Journal of Deep Learning, Computer Vision and Digital Image Processing
E-ISSN: 2986-8939 P-ISSN: 2986-8920
Vol. 4, No. 2, June 2026
DOI. https://doi.org/10.61255/decoding.v4i2.1193

Attention Span Classification of Social Media Users Using Multi-Kernel Support Vector Machine Based on Survey Data

Reza Pahlevi✉
(Universitas Prima Indonesia, Medan, Indonesia.)

Darren Lucius
(Universitas Prima Indonesia, Medan, Indonesia.)

Diasta Natanael Sembiring
(Universitas Prima Indonesia, Medan, Indonesia.)

Delima Sitanggang
(Universitas Prima Indonesia, Medan, Indonesia.)

OPEN ACCESS

ARTICLE HISTORY
Received: April 27, 2026
Revised: May 25, 2026
Accepted: June 27, 2026

KEYWORDS
Attention span;
Machine learning;
Mental health;
Social media;
Support vector machine

ABSTRACT

Purpose – This study aims to examine whether self-reported attention-related difficulty categories among social media users can be classified using a leakage-free machine learning framework. It addresses the risk of inflated performance in survey-based classification by excluding the same items used to construct the target label from the predictor set.

Methods – The study used the public Social Media and Mental Health (SMMH) Kaggle dataset with 478 valid respondents. A three-class label was constructed from Q10, Q12, and Q14 using percentile thresholds (P33 = 9.0; P66 = 12.0), producing High (n = 208), Medium (n = 152), and Low (n = 118) categories. These label-generating items were excluded from predictors. The remaining variables were processed in a scikit-learn Pipeline using MinMax scaling, ordinal encoding, and One-Hot Encoding. Multi-kernel SVM models and five baseline classifiers were evaluated using a stratified 70:30 split, cross-validation, F1 metrics, balanced accuracy, and permutation importance.

Findings – Random Forest achieved the highest performance, with 63.19% accuracy and 62.26% weighted F1. Linear SVM was the best SVM model, achieving 61.81% accuracy, 60.08% weighted F1, 58.99% macro F1, and 59.11% balanced accuracy. The strongest predictors were Restless Without Social Media, Use Without Purpose, and Interest Fluctuation.

Research implications – The findings are preliminary, dataset-specific, and based on a survey-derived composite label whose internal reliability still requires validation.

Originality – This study contributes a leakage-controlled classification approach for analyzing attention-related survey categories.

Correspondence Author: ✉repahlevi89@gmail.com
To cite this article: Pahlevi, R., Lucius, D., Sembiring, D. N., & Sitanggang, D. (2026). Attention span classification of social media users using multi-kernel support vector machine based on survey data. Journal of Deep Learning, Computer Vision and Digital Image Processing, 4(2), 143–156. https://doi.org/10.61255/decoding.v4i2.1193

INTRODUCTION

The rapid development of social media platforms over the past decade has fundamentally changed how individuals consume information and allocate their daily time. In 2024, the number of active social media users worldwide surpassed 5.1 billion, representing approximately 63% of the global population. Internet users in Indonesia spend an average of 7 hours and 38 minutes per day on digital devices, including more than 3 hours specifically on social media [1]. The increasing intensity of social media engagement has raised concerns about its potential influence on cognitive functioning, particularly sustained focus. Neurophysiological evidence from EEG studies indicates that short-form video use negatively impacts attention functions [2]. The present study operationalizes this concern through survey-based self-report rather than objective measurement, and all findings must be interpreted accordingly.

In this study, attention span is operationalized as a self-reported composite score derived from three survey items measuring cognitive distraction indicators. The broader impact of digital technology and social media on cognitive functions has been documented in recent reviews [3]. This is an important distinction: the construct measured here is a *survey-level proxy* for self-reported attentional difficulty, not a clinically validated neuropsychological measure of attention capacity. It should not be equated with clinical diagnoses such as ADHD or with laboratory-grade sustained attention assessments. In cognitive psychology, attention span encompasses selective, sustained, and executive attention components [4], [5], but these dimensions require standardized instruments beyond what a short survey can capture. The present study makes no claim beyond classifying a percentile-derived category of a survey composite score [6].

Prior work has reported associations between intensive social media use and attention-related difficulties. Hawi and Samaha found excessive social media engagement associated with self-reported attention difficulties in university students [7]. Media multitasking has also been linked to reduced attentional capacity across multiple population samples [8]. Twenge et al. documented longitudinal declines in psychological well-being correlated with increased screen exposure [9], [10]. Systematic reviews have similarly documented cognitive and mental health correlates of social media use across adolescent populations [11]. However, most existing research relies on correlational or descriptive approaches, limiting predictive classification capacity. Machine learning offers a complementary framework for classifying self-reported difficulty levels from behavioral survey data [12], [13].

Support Vector Machine (SVM) has demonstrated competitive performance in high-dimensional classification problems with moderate sample sizes [14], [15], including mental health classification [16] and psychological survey data classification [17]. Its ability to model both linear and nonlinear decision boundaries through different kernel functions makes it a natural candidate for behavioral survey analysis. This study maintains a multi-kernel SVM focus for four substantive reasons that go beyond overall accuracy. First, the multi-kernel comparison Linear, RBF, and Polynomial constitutes a self-contained research question about how decision-boundary geometry interacts with the leakage-free feature space; this question cannot be answered by Random Forest, which is a black-box ensemble that does not expose a kernel-based decision boundary. Second, SVM with a linear kernel provides an interpretable geometric separation that is informative for understanding which features drive class distinctions, complementing the permutation importance analysis reported in Section III-G. Third, with n=478 after preprocessing, SVM's margin-maximization mechanism provides an inherent regularization advantage over ensemble methods that are more sensitive to sample size. Fourth, the performance difference between Random Forest (63.19%) and Linear SVM (61.81%) amounts to approximately two observations on a test set of 144 samples a gap that falls within random sampling variability and cannot be declared statistically meaningful without confidence intervals. This study does not presuppose SVM superiority; five baseline classifiers including Random Forest are empirically compared to contextualize SVM's standing. As reported in the results, Random Forest achieves the numerically highest accuracy and is explicitly discussed as

the strongest baseline, while the multi-kernel SVM analysis addresses the primary research questions about kernel choice and leakage-free behavioral classification.

