

LightGBM Prediction of Digital Self-Harm from Self-Criticism, Emotion Dysregulation, and Online Disinhibition

Deborah. Cohen¹, Rachel. Higgins², Ariel. Paredes de Jesús³, Javier. Galvez-Martos^{3*}

¹ Department of Psychology, Queen's University, Kingston, Canada

² Department of Counseling, Leadership and Research Methods, University of Arkansas, Graduate Education Building, Fayetteville, AR 72701, USA

³ Department of Psychology, Universidad de La Frontera, Temuco 4780000, Chile

* Corresponding author email address: javier.galvez@ufrontera.cl

Article Info

Article type:

Original Research

How to cite this article:

Cohen, D., Higgins, R., Paredes de Jesús, A., & McCalla, L. (2026). LightGBM Prediction of Digital Self-Harm from Self-Criticism, Emotion Dysregulation, and Online Disinhibition. *Journal of Assessment and Research in Applied Counseling*, 8(4), 1-15.
<http://dx.doi.org/10.61838/kman.jarac.5824>



© 2026 the authors. Published by KMAN Publication Inc. (KMANPUB), Ontario, Canada. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

ABSTRACT

Objective: This study aimed to predict digital self-harm severity among young adults in Chile using a LightGBM model based on self-criticism, emotion dysregulation, and online disinhibition.

Methods and Materials: The study used a cross-sectional, correlational, and predictive design. The statistical population consisted of young adults living in Chile who were active users of digital communication platforms. The final sample included 512 participants aged 18 to 29 years. Data were collected using a Digital Self-Harm Behavior Checklist, the Forms of Self-Criticizing/Attacking and Self-Reassuring Scale, the Difficulties in Emotion Regulation Scale, the Online Disinhibition Scale, and a demographic and digital behavior questionnaire. Data analysis included descriptive statistics, reliability assessment, Pearson correlation coefficients, and predictive modeling using Light Gradient Boosting Machine. The dataset was divided into training and test sets, and model performance was evaluated using cross-validation, root mean square error, mean absolute error, and coefficient of determination. SHAP values were used to interpret the relative contribution of predictors.

Findings: Digital self-harm was positively correlated with inadequate self ($r = 0.43$, $p < 0.01$), hated self ($r = 0.57$, $p < 0.01$), emotion dysregulation total score ($r = 0.52$, $p < 0.01$), benign online disinhibition ($r = 0.21$, $p < 0.01$), and toxic online disinhibition ($r = 0.61$, $p < 0.01$), and negatively correlated with reassured self ($r = -0.32$, $p < 0.01$). The LightGBM model showed the strongest predictive performance, with test RMSE = 3.96, MAE = 2.98, and $R^2 = 0.59$. SHAP analysis identified toxic online disinhibition, hated self, limited access to emotion regulation strategies, impulse control difficulties, and nonacceptance of emotional responses as the most important predictors of digital self-harm severity.

Conclusion: The findings indicate that digital self-harm is best understood as a multidimensional phenomenon shaped by hostile self-criticism, impaired emotion regulation, and toxic online disinhibition.

Keywords: Digital self-harm; LightGBM; self-criticism; emotion dysregulation; online disinhibition; machine learning; Chile.

1. Introduction

Digital environments have become central spaces for identity construction, emotional expression, social interaction, conflict, intimacy, and aggression. As social communication increasingly occurs through social networking platforms, instant messaging applications, online forums, gaming spaces, and anonymous or pseudonymous digital accounts, psychological distress and interpersonal hostility are no longer confined to offline settings. Online environments can intensify, transform, and sometimes conceal harmful behavior because they combine rapid communication, reduced social cues, perceived anonymity, audience visibility, and the possibility of delayed or limited accountability. Research on online deviance has shown that digital spaces may facilitate behaviors such as trolling, cyberstalking, body shaming, slut shaming, cyberhate, cyber-racism, online aggression, and other forms of cyber violence, indicating that the internet can function as both a site of social connection and a context for psychological risk (Pan et al., 2022; Scott & Barlett, 2023; Vale & Matos, 2023; Zhang et al., 2024; Zhao, 2023). Within this broader field, digital self-harm represents a particularly complex and underexamined phenomenon because the individual is not only exposed to harmful digital content but also participates in generating or directing harmful online content toward the self.

Digital self-harm can be understood as the deliberate creation, posting, sending, or circulation of negative, humiliating, hostile, or abusive online content about oneself, often through anonymous or pseudonymous accounts. Unlike conventional cyberbullying, where the perpetrator and victim are usually conceptualized as different persons, digital self-harm disrupts the usual perpetrator-victim distinction by placing self-directed aggression within the architecture of online communication. This phenomenon is theoretically important because it may express emotional pain, self-punishment, shame, anger, attention-seeking, interpersonal signaling, identity experimentation, or difficulty regulating distress. It also occupies an ambiguous position between self-injury, online aggression, and digital identity performance. Earlier discussions of the internet and the meaning of self suggest that online settings may alter the boundaries of self-expression, self-presentation, and regressive behavior, thereby making digital contexts psychologically relevant for understanding how individuals enact internal distress externally (Sharma et al., 2021). In this sense, digital self-harm is not simply an unusual online

behavior, but a psychologically meaningful expression of self-directed hostility enacted through technological affordances.

