sufficient for some studies, this approach will miss out on the individuals with moderate or nascent symptoms who do not fit into a high-risk category. A three-level severity-based approach appears to be more suitable for practical use since it identifies a moderate-risk class, which is critical for early screening, detection within educational institutions, and prevention.

Moreover, it seems that the application of an ensemble learning approach can improve the model's ability to make accurate predictions. In particular, the combination of CatBoost and XGBoost approaches, which are based on gradient boosting decision trees algorithm, along with the PSO algorithm, helped achieve the best classification outcomes. CatBoost has advantages in tabular data modeling, while XGBoost demonstrates excellent results in dealing with complex and nonlinear decision boundaries. Presumably, their combination has improved the model's ability to differentiate the risk groups.

Furthermore, the superiority of the best-performing model compared with other variants of the hybrid ensemble learning algorithm can be explained by the higher optimization quality provided by the particle swarm

optimization algorithm. This is confirmed by the fact that the results obtained with the help of another optimization algorithm, namely the grey wolf optimizer, demonstrated high similarity.

In addition to methodological strengths, another advantage is the use of multidimensional data. For example, the proposed model does not use only screen-time duration but also incorporates a variety of factors, such as adapted Bergen-based symptoms, depression risk indicators based on the PHQ-A scale, self-reported behavioral traits, and objective behavioral variables. The use of these variables is reasonable because screen time is the amount of exposure to the internet but not a measure of risk.

Furthermore, the inclusion of the objective behavioral indicator variables as predictors is a reasonable choice. The exclusion of this type of variable from the targets allowed avoiding the circularity. Therefore, the proposed model should be considered a decision-support model for early detection of adolescents at risk of developing social media addiction.

Nevertheless, several limitations require mention. To begin with, the cross-sectional design of the study prevents the establishment of causality among social media usage, depressive symptoms, and social media addiction risk. Moreover, objective screen-time records were available only for a subsample of participants and were extracted from screenshots over a ten-day period, which may limit the generalizability of findings related to objective behavioral indicators. In addition, despite good psychometric properties, the adapted Bergen-based measure requires further validation, including measurement invariance analysis across gender and educational stage. Finally, the findings represent internal validation within a cross-sectional dataset; therefore, future studies should test the model prospectively and across independent Iraqi or regional adolescent samples before broader practical deployment.

5. Conclusion and Recommendations

In summary, the current paper describes the development and internal validation of a novel machine learning model that measures and classifies the risk of social media addiction in adolescents. Classification into low-risk, moderate-risk, and high-risk groups provides a more sophisticated measure compared with a simple classification into addicts and non-addicts. The model that demonstrated the highest level of performance includes a combination of CatBoost, XGBoost, and PSO with a test accuracy of 91.77%, F1-score of 0.9173, and MCC of 0.8452.

It is concluded that combining adapted scales and tests for measuring psychological aspects of addiction, PHQ-A criteria, reported behavioral characteristics, and actual screen-time use can contribute to the correct risk stratification for adolescents with respect to their probability to become social media addicts. Moreover, one of the strengths of the proposed approach is the possibility to classify individuals as moderate-risk and monitor them before they reach a high-risk state.

To further investigate the issue, it is recommended that future studies evaluate the model using larger data samples and conduct prospective analysis of patients. Researchers could also analyze the contribution of particular factors with help of explainable AI methods such as SHAP [19] and LIME [20]. On the other hand, in practice, the discussed model may assist schools, scientists, and psychologists as an instrument for early screening and decision-making support, although it cannot replace clinical diagnosis and psychological testing.

References

- [1] B. Keles, N. McCrae, and A. Grealish, "A systematic review: The influence of social media on depression, anxiety and psychological distress in adolescents," *International Journal of Adolescence and Youth*, vol. 25, no. 1, pp. 79–93, 2020, doi: 10.1080/02673843.2019.1590851.
- [2] M. D. Griffiths, "A components model of addiction within a biopsychosocial framework," *Journal of Substance Use*, vol. 10, no. 4, pp. 191–197, 2005, doi: 10.1080/14659890500114359.
- [3] C. S. Andreassen, J. Billieux, M. D. Griffiths, D. J. Kuss, Z. Demetrovics, E. Mazzoni, and S. Pallesen, "The relationship between addictive use of social media and video games and symptoms of psychiatric disorders: A large-scale cross-sectional study," *Psychology of Addictive Behaviors*, vol. 30, no. 2, pp. 252–262, 2016, doi: 10.1037/adb0000160.
- [4] J. M. Nagata *et al.*, "Social media use and depressive symptoms during early adolescence," *JAMA Network Open*, vol. 8, no. 5, e2511704, 2025, doi: 10.1001/jamanetworkopen.2025.11704.
- [5] Y. Xiao *et al.*, "Addictive screen use trajectories and suicidal behaviors, suicidal ideation, and mental health in youths," *JAMA*, 2025, doi: 10.1001/jama.2025.7829.