The primary methodological contribution of this study is the explicit exclusion of Q10, Q12, and Q14 the three items that define the composite label from the predictor set. This directly addresses the label-feature dependency problem [18], [19], [20], [21] that would otherwise artificially inflate model performance estimates and was the central flaw of an earlier version of this work; the leakage-free analysis reported here supersedes any previously reported contaminated results. By removing these items, the study provides a more realistic estimate of whether the remaining behavioral and psychological indicators carry classifiable signal for the constructed attention category. A secondary contribution is the use of One-Hot Encoding for nominal predictors within a full scikit-learn Pipeline, ensuring preprocessing parameters are fitted exclusively on training data. A limitation acknowledged upfront is that the constructed label is a three-item composite whose psychometric reliability has not been formally tested in this study; the internal consistency of Q10, Q12, and Q14 as a scale warrants future assessment. Additionally, several non-label predictors including Q13 (Worry Level), Q18 (Depressed Frequency), and Q20 (Sleep Issues) are psychologically adjacent to the target construct, meaning the model predicts a cognitive-distraction composite partly from related self-reported psychological indicators rather than purely from behavioral social media usage variables. Readers should bear this in mind when interpreting the classification results.

Three research questions guide this study: (1) How do Linear, RBF, and Polynomial SVM kernels perform when classifying attention-related difficulty categories using only non-label-generating predictors? (2) Which kernel achieves the best leakage-free performance across accuracy, macro F1, balanced accuracy, and cross-validation? (3) Which behavioral and psychological variables are most informative for this classification according to permutation importance?

METHOD

Dataset

The Social Media and Mental Health (SMMH) dataset was obtained from the Kaggle platform. It comprises 481 respondents and 21 variables, including 12 mental health indicators measured on a five-point Likert scale (1=rarely/never, 5=very often). After preprocessing, 478 active social media users were retained. Table 1 summarizes the dataset characteristics.

Table 1. SMMH Dataset Characteristics

Characteristic	Description
Dataset name	Social Media & Mental Health (SMMH)
Source	Kaggle (publicly available)
Initial respondents	481
Valid respondents	478 (after preprocessing)
Variables	21 (1 timestamp, 8 behavioral/demographic, 12 Likert mental health items)
Gender	Female: 54.6%, Male: 44.1%, Non-binary/Other: 1.3%
Dominant occupation	University Student (60.7%), Salaried Worker (27.4%)
Modal daily SM time	More than 5 hours (24.1%)

Label Construction

The Attention Span category label was constructed computationally from three Likert items: Q10 (Distraction While Busy), Q12 (Easily Distracted Score), and Q14 (Difficulty Concentrating), each scored 1–5, producing a composite range of 3–15. Percentile-based thresholding was applied: P33=9.0 defines the High boundary (lower cognitive distraction), P66=12.0 defines the Low boundary (higher cognitive distraction). Table 2 details the scheme. The selection of Q10, Q12, and Q14 as the composite constituents was conceptually grounded: Q10 (Distraction While Busy)

captures the behavioral tendency to shift attention away from an ongoing task when social media is present; Q12 (Easily Distracted Score) directly operationalizes the self-perceived ease of attentional disengagement; and Q14 (Difficulty Concentrating) reflects subjective impairment in sustained focus. Together, these three items cover the core facets of self-reported attentional difficulty task interference, distractibility, and concentration impairment making them the most face-valid candidates in the SMMH instrument for constructing a cognitive-distraction composite. It is important to note that this three-item composite has not been subjected to formal reliability analysis (e.g., Cronbach's alpha or McDonald's omega) in the present study [22], [23]. The label should therefore be treated as an exploratory survey-derived proxy rather than a psychometrically validated scale. Future work should confirm inter-item consistency and validate the composite against standardized attention instruments.

This label does not constitute a validated clinical attention span measure. It is a survey-derived composite score representing self-reported attentional difficulty. Deployment as a clinical instrument would require validation against standardized neuropsychological tests such as the Continuous Performance Test or Conners' Rating Scales. Critically, Q10, Q12, and Q14 are entirely excluded from the predictor set in all experiments to prevent label-feature dependency [18].

Table 2. Attention Span Label Formation Scheme

Parameter	Value
Label components (EXCLUDED from predictors)	Q10 + Q12 + Q14 (Likert 1–5 each)
Composite range	3 (min) to 15 (max)
P33	9.0
P66	12.0
High (score ≤ 9.0)	208 respondents (43.5%) — lower cognitive distraction
Medium (9.0 < score ≤ 12.0)	152 respondents (31.8%) — moderate cognitive distraction
Low (score > 12.0)	118 respondents (24.7%) — higher cognitive distraction

Data Preprocessing

Three preprocessing steps were applied: (1) 30 missing values in Organization_Type were imputed with "Unknown"; this variable was subsequently recoded into Occupation_Status and retained as a nominal predictor, so no records were discarded due to missingness; (2) duplicate checking confirmed no duplicate records; (3) three respondents who reported not using social media were removed. Gender labels were normalized to Male, Female, and Non-binary by consolidating variant text representations. No preprocessing parameters were fitted on test data at any stage.