The growing literature on cyber aggression provides a necessary background for understanding digital self-harm. Studies on cyberbullying, workplace cyberbullying, cyber intimate partner aggression, cyberdating abuse, revenge pornography, virtual sexual assault, and bystander behavior have shown that harmful digital behavior is shaped by individual traits, relational conditions, contextual affordances, and social norms within online settings (Booth et al., 2023; Lee, 2026; Rehman, 2024; Sánchez-Hernández et al., 2024; Wanjiku, 2022; You, 2023). Although these studies primarily focus on aggression directed toward others, they are relevant to digital self-harm because both other-directed and self-directed online aggression may involve disinhibited expression, reduced inhibition, emotional arousal, moral disengagement, and distorted perceptions of interpersonal consequences. Furthermore, research on adolescent cyberbullying has emphasized the role of personal and situational factors, internal states, and family environments in explaining online harmful behavior (Wang & Ngai, 2021; Zheng et al., 2025). These findings suggest that digital self-harm should be studied through a multidimensional framework that incorporates individual psychological vulnerabilities and online behavioral tendencies rather than through a single-factor explanation.

One of the most important constructs for understanding harmful online behavior is online disinhibition. Online disinhibition refers to the tendency to behave, disclose, express emotions, or act in ways online that one might not enact in face-to-face contexts. This construct includes benign forms, such as openness and self-disclosure, and toxic forms, such as aggression, hostility, impulsivity, and norm violation. The adaptation and measurement of online disinhibition have supported its utility as a psychological construct for examining digital conduct across cultural contexts (Mantara et al., 2023). Empirical studies have connected online disinhibition to cyberbullying, trolling, cyberdating abuse, and aggression in online gaming, demonstrating that reduced restraint in digital settings can become a pathway toward harmful online expression (Ding et al., 2025; Gan et al., 2024; Maftei et al., 2023; Mao & Hu, 2024). In the context of digital self-harm, toxic online disinhibition may be especially relevant because it can lower the psychological barrier against posting hostile or humiliating statements, even when the target of those statements is the self.

Online disinhibition is not merely a technical consequence of anonymity; it is also shaped by emotion, motivation, personality, and social context. For example, studies on online aggression have identified roles for trait anger, revenge motivation, social dominance orientation, moral disengagement, and self-control, suggesting that disinhibition becomes particularly harmful when it interacts with hostile affective and cognitive tendencies (Ding et al., 2025; Gan et al., 2024; Nocera et al., 2022; Ramadhani & Merida, 2024). Research on anti-social behavior on social media similarly indicates that trolling and other hostile online behaviors among young adults are not random acts, but are embedded in psychological and social processes that shape how individuals interpret, justify, and perform digital conduct (Soares et al., 2023). In the case of digital self-harm, these mechanisms may operate inwardly: anger, shame, or punitive self-appraisal may be transformed into self-directed hostile content, while online disinhibition provides the behavioral channel through which the internal state is made visible.

Self-criticism is another central construct in understanding digital self-harm. Although digital self-harm occurs in an online context, its psychological origin may lie in hostile self-evaluation, feelings of inadequacy, self-blame, or a punitive attitude toward the self. Individuals with strong self-critical tendencies may interpret mistakes, rejection, failure, or interpersonal conflict as evidence of personal worthlessness, and such internalized hostility may increase vulnerability to self-directed harmful behaviors. The relevance of self-directed hostility is supported indirectly by work on psychopathy, suicide, aggression, and pathological personality domains, which shows that maladaptive personality features, affective instability, and dysregulated self-related processes can increase risk for severe behavioral and emotional outcomes (Bant et al., 2025; Hudon et al., 2025; Nazari et al., 2025; Rose et al., 2025). A transdiagnostic perspective also supports the idea that psychological symptoms and maladaptive behaviors often emerge from overlapping internal processes rather than isolated diagnostic categories (Carmichael et al., 2023). Accordingly, digital self-harm may be conceptualized as an online behavioral manifestation of broader self-directed distress and maladaptive self-regulation.

Emotion dysregulation provides a further explanatory pathway. Difficulties in emotion regulation may include nonacceptance of emotional responses, inability to maintain goal-directed behavior under distress, impulse control difficulties, limited access to effective regulation strategies,

lack of emotional awareness, and lack of emotional clarity. These difficulties can make individuals more likely to respond to distress through immediate, impulsive, or maladaptive behaviors, particularly in digital environments where posting, messaging, and account creation are rapidly available. Research on personality functioning has emphasized the mediating role of emotion dysregulation in the relationship between pathological traits and maladaptive psychological outcomes, suggesting that emotion regulation difficulties may translate dispositional vulnerabilities into harmful behaviors (Nazari et al., 2025). In related behavioral domains, self-regulation and pursued rewards have been identified as important for understanding gaming disorder symptoms and protective processes, indicating that regulation capacities are central to digital behavior more broadly (Achab et al., 2022). Therefore, digital self-harm may be particularly likely when self-critical thoughts and painful emotions occur alongside poor regulatory capacity.