- [6] R. Y. Abdullah, B. M. S. Ismail, H. Q. Ezzat, and H. A. Sadeeq, "Social media addiction among high school students in Iraqi Kurdistan Region," *Egyptian Journal of Community Medicine*, vol. 42, no. 4, pp. 223–232, 2024, doi: 10.21608/ejcm.2024.271694.1286.
- [7] D. Zarate, B. A. Hobson, E. March, M. D. Griffiths, and V. Stavropoulos, "Psychometric properties of the Bergen Social Media Addiction Scale: An analysis using item response theory," *Addictive Behaviors Reports*, vol. 17, p. 100473, 2023, doi: 10.1016/j.abrep.2022.100473.
- [8] J. G. Johnson, E. S. Harris, R. L. Spitzer, and J. B. W. Williams, "The Patient Health Questionnaire for Adolescents: Validation of an instrument for the assessment of mental disorders among adolescent primary care patients," *Journal of Adolescent Health*, vol. 30, no. 3, pp. 196–204, 2002, doi: 10.1016/S1054-139X(01)00333-0.
- [9] G. S. Collins *et al.*, "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, e078378, 2024, doi: 10.1136/bmj-2023-078378.
- [10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [12] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [13] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.
- [15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [16] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, pp. 1942–1948, 1995, doi: 10.1109/ICNN.1995.488968.
- [17] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014, doi: 10.1016/j.advengsoft.2013.12.007.
- [18] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications," *Future Generation Computer Systems*, vol. 97, pp. 849–872, 2019, doi: 10.1016/j.future.2019.02.028.
- [19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [20] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [21] D. Marengo, M. A. Fabris, C. Longobardi, and M. Settanni, "Smartphone and social media use contributed to individual tendencies towards social media addiction in Italian adolescents during the COVID-19 pandemic," *Addictive Behaviors*, vol. 126, Art. no. 107204, 2022, doi: 10.1016/j.addbeh.2021.107204.
- [22] B. Xiao, N. Parent, L. Rahal, and J. Shapka, "Using Machine Learning to Explore the Risk Factors of Problematic Smartphone Use among Canadian Adolescents during COVID-19: The Important Role of Fear of Missing Out (FoMO)," *Applied Sciences*, vol. 13, no. 8, Art. no. 4970, 2023, doi: 10.3390/app13084970.
- [23] K. Kim, Y. Yoon, and S. Shin, "Explainable prediction of problematic smartphone use among South Korea's children and adolescents using a machine learning approach," *International Journal of Medical Informatics*, vol. 186, Art. no. 105441, 2024, doi: 10.1016/j.ijmedinf.2024.105441.
- [24] J. Jović, A. Čorac, A. Stanimirović, M. Nikolić, M. Stojanović, Z. Bukumirić, and D. Ignjatović Ristić, "Using machine learning algorithms and techniques for defining the impact of affective temperament types, content search and activities on the internet on the development of problematic internet use in adolescents' population," *Frontiers in Public Health*, vol. 12, Art. no. 1326178, 2024, doi: 10.3389/fpubh.2024.1326178.
- [25] M. N. Mim, M. Firoz, M. M. Islam, M. Hasan, and M. T. Habib, "A study on social media addiction analysis on the people of Bangladesh using machine learning algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, 2024, doi: 10.11591/eei.v13i5.5680.
- [26] T. Ehsan and J. Basit, "Machine Learning for Detecting Social Media Addiction Patterns: Analyzing User Behavior and Mental Health Data," *International Journal of Innovations in Science & Technology*, vol. 6, no. 4, pp. 1789–1807, 2024.
- [27] K. Hylkilä, M. Kääriäinen, A. Peltonen, S. Castrén, T. Mustonen, J. Konttila, and N. Männikkö, "Young adults' risk profiles and predictive factors of Problematic Social Media Use (PSMU): A cross-sectional study," *Current Psychology*, vol. 44, pp. 6848–6862, 2025, doi: 10.1007/s12144-025-07662-w.
- [28] A. Ali, M. S. Hosain, M. A. B. Siddik, M. Hasan, M. A. Habib, M. A. Kabir, M. M. Rahman, P. A. Shanto, N. Hasan, and A. Mahmud, "Classifying Internet Addiction Using Machine Learning Approach: A Study Among Adolescents in Bangladesh," *Public Health Challenges*, vol. 4, no. 4, Art. no. e70165, 2025, doi: 10.1002/puh2.70165.
- [29] Nurjoko, A. D. Praditya, N. Triyasri, M. R. Octa, and A. Rahardi, "Perbandingan Performa Model Naïve Bayes dan Regresi Logistik dalam Klasifikasi Kecanduan Media Sosial pada Siswa," *Journal of Data Science Methods and Applications*, vol. 1, no. 2, pp. 92–101, 2025, doi: 10.30873/jodmapps.v1i2.pp92-101.
- [30] A. T. B. S. Tegar, H. Hasanah, and I. Oktaviani, "Predicting Social Media Addiction Using Machine Learning and Interactive Visualization with Streamlit," *bit-Tech*, vol. 8, no. 1, pp. 778–788, 2025, doi: 10.32877/bt.v8i1.2715.