Leakage-Free Feature Set

The predictor set was constructed by explicitly excluding Q10, Q12, and Q14. The remaining 15 features (after encoding expansion) span four encoding strategies, as detailed in Table 3.

Table 3. Leakage-Free Predictor Set (Q10, Q12, Q14 Excluded from All Experiments)

No	Variable	Survey Item	Scale	Encoding
1	Q9_Purposeless	Use Without Purpose Frequency	1–5 Likert	MinMaxScaler (numeric)
2	Q11_Restless	Restless Without Social Media	1–5 Likert	MinMaxScaler (numeric)
3	Q13_Worries	Worry Level Score	1–5 Likert	MinMaxScaler (numeric)

No	Variable	Survey Item	Scale	Encoding
4	Q15_Compare	Compare With Others Frequency	1–5 Likert	MinMaxScaler (numeric)
5	Q16_CompareFeeling	Feeling After Comparison	1–5 Likert	MinMaxScaler (numeric)
6	Q17_Validation	Seek Validation Frequency	1–5 Likert	MinMaxScaler (numeric)
7	Q18_Depressed	Feel Depressed Frequency	1–5 Likert	MinMaxScaler (numeric)
8	Q19_Fluctuate	Interest Fluctuation Score	1–5 Likert	MinMaxScaler (numeric)
9	Q20_Sleep	Sleep Issues Score	1–5 Likert	MinMaxScaler (numeric)
10	Num_Platforms	Count of social media platforms	Integer	MinMaxScaler (numeric)
11	Age	Respondent age (years)	Integer	MinMaxScaler (numeric)
12	Daily_Time	Daily time on social media	6-level ordered	Ordinal Encoding + MinMaxScaler
13	Gender	Gender identity	Nominal (3 cats.)	One-Hot Encoding
14	Relationship_Status	Relationship status	Nominal	One-Hot Encoding
15	Occupation_Status	Occupation category	Nominal	One-Hot Encoding

Encoding and Pipeline Design

One-Hot Encoding was applied to nominal categorical variables (Gender, Relationship_Status, Occupation_Status) to avoid imposing artificial ordinal distances that would distort the geometry of SVM's feature space. Ordinal Encoding was applied to Daily_Time_on_Social_Media using the logically ordered sequence: Less than an Hour → Between 1 and 2 hours → ... → More than 5 hours. Numeric variables were scaled using MinMaxScaler, a normalization approach that has demonstrated consistent performance across ML classification tasks [24].

All preprocessing was implemented within a scikit-learn Pipeline using ColumnTransformer, ensuring that scaler and encoder parameters were fitted exclusively on training data and that the same parameters were applied to test data a strict anti-leakage procedure [25], [26].

Train–Test Split

A stratified 70:30 split (random_state=42) produced 334 training and 144 test samples, preserving class proportions in both subsets. The same split was applied identically to all models.

SVM Models

Three Support Vector Machine (SVM) classifiers were developed, namely a Linear SVM with C = 1.0, an RBF SVM with C = 1.0 and gamma = "scale", and a Polynomial SVM with degree = 3 and C = 1.0. For all models, probability estimation was enabled by setting probability = True. Hyperparameter optimization was conducted using GridSearchCV with stratified 5-fold cross-validation. The Linear

SVM was tuned over C values of {0.001, 0.01, 0.1, 1, 10, 100}. The RBF SVM was optimized using C values of {0.1, 1, 10, 100} and gamma values of {scale, auto, 0.01, 0.1}. Meanwhile, the Polynomial SVM was tuned using C values of {0.1, 1, 10}, polynomial degrees of {2, 3}, and gamma values of {scale, auto}.

Baseline Classifiers

Five baselines were implemented with identical preprocessing pipelines: Logistic Regression (max_iter=1000), Decision Tree (default), Random Forest (n_estimators=100), Gradient Boosting (default), and k-Nearest Neighbors (k=5). All used the same 70:30 split and evaluation metrics as SVM.

Evaluation Metrics

Models were evaluated on: (1) Accuracy; (2) Weighted F1-Score weighted by class support; (3) Macro F1-Score unweighted across classes, sensitive to minority class performance; (4) Balanced Accuracy mean per-class recall; (5) Confusion Matrix; (6) Permutation Importance (n_repeats=30) for the best-performing SVM [27], [28]. Macro F1 and balanced accuracy were prioritized for assessing fairness across the moderately imbalanced three-class problem.

Software

Python 3.10; scikit-learn 1.3.2 [25]; pandas 2.0.3; numpy 1.24.4; matplotlib 3.7.2; seaborn 0.12.2. All experiments run in Google Colab with random_state=42 for full reproducibility.

RESULTS AND DISCUSSION

Label Distribution

Figure 1 shows the distribution of constructed Attention Span labels. The High class (43.5%) is the majority, followed by Medium (31.8%) and Low (24.7%). This moderate class imbalance was handled through stratified splitting and evaluation using macro F1 and balanced accuracy alongside weighted metrics.