The social and communicative structure of digital platforms may further amplify the risk of digital self-harm. Online environments allow users to create multiple identities, obscure authorship, seek reactions from others, and test social responses while maintaining partial psychological distance from the behavior. Studies on social media aggression, cyberhate, and online misinformation demonstrate that digital platforms can intensify hostility by creating conditions where harmful speech is rapidly amplified, socially reinforced, or detached from immediate interpersonal consequences (Giannakopoulos & Prassou, 2025; Lai, 2024). Research on psychiatric participation in online forums and online counseling also highlights the need to understand digital communication not only as a risk environment but also as a space where distress, disclosure, support-seeking, and intervention may occur (Agathokleous & Taiwo, 2022; Looi et al., 2023). This dual nature is important for digital self-harm because individuals who harm themselves online may simultaneously be expressing distress, seeking recognition, testing whether others care, or reproducing internal self-attacks through the language and mechanisms of digital interaction.

Emerging adulthood is a particularly relevant developmental period for examining digital self-harm because it is marked by identity exploration, heightened online engagement, affective reactivity, social comparison, intimate relationship formation, and increased autonomy in digital communication. Studies on risky sexting, cyber aggression, and adolescent or young adult online behavior show that risky digital conduct is often associated with

personality, relational, gendered, and contextual factors (Giannakopoulos & Prassou, 2025; Maftai et al., 2023; Morelli et al., 2025). In addition, research on cyberbullying perpetration and family factors demonstrates that online harmful behavior may be shaped by both individual vulnerabilities and relational environments (Wang & Ngai, 2021; Zheng et al., 2025). For young adults in Chile, where digital communication is deeply embedded in education, peer interaction, and social life, studying digital self-harm can contribute to a culturally relevant understanding of how internal psychological distress is translated into online behavior.

Despite the growing body of research on cyber aggression, online disinhibition, cyberbullying, and digital violence, important gaps remain. Much of the existing literature focuses on aggression toward others, whereas digital self-harm requires attention to the interaction between self-directed cognition and online behavior. In addition, many studies rely on linear or variable-centered models that may not fully capture the nonlinear and interactive nature of psychological risk. Digital self-harm is likely to emerge from a configuration of vulnerabilities rather than from a single predictor. For example, high self-criticism may be most risky when accompanied by emotion dysregulation and toxic online disinhibition; similarly, online disinhibition may be most harmful among individuals who experience intense self-directed hostility or limited access to regulation strategies. A machine learning approach is therefore appropriate because it can model complex patterns, interactions, and nonlinear associations among predictors. Light Gradient Boosting Machine modeling is especially useful for tabular psychological data because it can identify high-contribution predictors while preserving strong predictive performance.

The integration of self-criticism, emotion dysregulation, and online disinhibition into one predictive model provides a theoretically coherent framework for studying digital self-harm. Self-criticism reflects the internal evaluative and punitive dimension of risk; emotion dysregulation reflects the affective and self-regulatory dimension; and online disinhibition reflects the behavioral-contextual mechanism through which internal distress may be enacted digitally. These constructs also align with broader evidence that harmful online behavior is shaped by aggression, self-control, moral disengagement, personality traits, relational context, and platform-specific affordances (Hudon et al., 2025; Nocera et al., 2022; Ramadhani & Merida, 2024; Sánchez-Hernández et al., 2024). By applying a LightGBM

model, the present study moves beyond simple associations and examines the relative predictive contribution of each psychological and digital-behavioral factor. Such an approach can help identify which features are most informative for estimating digital self-harm severity and may guide future screening, prevention, and digitally delivered mental health interventions.

The aim of this study was to predict digital self-harm severity among young adults in Chile using a LightGBM model based on self-criticism, emotion dysregulation, and online disinhibition.

2. Methods and Materials

2.1. Study Design and Participants

The present study was conducted using a cross-sectional, correlational, and predictive design to examine whether digital self-harm could be predicted from self-criticism, emotion dysregulation, and online disinhibition through a Light Gradient Boosting Machine model. The target population consisted of young adults living in Chile who were active users of social networking platforms, messaging applications, online forums, or other interactive digital environments. The final sample included 512 participants from Chile, recruited from universities, technical institutes, community online groups, and public digital advertisements in Santiago, Valparaíso, Concepción, La Serena, and Temuco. Participants ranged in age from 18 to 29 years, with a mean age of 22.64 years and a standard deviation of 3.18. Inclusion criteria were being between 18 and 29 years of age, residing in Chile at the time of data collection, having used at least one social media or online communication platform during the previous six months, and providing informed consent to participate. Participants were excluded if they submitted incomplete questionnaires, failed attention-check items, reported inconsistent demographic responses, or completed the survey in an unrealistically short time that suggested non-attentive responding. Data were collected anonymously through an encrypted online survey platform. Before beginning the survey, participants received written information about the purpose of the study, the voluntary nature of participation, the approximate duration of completion, confidentiality procedures, and their right to withdraw without penalty. Because the study included questions related to self-directed harmful online behavior and psychological distress, participants were also provided with information about mental health support services and

emergency resources in Chile at the beginning and end of the questionnaire.