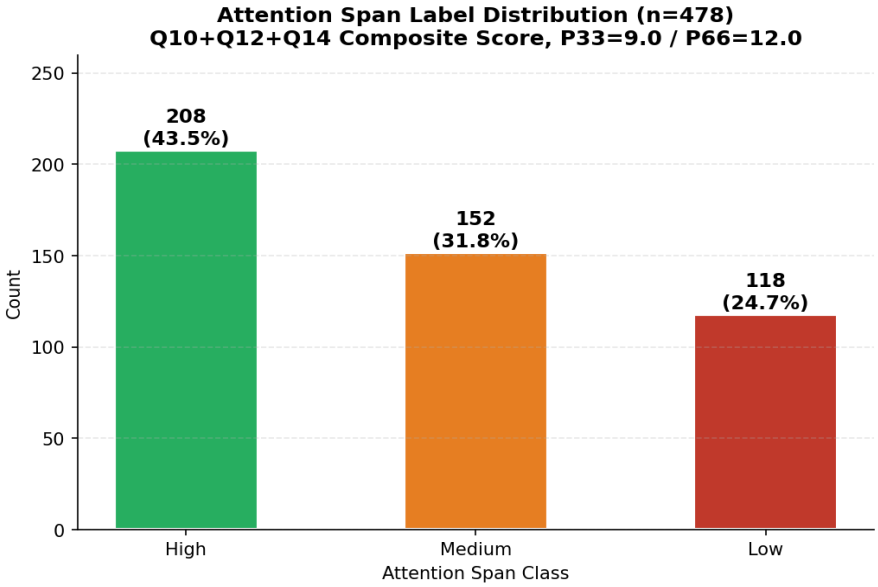

Figure 1. Attention Span label distribution (n=478). Constructed from Q10+Q12+Q14 composite score; P33=9.0, P66=12.0.

Leakage-Free SVM Performance at Default Parameters

Table 4 presents results with Q10, Q12, Q14 excluded. Linear SVM achieves the highest performance on all four metrics, making it the best SVM configuration in the leakage-free setting. This outcome directly contrasts with the original (contaminated) result where Polynomial SVM ranked first with 97.22% accuracy a consequence of label-feature dependency rather than genuine kernel superiority.

When the cognitive items that define the label are removed, Polynomial SVM drops to 55.56%, confirming that its original dominance was arithmetically expected rather than empirically informative.

Table 4. Leakage-Free SVM Performance Default Parameters

SVM Kernel	Accuracy	Weighted F1	Macro F1	Balanced Acc.
Linear (C=1.0)	61.81%	60.08%	58.99%	59.11%
RBF (C=1.0, γ=scale)	58.33%	56.40%	53.65%	53.24%
Polynomial (d=3, C=1.0)	55.56%	51.34%	48.62%	49.79%

The Linear kernel outperforming Polynomial in the leakage-free setting is consistent with, though not conclusive proof of, a relatively linear feature structure in the residual predictor space. This interpretation is plausible but dataset-specific, as no formal test of feature-space geometry was conducted. Importantly, the test set was held out entirely from model selection; all hyperparameter decisions were made using cross-validation on training data only, and the test set was evaluated once after all model decisions were finalized.

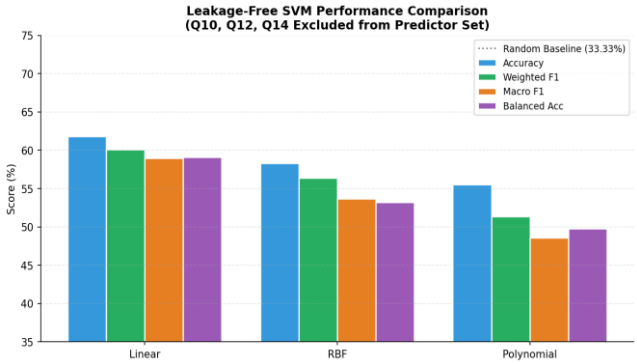

Figure 2. Leakage-free SVM performance comparison. All four metrics shown for each kernel.

Per-Class Performance and Confusion Matrices

Figure 3 shows confusion matrices for all three kernels. Table V reports per-class metrics for Linear SVM.

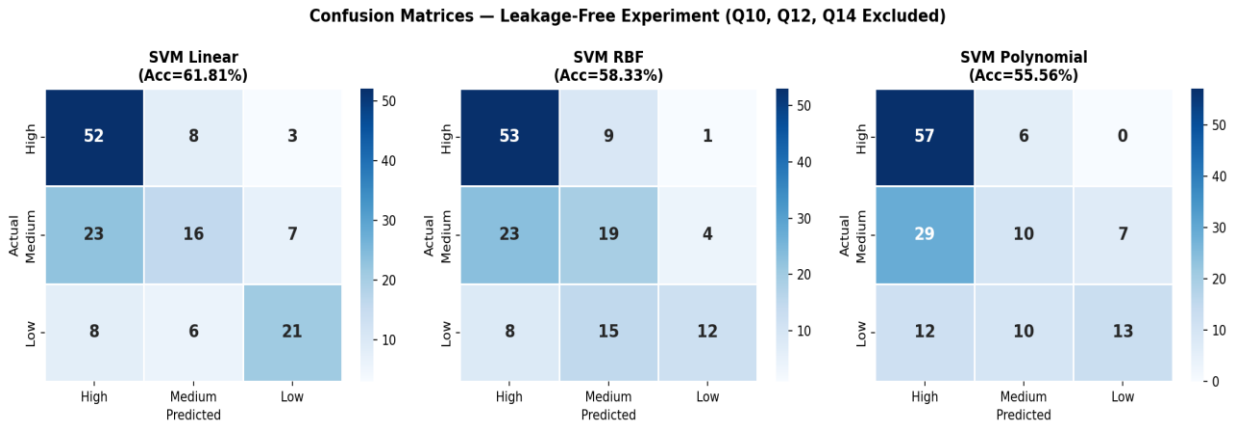

Figure 3. Confusion matrices for all three SVM kernels leakage-free experiment (Q10, Q12, Q14 excluded).

Table 5. Per-Class Metrics Linear SVM, Leakage-Free

Class	Precision	Recall	F1-Score	Support
High	0.63	0.83	0.71	63
Medium	0.53	0.35	0.42	46
Low	0.68	0.60	0.64	35
Macro average	0.61	0.59	0.59	144
Weighted average	0.61	0.62	0.60	144

The High class achieves the highest recall (0.83), consistent with its majority status and having the most training signal. The Medium class has the lowest F1 (0.42), which warrants explicit discussion. The Medium class occupies the middle percentile band of the composite score distribution; percentile-based thresholding places boundaries at the most ambiguous score regions, where adjacent classes share similar behavioral profiles. This reflects labeling-boundary ambiguity rather than model failure, and may be reduced by using validated scales with established cut-points in future work. The Low class achieves F1=0.64, suggesting that high-distraction behavioral patterns especially Q11 (Restless Without Social Media) are moderately separable.

5-Fold Cross-Validation

Table 6 and Figure 4 show stratified 5-fold CV results. All three kernels converge to essentially the same mean CV accuracy (≈57.1%), confirming no kernel holds a decisive advantage when label contamination is removed. The consistency between CV scores (≈57%) and test accuracy (56–62%) also confirms the absence of overfitting. However, it is important to acknowledge that 5-fold CV on a dataset of 478 records produces folds of approximately 96 samples each, which limits the precision of cross-validated estimates.

Table 6. 5-Fold Stratified Cross-Validation Leakage-Free

Kernel	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
Linear	51.04%	51.04%	64.58%	56.84%	62.11%	57.12%	±5.56%
RBF	56.25%	54.17%	64.58%	50.53%	60.00%	57.11%	±4.84%
Polynomial	54.17%	53.12%	61.46%	54.74%	62.11%	57.12%	±3.85%

The standard deviations of 3.85–5.56 percentage points indicate non-trivial fold-to-fold variability, meaning that differences of 1–2% between classifiers should not be interpreted as reliable performance gaps. Repeated stratified cross-validation with bootstrapped confidence intervals would be needed to establish whether the observed ranking of classifiers is stable across re-samplings of the data.

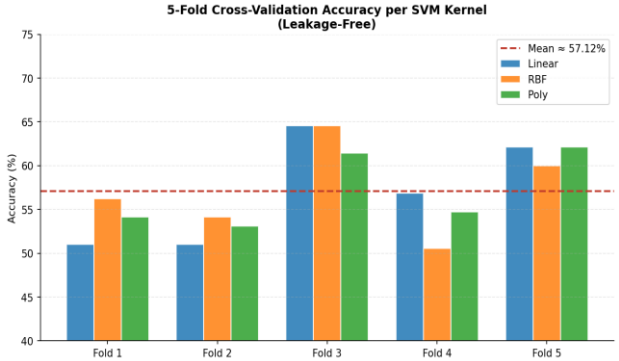

Figure 4. 5-Fold CV accuracy per kernel leakage-free experiment.

GridSearchCV Results

Table 7 shows the best hyperparameter configurations from GridSearchCV. Post-tuning improvements are modest or absent: tuned Linear SVM (58.33%) does not outperform default Linear SVM (61.81%) on the test set, and similar patterns hold for RBF and Polynomial kernels. This result has two interpretations. First, it confirms that the predictive ceiling of the leakage-free feature set cannot be materially raised through hyperparameter optimization alone the signal available in the remaining 15 features appears largely exhausted at default settings. Second, it suggests that the default parameters (C=1.0, gamma="scale") are already well-calibrated for this dataset size and feature dimensionality. Crucially, all hyperparameter selection was performed using cross-validation on the training set only; the test set was held out entirely and evaluated exactly once after all model decisions were finalized, ensuring that reported test performance reflects a clean, uncontaminated generalization estimate.

Table 7. GridSearchCV Results Leakage-Free

Kernel	Best Parameters	Test Accuracy	Weighted F1
Linear	C = 10	58.33%	56.29%
RBF	C = 10, gamma = auto	56.94%	55.47%
Polynomial	C = 10, degree = 2, gamma = auto	56.94%	54.06%

Baseline Comparison

Table 8 and Figure 5 present the full classifier comparison. Random Forest achieves the highest accuracy (63.19%) and weighted F1 (62.26%), modestly outperforming Linear SVM (61.81%) and Logistic Regression (61.11%). The performance differences across the top three classifiers are within 2 percentage points and lack confidence intervals. Given the test set size of 144 samples, a 2-percentage-point accuracy difference corresponds to approximately 3 observations — well within the margin of random sampling variability. No single model can therefore be declared decisively superior without uncertainty quantification. Decision Tree and k-NN are the weakest classifiers, falling below 52% accuracy, which is itself near the 33.3% random baseline when adjusted for class imbalance. The manuscript maintains its focus on multi-kernel SVM to examine how kernel choice interacts with the leakage-free feature space; Random Forest is the strongest baseline comparator. Linear SVM is competitive but not uniquely superior; its practical advantage lies in its interpretable geometric decision boundary and predictable behavior with small datasets. The finding that Random Forest marginally outperforms SVM correctly contextualizes SVM's competitive standing.