2.2. Measures

Digital self-harm was assessed using a Digital Self-Harm Behavior Checklist adapted from the operational definition proposed by Patchin and Hinduja, in which digital self-harm refers to anonymously posting, sending, or sharing hurtful, humiliating, hostile, or abusive content about oneself in online spaces. The checklist assessed whether participants had engaged in self-directed harmful online behaviors during the previous 12 months, including anonymously posting mean or degrading statements about themselves, creating or using anonymous accounts to insult themselves, sending hostile messages to themselves, encouraging others to criticize or attack them online, or staging online victimization directed at the self. Each item was rated on a five-point frequency scale ranging from 0 to 4, with 0 indicating never and 4 indicating very often. Higher total scores indicated greater severity of digital self-harm. For predictive modeling, the continuous total score was used as the main outcome variable, allowing the LightGBM model to estimate the level of digital self-harm severity rather than merely distinguishing between presence and absence. The checklist was translated into Spanish and culturally adapted for Chilean participants using forward translation, expert review, back-translation, and pilot testing with a small group of young adults to ensure semantic clarity, cultural appropriateness, and sensitivity of wording. In the present study, internal consistency was evaluated using Cronbach's alpha and McDonald's omega before inclusion of the score in the machine learning model.

Self-criticism was measured using the Forms of Self-Criticizing/Attacking and Self-Reassuring Scale developed by Gilbert, Clarke, Hempel, Miles, and Irons. This instrument assesses how individuals respond to personal failure, mistakes, or perceived inadequacy, and it is widely used to evaluate maladaptive self-critical processes and the ability to adopt a reassuring stance toward the self. The scale contains 22 items rated on a five-point Likert scale, ranging from low endorsement to high endorsement of each statement. It includes the subscales of Inadequate Self, which reflects feelings of inferiority, disappointment, and personal failure; Hated Self, which reflects more hostile, punitive, and self-attacking attitudes; and Reassured Self, which reflects the ability to be compassionate, supportive, and encouraging toward oneself during distress or setbacks.

In the present study, the Inadequate Self and Hated Self subscales were used as the primary indicators of self-criticism, while Reassured Self was retained as a theoretically relevant protective dimension and reverse-oriented feature in the predictive model. Higher scores on the self-critical dimensions indicated stronger self-critical tendencies. The Spanish version used in this study was reviewed for linguistic clarity and cultural fit in Chilean Spanish, and reliability coefficients were calculated for the total and subscale scores.

Emotion dysregulation was assessed using the Difficulties in Emotion Regulation Scale developed by Gratz and Roemer. The DERS is a 36-item self-report instrument designed to assess multiple dimensions of difficulty in understanding, accepting, and regulating emotional experiences. Items are rated on a five-point Likert scale, typically ranging from almost never to almost always, with higher scores indicating greater emotion regulation difficulties. The scale yields a total emotion dysregulation score and six subscale scores: Nonacceptance of Emotional Responses, Difficulties Engaging in Goal-Directed Behavior, Impulse Control Difficulties, Lack of Emotional Awareness, Limited Access to Emotion Regulation Strategies, and Lack of Emotional Clarity. These dimensions are especially relevant to digital self-harm because online self-directed aggression may occur in moments of emotional intensity, shame, impulsivity, or perceived inability to manage distress through adaptive strategies. In this study, both the DERS total score and its six subscale scores were considered as candidate predictors, allowing the model to determine whether global emotion dysregulation or specific regulatory deficits contributed more strongly to the prediction of digital self-harm. The Spanish-language form was checked for comprehensibility among Chilean young adults, and internal consistency estimates were computed for the total scale and each subscale.

Online disinhibition was measured using the Online Disinhibition Scale developed by Udris, which was designed to assess the extent to which individuals express themselves online in ways that may be less restrained than in face-to-face interaction. The scale evaluates both benign and toxic forms of online disinhibition. Benign online disinhibition refers to greater openness, emotional expression, and disclosure in digital spaces, whereas toxic online disinhibition refers to hostile, impulsive, aggressive, or norm-violating online behavior. Items are rated on a Likert-type response format, with higher scores representing stronger disinhibition in online contexts. The distinction

between benign and toxic online disinhibition was important for the present study because digital self-harm may involve both increased online self-disclosure and harmful self-directed aggression enacted through anonymous or pseudonymous digital behavior. The total score and subscale scores were entered as predictors in the machine learning analysis. The instrument was translated and adapted into Chilean Spanish through expert review and pilot testing, and reliability indices were examined before analysis. In addition to the main psychological instruments, a demographic and digital behavior form was used to collect information on age, educational status, region of residence, average daily internet use, primary online platforms, history of anonymous account use, and frequency of social media engagement during the previous six months.