Table 8. Full Classifier Comparison Leakage-Free Experiment

Classifier	Accuracy	Weighted F1	Macro F1	Notes
Random Forest	63.19%	62.26%	61.17%	Best overall
SVM Linear (C=1.0)	61.81%	60.08%	58.99%	Best SVM
Logistic Regression	61.11%	59.56%	58.50%	Competitive
SVM RBF (C=1.0)	58.33%	56.40%	53.65%	
Gradient Boosting	55.56%	55.20%	54.26%	
SVM Polynomial (d=3)	55.56%	51.34%	48.62%	
Decision Tree	51.39%	51.23%	50.56%	Weak
k-NN (k=5)	52.08%	50.57%	48.19%	Weak

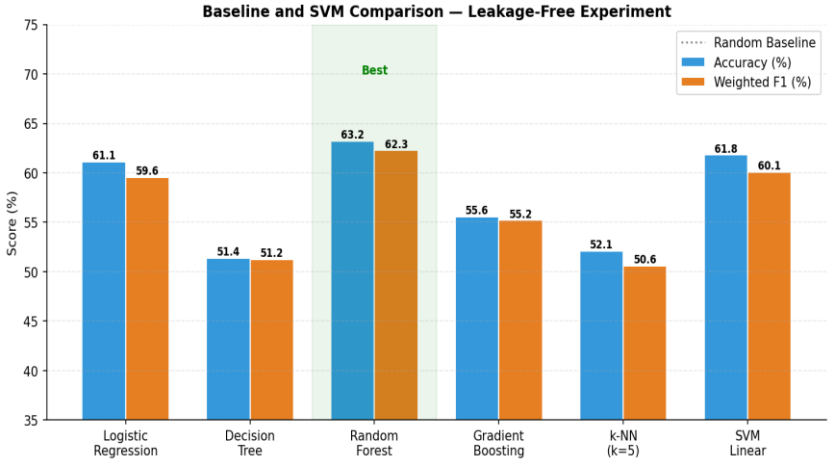

Figure 5. Full classifier comparison accuracy and weighted F1 for all models, leakage-free experiment.

Permutation Feature Importance

Permutation importance was computed for Linear SVM on the held-out test set (n_repeats=30). Figure 6 and Table 9 present results.

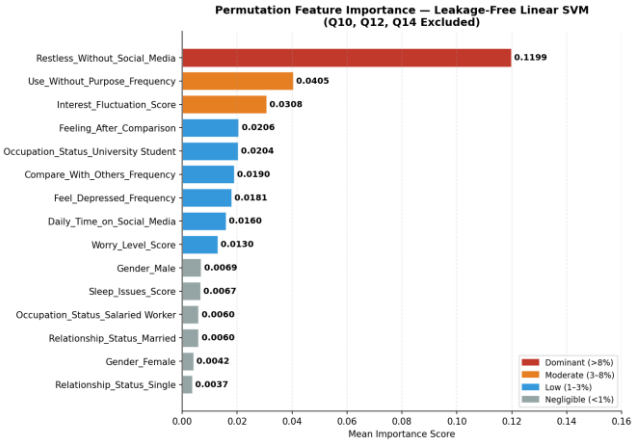

Figure 6. Permutation feature importance Linear SVM, leakage-free experiment. Top 15 features shown.

The dominant predictor in the leakage-free setting is Restless Without Social Media (Q11, importance=0.1199). This finding is theoretically plausible: behavioral restlessness when separated from social media may be associated with compulsive checking patterns that correlate with difficulty sustaining independent focus [29], [30], [31]. However, because the data are cross-sectional and self-reported, this association cannot be interpreted as a causal mechanism; Q11 is a possible behavioral correlate of higher composite distraction scores rather than a confirmed psychological driver. Importantly, Q11 is not part of the label formula; its high importance score represents dataset-specific predictive signal, not arithmetic reconstruction of the label.

Table 9. Top 10 Permutation Importance Linear SVM, Leakage-Free

Rank	Feature	Importance	Interpretation
1	*Q11_Restless_Without_SM*	0.1199	Restlessness when not using social media — highest leakage-free predictor
2	*Q9_Use_Without_Purpose*	0.0405	Habitual/purposeless scrolling behavior
3	*Q19_Interest_Fluctuation*	0.0308	Fluctuating motivation and interest levels

Rank	Feature	Importance	Interpretation
4	*Q16_Feeling_After_Comparison*	0.0206	Emotional state following upward social comparison
5	*Occupation: University Student*	0.0204	Student status — contextual behavioral pattern
6	*Q15_Compare_With_Others*	0.0190	Frequency of social comparison behavior
7	*Q18_Feel_Depressed*	0.0181	Self-reported frequency of depressive feelings
8	*Daily_Time_on_Social_Media*	0.0160	Ordinal daily usage duration
9	*Q13_Worry_Level*	0.0130	Self-reported worry/anxiety
10	*Num_Platforms*	0.0069	Number of platforms used

The second predictor, Use Without Purpose (Q9, 0.0405), has been associated with attentional distraction in prior work [32], [33], [34], [35]; the present study identifies it as a possible correlate of higher composite distraction scores, though causal claims are not warranted from cross-sectional survey data. Interest Fluctuation (Q19, 0.0308) may be associated with novelty-seeking tendencies that co-occur with heavy social media engagement; this interpretation is offered as a hypothesis for future research. The modest importance of Daily_Time_on_Social_Media (rank 8, 0.0160) suggests that total usage duration is a weaker predictor than qualitative behavioral indicators once label contamination is removed.

Discussion

Three levels of interpretation should be maintained. At the technical classification level: Linear SVM achieves 61.81% accuracy on a three-class problem with 33.3% random baseline a 71.7% relative improvement above chance, confirmed by CV (57.1%). This indicates that social media behavioral variables carry a modest, dataset-specific classification signal. The signal is real but limited; performance differences across classifiers are small and lack confidence intervals.