2.3. Data analysis

Data analysis was performed in several sequential stages. First, the dataset was screened for incomplete responses, duplicate submissions, careless responding, extreme completion times, and inconsistent answers. Cases with excessive missing data were removed, while isolated missing values were handled using model-compatible imputation procedures based on the distributional characteristics of each variable. Descriptive statistics, including mean, standard deviation, minimum, maximum, skewness, and kurtosis, were calculated for all main study variables. Internal consistency of the psychological instruments was evaluated using Cronbach's alpha and McDonald's omega, and bivariate correlations were examined to provide an initial understanding of the associations among digital self-harm, self-criticism, emotion dysregulation, and online disinhibition. Because the main objective of the study was predictive rather than purely explanatory, traditional assumptions of linear regression were not treated as prerequisites for model estimation; however, distributional patterns and outliers were inspected to guide preprocessing decisions and to improve interpretability.

The main predictive analysis was conducted using the Light Gradient Boosting Machine algorithm. LightGBM was selected because it is an efficient gradient-boosted decision-tree framework that can model nonlinear relationships, high-order interactions, and complex predictor patterns in tabular psychological data. The outcome variable was the continuous digital self-harm severity score. Predictor variables included the self-criticism indicators,

emotion dysregulation total and subscale scores, online disinhibition total and subscale scores, and selected digital behavior covariates such as daily internet use and anonymous account use. Before model training, continuous predictors were checked for implausible values, categorical variables were encoded appropriately, and the dataset was divided into training and test sets using an 80:20 split. The training set was used for model development and hyperparameter tuning, while the test set was held out for final evaluation. To reduce overfitting and improve generalizability, five-fold cross-validation was applied within the training data. Hyperparameters, including the number of estimators, learning rate, maximum tree depth, number of leaves, minimum child samples, feature fraction, bagging fraction, and regularization parameters, were tuned using cross-validated performance. Early stopping was used when model performance no longer improved across validation iterations.

Model performance was evaluated using root mean square error, mean absolute error, and the coefficient of determination. These indices were selected because the primary outcome was treated as a continuous severity score, and together they provided information about average prediction error, sensitivity to larger errors, and the proportion of variance in digital self-harm severity explained by the model. To improve interpretability, feature importance values were extracted from the final LightGBM model, and SHAP values were calculated to examine the relative contribution and direction of influence of self-criticism, emotion dysregulation, and online disinhibition variables. SHAP summary and dependence analyses were used to identify the strongest predictors and to clarify whether specific psychological dimensions, such as hated self, impulse control difficulties, limited access to emotion regulation strategies, or toxic online disinhibition, had particularly meaningful roles in predicting digital self-harm. The final interpretation emphasized both predictive performance and psychological plausibility, ensuring that the machine learning findings were discussed in relation to established theory on self-directed hostility, emotion regulation difficulties, and disinhibited online behavior.

3. Findings and Results

The final analysis was conducted on data obtained from 512 young adults residing in Chile. The participants' ages ranged from 18 to 29 years, with a mean age of 22.64 years and a standard deviation of 3.18. Of the total sample, 291

participants were women, representing 56.84% of the sample, 208 were men, representing 40.63%, and 13 participants identified as non-binary or preferred not to report gender, representing 2.54%. Regarding educational status, 327 participants were undergraduate university students, 91 were students in technical or professional institutes, 54 were graduate students, and 40 were not currently enrolled in formal education but met the age and internet-use criteria for participation. In terms of geographical distribution, the largest proportion of participants lived in Santiago, followed by Valparaíso, Concepción, La Serena, and Temuco. The mean reported daily internet use was 6.47 hours, with a standard deviation

of 2.31, indicating a high level of routine digital engagement among participants. Social networking platforms, instant messaging applications, and short-video platforms were the most frequently used online environments. In addition, 188 participants, corresponding to 36.72% of the sample, reported that they had used an anonymous or pseudonymous online account during the previous six months. This demographic and digital profile indicated that the sample was appropriate for examining digital self-harm in relation to self-criticism, emotion dysregulation, and online disinhibition, because the participants were active users of online communication environments in which anonymous, self-directed, or disinhibited behavior could occur.

Table 1

Descriptive Statistics and Reliability Indices of the Main Study Variables

Variable	Possible Range	Mean	Standard Deviation	Skewness	Kurtosis	Cronbach's Alpha	McDonald's Omega
Digital self-harm	0–32	8.76	6.14	0.91	0.64	0.88	0.89
Inadequate self	0–36	18.42	7.31	0.28	-0.41	0.86	0.87
Hated self	0–20	7.94	5.22	0.77	0.19	0.89	0.90
Reassured self	0–32	17.86	6.48	-0.22	-0.36	0.84	0.85
Emotion dysregulation total	36–180	104.73	24.96	0.34	-0.18	0.93	0.94
Nonacceptance of emotional responses	6–30	17.91	5.87	0.31	-0.47	0.85	0.86
Difficulties in goal-directed behavior	5–25	14.83	4.72	0.22	-0.39	0.81	0.82
Impulse control difficulties	6–30	16.76	5.94	0.46	-0.11	0.87	0.88
Lack of emotional awareness	6–30	16.11	5.36	0.18	-0.52	0.79	0.80
Limited access to regulation strategies	8–40	24.62	7.18	0.37	-0.26	0.88	0.89
Lack of emotional clarity	5–25	14.50	4.69	0.29	-0.31	0.82	0.83
Online disinhibition total	11–55	31.28	8.42	0.16	-0.44	0.86	0.87
Benign online disinhibition	5–25	15.63	4.51	-0.08	-0.38	0.78	0.79
Toxic online disinhibition	6–30	15.65	5.39	0.43	-0.14	0.84	0.85

The descriptive findings presented in Table 1 indicate that the participants showed a moderate level of digital self-harm, with a mean score of 8.76 out of a possible range of 0 to 32. The positive skewness value for digital self-harm suggests that although many participants reported low to moderate levels of this behavior, a meaningful subgroup reported elevated engagement in self-directed harmful online behaviors. Among the self-criticism dimensions, inadequate self had a higher mean score than hated self, indicating that feelings of personal inadequacy, disappointment, and failure were more common than explicitly hostile or punitive attitudes toward the self. Reassured self showed a moderate mean score, suggesting that although many participants retained some capacity for self-support, this protective tendency coexisted with self-

critical tendencies in the sample. The total emotion dysregulation score was also moderate to high, and the highest mean among its subscales was observed for limited access to regulation strategies, followed by nonacceptance of emotional responses and impulse control difficulties. These results suggest that difficulty identifying effective strategies for managing emotional distress was a particularly salient emotional regulation problem among the participants. Online disinhibition scores showed that both benign and toxic forms of disinhibition were present, although toxic online disinhibition was more theoretically relevant to the prediction of digital self-harm. The reliability coefficients were satisfactory to excellent across all instruments and subscales, with Cronbach's alpha values ranging from 0.78 to 0.93 and McDonald's omega values ranging from 0.79 to

0.94. These results indicate that the measures demonstrated adequate internal consistency and were suitable for subsequent correlational and machine learning analyses.

Table 2

Bivariate Correlations Between Digital Self-Harm, Self-Criticism, Emotion Dysregulation, and Online Disinhibition

Variable	1	2	3	4	5	6	7
1. Digital self-harm	1.00						
2. Inadequate self	0.43**	1.00					
3. Hated self	0.57**	0.61**	1.00				
4. Reassured self	-0.32**	-0.48**	-0.44**	1.00			
5. Emotion dysregulation total	0.52**	0.56**	0.59**	-0.41**	1.00		
6. Benign online disinhibition	0.21**	0.18**	0.24**	-0.09*	0.26**	1.00	
7. Toxic online disinhibition	0.61**	0.45**	0.58**	-0.29**	0.55**	0.39**	1.00

The correlation analysis shown in Table 2 demonstrated that digital self-harm was significantly associated with all main psychological predictors. The strongest bivariate association was observed between digital self-harm and toxic online disinhibition, $r = 0.61$, $p < 0.01$, indicating that participants who reported more hostile, impulsive, or norm-violating behavior in online environments were also more likely to report higher levels of self-directed harmful online behavior. Digital self-harm was also strongly and positively correlated with hated self, $r = 0.57$, $p < 0.01$, and emotion dysregulation total, $r = 0.52$, $p < 0.01$. These findings suggest that digital self-harm was closely related to both self-attacking attitudes and difficulties in regulating emotional responses. Inadequate self also showed a moderate positive association with digital self-harm, $r = 0.43$, $p < 0.01$, indicating that broader feelings of inferiority and personal

failure were relevant, although less strongly than explicitly hostile self-criticism. Reassured self was negatively correlated with digital self-harm, $r = -0.32$, $p < 0.01$, suggesting that the ability to respond to oneself with support and reassurance may function as a protective factor against self-directed harmful online behavior. Benign online disinhibition showed a weaker but statistically significant positive correlation with digital self-harm, $r = 0.21$, $p < 0.01$. This finding suggests that general openness and disclosure in digital spaces may be related to digital self-harm, but the toxic component of online disinhibition was substantially more important. Overall, the correlational pattern supported the theoretical assumption that digital self-harm is not merely a function of internet use, but is more closely connected to self-critical cognition, emotional dysregulation, and disinhibited online behavior.