At the survey-score reconstruction level: the model classifies percentile groupings of Q10+Q12+Q14 composite scores. The predictive ceiling is partly determined by the psychometric reliability of these three items as a composite which has not been formally reported in this study and by the discriminative power of the remaining behavioral variables. The difficulty of the Medium class (F1=0.42) reflects genuine ambiguity at the score distribution boundary. A key limitation is that the SMMH dataset is a public Kaggle dataset with self-selected responses and unknown sampling procedures, limiting generalizability. The findings are therefore dataset-specific and require replication.

At the cognitive and clinical level: no inferences about neurological attention capacity are warranted. The practical value of these findings, if any, lies in identifying behavioral indicators restlessness (Q11), purposeless use (Q9), and interest fluctuation (Q19) that show a modest statistical association with higher composite distraction scores in this dataset. The current model should be treated as a preliminary analytical exploration, not a deployable screening tool. Any practical application would require: (1) replication with independently collected data; (2) validation of the three-item composite against standardized attention instruments; (3) formal reliability testing of the label; and (4) confidence intervals from repeated stratified CV or bootstrapping.

CONCLUSION

This study provides preliminary evidence that self-reported attention-related difficulty categories, derived from a survey composite score, can be classified at a modest but consistent level above chance using a leakage-controlled multi-classifier framework. This work should be treated as an exploratory analytical study rather than a validated predictive tool. After excluding Q10, Q12, and Q14 from the predictor set and implementing One-Hot Encoding within a full scikit-learn Pipeline,

Linear SVM achieved 61.81% test accuracy, 60.08% weighted F1, 58.99% macro F1, and 57.12% mean CV accuracy consistently above the 33.3% random baseline. Random Forest (63.19%) was the best overall classifier, marginally outperforming Linear SVM and Logistic Regression; performance differences across the top three classifiers were small (within 2 percentage points) and lack uncertainty intervals.

Permutation importance identified Restless Without Social Media (Q11, 0.1199), Use Without Purpose (Q9, 0.0405), and Interest Fluctuation (Q19, 0.0308) as the three dominant leakage-free predictors. These are possible behavioral correlates of the composite distraction score; causal or mechanistic interpretations should not be drawn from cross-sectional self-report data. All results should be interpreted as preliminary, dataset-specific classification of a survey-derived composite score. The label itself is a three-item construct whose internal reliability has not been formally confirmed. Results do not constitute clinically validated attention span assessment and should not be used for screening without extensive external validation.

Future directions include: (1) reporting internal consistency (Cronbach's alpha or McDonald's omega) for the Q10+Q12+Q14 composite; (2) using standardized attention instruments (e.g., Continuous Performance Test, Conners' Scales) as validated external labels; (3) applying SMOTE or class-weight adjustment for the minority class; (4) repeated stratified CV with bootstrapped confidence intervals; (5) ablation studies across predictor subsets; (6) longitudinal data collection to assess temporal causality; and (7) model evaluation on independently collected, geographically diverse datasets.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Dr. Delima Sitanggang, S.Kom., M.Kom. as the research supervisor, and to the Faculty of Science and Technology, Universitas Prima Indonesia, for the support and facilities provided throughout this research.

AUTHOR CONTRIBUTION STATEMENT

RP conceived the study, performed data preprocessing, developed the machine learning models, conducted the experiments, analyzed the results, and wrote the manuscript. DL contributed to methodology validation, data interpretation, and manuscript review. DNS contributed to machine learning process and the results validation. DS contributed to research supervision, conceptual guidance, and final manuscript evaluation.

AI DISCLOSURE STATEMENT

The Author used ClaudeAI and ChatGPT during the preparation of this work for language refinement, reviewing, and editing assistance. After using the tool/service, the author throughly reviewed and edited the content as needed and take full responsibility for the content of the publication. The authors declare that this research was prepared, researched, written, and edited without the aid of artificial intelligence (AI) techniques.

REFERENCES

[1] J. Firth *et al.*, "The 'online brain': how the Internet may be changing our cognition," *World Psychiatry*, vol. 18, no. 2, hal. 119–129, 2019, doi: 10.1002/wps.20617.

[2] T. Yan, C. Su, W. Xue, Y. Hu, dan H. Zhou, "Mobile phone short video use negatively impacts attention functions: An EEG study," *Front. Psychol.*, vol. 14, hal. 1106297, 2023, doi: 10.3389/fpsyg.2023.1106297.

[3] M. Shanmugasundaram, M. Tamilarasu, dan S. S. Mohan, "The impact of digital technology, social media, and artificial intelligence on cognitive functions: A review," *Front. Cogn.*, vol. 2, hal. 1203077, 2023, doi: 10.3389/fcogn.2023.1203077.

[4] V. S. Ramachandran, Ed., *Encyclopedia of Human Behavior*, 2nd ed., vol. 3. Amsterdam, Netherlands: Elsevier, 2012.

[5] G. Mark, D. Gudith, dan U. Klocke, "The cost of interrupted work: More speed and stress," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Florence, Italy, 2008, hal. 107–110. doi: 10.1145/1357054.1357072.

[6] L. D. Rosen, L. M. Carrier, dan N. A. Cheever, "Facebook and texting made me do it: Media-induced task-switching while studying," *Comput. Human Behav.*, vol. 29, no. 3, hal. 948–958, 2013, doi: 10.1016/j.chb.2012.12.001.

[7] N. S. Hawi dan M. Samaha, "The relations among social media addiction, self-esteem, and life satisfaction in university students," *Soc. Sci. Comput. Rev.*, vol. 35, no. 5, hal. 576–586, 2017, doi: 10.1177/0894439316660340.

[8] K. Rioja, S. Cekic, D. Bavelier, dan S. E. Baumgartner, "Unraveling the link between media multitasking and attention across three samples," *J. Commun.*, vol. 73, no. 5, hal. 427–439, 2023, doi: 10.1093/joc/jqad025.