Table 3

Predictive Performance of the LightGBM Model for Digital Self-Harm Severity

Model	Cross-Validation RMSE	Cross-Validation MAE	Cross-Validation R ²	Test RMSE	Test MAE	Test R ²
Linear regression	4.91	3.76	0.36	5.04	3.88	0.34
Ridge regression	4.84	3.69	0.38	4.96	3.80	0.36
Random forest regression	4.31	3.21	0.50	4.46	3.34	0.48
Support vector regression	4.45	3.35	0.47	4.58	3.46	0.45
LightGBM regression	3.82	2.87	0.61	3.96	2.98	0.59

The model comparison presented in Table 3 showed that the LightGBM regression model provided the strongest predictive performance for digital self-harm severity. In the cross-validation phase, LightGBM achieved an RMSE of 3.82, MAE of 2.87, and R² of 0.61, indicating that the model explained approximately 61% of the variance in digital self-harm scores within the training-validation procedure. When evaluated on the independent test set, the model maintained

strong generalization performance, with an RMSE of 3.96, MAE of 2.98, and R² of 0.59. The relatively small difference between cross-validation and test performance suggests that the model did not substantially overfit the training data. Compared with traditional linear regression and ridge regression, LightGBM produced a considerably lower prediction error and higher explained variance, indicating that nonlinear associations and interaction effects were

important in predicting digital self-harm. LightGBM also outperformed random forest regression and support vector regression, although the difference between tree-based models was smaller than the difference between LightGBM and linear models. These findings indicate that the relationship between self-criticism, emotion dysregulation, online disinhibition, and digital self-harm was not

adequately captured by purely linear models. Instead, the results support the use of gradient-boosted decision trees for modeling complex psychological risk patterns, especially when predictors may operate through threshold effects, cumulative vulnerability, or interactions among self-critical cognition, affective instability, and online behavioral disinhibition.

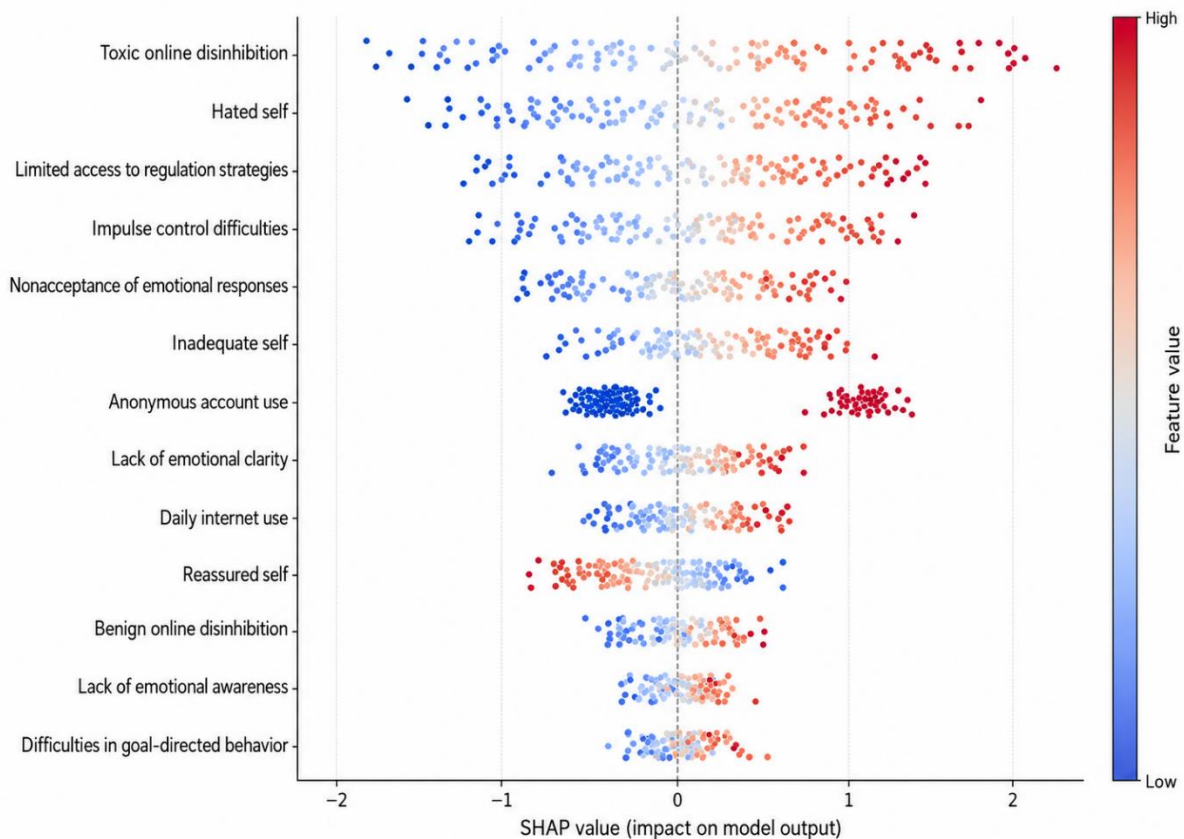
Table 4

Relative Importance of Predictors in the Final LightGBM Model

Predictor	Mean Absolute SHAP Value	Relative Importance Rank	Direction of Association With Digital Self-Harm
Toxic online disinhibition	1.42	1	Positive
Hated self	1.18	2	Positive
Limited access to emotion regulation strategies	0.97	3	Positive
Impulse control difficulties	0.86	4	Positive
Nonacceptance of emotional responses	0.74	5	Positive
Inadequate self	0.63	6	Positive
Anonymous account use	0.51	7	Positive
Lack of emotional clarity	0.46	8	Positive
Daily internet use	0.38	9	Positive
Reassured self	0.34	10	Negative
Benign online disinhibition	0.29	11	Positive
Lack of emotional awareness	0.22	12	Positive
Difficulties in goal-directed behavior	0.19	13	Positive