[9] J. M. Twenge, G. N. Martin, dan W. K. Campbell, "Decreases in psychological well-being among American adolescents after 2012 and links to screen time," *Emotion*, vol. 18, no. 6, hal. 765–780, 2018, doi: 10.1037/emo0000403.

[10] J. M. Twenge, "More time on technology, less happiness? Associations between digital media use and psychological well-being," *Curr. Dir. Psychol. Sci.*, vol. 28, no. 4, hal. 372–379, 2019, doi: 10.1177/0963721419838244.

[11] L. Nguyen, A. T. Nguyen, M. M. McDonald, dan C. M. Burns, "Feeds, Feelings, and Focus: A Systematic Review and Meta-Analysis Examining the Cognitive and Mental Health Correlates of Social Media Use in Adolescents," *Adolescents*, vol. 4, no. 1, hal. 58–78, 2024, doi: 10.3390/adolescents4010005.

[12] N. K. Iyortsuun, S.-H. Kim, M. Jhon, H.-J. Yang, dan S. Pant, "A review of machine learning and deep learning approaches on mental health diagnosis," *Healthcare*, vol. 11, no. 3, hal. 285, 2023, doi: 10.3390/healthcare11030285.

[13] J. Kim, D. Lee, dan E. Park, "Machine learning for mental health in social media: Bibliometric study," *J. Med. Internet Res.*, vol. 23, no. 5, hal. e24870, 2021, doi: 10.2196/24870.

[14] C. Cortes dan V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, hal. 273–297, 1995, doi: 10.1007/BF00994018.

[15] B. Schölkopf dan A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.

[16] A. Golchha, M. Kuchibhotla, dan A. Mukherjee, "Mental health classifier using SVM," in *Lecture Notes in Computer Science*, vol. 15346, Cham, Switzerland: Springer, 2024, hal. 127–138. doi: 10.1007/978-981-96-0573-6_10.

[17] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Informatica*, vol. 31, no. 3, pp. 249–268, 2007.

[18] A. F. Ward, K. Duke, A. Gneezy, dan M. W. Bos, "Brain Drain: The mere presence of one's own smartphone reduces available cognitive capacity," *J. Assoc. Consum. Res.*, vol. 2, no. 2, hal. 140–154, 2017, doi: 10.1086/691462.

[19] T. Haliti-Sylaj dan A. Sadiku, "Impact of short reels on attention span and academic performance of undergraduate students," *Eur. J. Appl. Sci.*, vol. 11, no. 5, hal. 266–277, 2023, doi: 10.14738/aivp.115.15623.

[20] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[21] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, pp. 1–21, 2012, doi:

10.1145/2382577.2382579.

[22] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory*, 3rd ed. New York, NY, USA: McGraw-Hill, 1994.

[23] D. L. Streiner, "Starting at the beginning: An introduction to coefficient alpha and internal consistency," *J. Pers. Assess.*, vol. 80, no. 1, pp. 99–103, 2003, doi: 10.1207/S15327752JPA8001_18.

[24] S. Aksoy and R. M. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognit. Lett.*, vol. 22, no. 5, pp. 563–582, 2001, doi: 10.1016/S0167-8655(00)00112-4.

[25] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, hal. 2825–2830, 2011.

[26] L. Buitinck et al., "API design for machine learning software: Experiences from the scikit-learn project," in *Proc. ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[27] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010, doi: 10.1093/bioinformatics/btq134.

[28] A. Fisher, C. Rudin, dan F. Dominici, "All models are wrong, but many are useful," *J. Mach. Learn. Res.*, vol. 20, no. 177, hal. 1–81, 2019.

[29] T. E. Robinson dan K. C. Berridge, "Incentive-sensitization and addiction," *Addiction*, vol. 96, no. 1, hal. 103–114, 2001, doi: 10.1080/09652140020016996.

[30] B. T. Sharpe dan R. A. Spooner, "Dopamine-scrolling: A modern public health challenge requiring urgent attention," *Perspect. Public Health*, vol. 144, no. 6, hal. 315–316, 2024, doi: 10.1177/17579139241240564.

[31] C. Montag, B. Lachmann, M. Herrlich, dan K. Zweig, "Addictive features of social media/messenger platforms and freemium games," *Int. J. Environ. Res. Public Health*, vol. 16, no. 14, hal. 2612, 2019, doi: 10.3390/ijerph16142612.

[32] A. C. Drody, E. J. Pereira, dan D. Smilek, "A desire for distraction: Uncovering the rates of media multitasking during online research study tasks," *Attention, Perception, Psychophys.*, vol. 85, no. 7, hal. 2431–2450, 2023, doi: 10.3758/s13414-023-02721-5.

[33] J. Xie, Y. Wang, Z. Lin, dan X. Li, "The effect of short-form video addiction on undergraduates' academic procrastination: A moderated mediation model," *Front. Psychol.*, vol. 14, hal. 1028045, 2023, doi: 10.3389/fpsyg.2023.1028045.

[34] M. Wadsley dan N. Ihssen, "The predictive utility of reward-based motives underlying excessive and problematic social networking site use," *Psychol. Addict. Behav.*, vol. 36, no. 4, hal. 373–382, 2022, doi: 10.1037/adb0000693.

[35] Z. Yang, X. Asbury, dan M. Griffiths, "An investigation of problematic smartphone use among Chinese university students: Associations with academic anxiety, procrastination, self-regulation and subjective wellbeing," *Int. J. Ment. Health Addict.*, vol. 17, no. 3, hal. 596–614, 2019, doi: 10.1007/s11469-018-9973-8.