The feature importance results presented in Table 4 provide a more detailed explanation of the final LightGBM model. Toxic online disinhibition emerged as the most influential predictor of digital self-harm severity, with the highest mean absolute SHAP value. This indicates that online tendencies toward hostility, impulsivity, aggression, or norm violation made the largest contribution to the model's predictions. Hated self was the second most important predictor, suggesting that explicitly punitive and self-attacking attitudes were more predictive of digital self-harm than broader feelings of inadequacy alone. Among the emotion dysregulation dimensions, limited access to emotion regulation strategies and impulse control difficulties were especially important, ranking third and fourth in the model. This pattern indicates that participants who perceived themselves as lacking effective ways to regulate distress, and those who had difficulty controlling behavior under emotional activation, were more likely to

receive higher predicted digital self-harm scores. Nonacceptance of emotional responses also contributed meaningfully to prediction, suggesting that rejecting or judging one's emotional experiences may intensify vulnerability to self-directed harmful online behavior. Anonymous account use was also retained as a relevant behavioral predictor, which is theoretically consistent with the role of anonymity in enabling digital self-harm. Reassured self showed a negative direction of association, indicating that the ability to respond to oneself with warmth, reassurance, and emotional support reduced predicted digital self-harm severity. Benign online disinhibition and daily internet use contributed less strongly than toxic disinhibition, which suggests that digital self-harm was not primarily explained by high internet exposure or online openness alone, but by the combination of hostile self-directed cognition, impaired emotion regulation, and toxic disinhibited behavior in online environments.

Figure 1*SHAP Summary Plot of the LightGBM Predictors of Digital Self-Harm Severity*

The SHAP summary plot further clarified the direction and magnitude of the predictors included in the LightGBM model. The figure showed that higher values of toxic online disinhibition were consistently associated with higher predicted digital self-harm scores, confirming its role as the most important risk factor in the model. Similarly, higher hated self scores shifted predictions upward, indicating that participants with stronger self-attacking tendencies were more likely to be classified by the model as having elevated digital self-harm severity. The emotion dysregulation variables showed a nonlinear but generally positive pattern: limited access to regulation strategies, impulse control difficulties, and nonacceptance of emotional responses increased predicted risk most strongly when scores were in the upper range. This suggests that emotion regulation problems may become especially predictive when they reach clinically meaningful or functionally disruptive levels. Anonymous account use also increased predicted digital self-harm, but its contribution was smaller than that of the core psychological variables. In contrast, higher reassured self scores shifted predictions downward, indicating a

protective effect in the model. Overall, the SHAP interpretation demonstrated that the LightGBM model did not rely on a single predictor, but integrated a broader vulnerability profile in which toxic online behavior, hostile self-criticism, and difficulty regulating distress jointly contributed to higher digital self-harm severity.

4. Discussion

The present study examined the prediction of digital self-harm severity among young adults in Chile using a LightGBM model based on self-criticism, emotion dysregulation, and online disinhibition. The findings showed that digital self-harm was not a random or isolated digital behavior, but was meaningfully associated with a broader psychological vulnerability profile. In the descriptive results, participants reported a moderate level of digital self-harm, with variability indicating that while many participants showed low or occasional engagement in self-directed harmful online behaviors, a clinically and socially relevant subgroup reported higher severity. The correlational findings further indicated that digital self-harm was

positively associated with inadequate self, hated self, emotion dysregulation, benign online disinhibition, and toxic online disinhibition, while it was negatively associated with reassured self. These findings suggest that individuals who experience stronger self-attacking thoughts, greater emotional regulation difficulties, and greater disinhibition in online spaces are more vulnerable to engaging in online behaviors that reproduce harm toward the self. This pattern is consistent with the broader literature showing that harmful digital conduct is often shaped by the interaction of internal psychological states and online affordances, rather than by internet exposure alone (Wang & Ngai, 2021; Zhang et al., 2024; Zhao, 2023).

The predictive findings provided stronger evidence for the integrated role of these constructs. The LightGBM model outperformed linear regression, ridge regression, random forest regression, and support vector regression, achieving the strongest test performance, with an explained variance of 59% in digital self-harm severity. This result indicates that the predictors contributed to digital self-harm in a complex and probably nonlinear manner. In psychological terms, digital self-harm may emerge when multiple risk conditions converge: a person experiences intense self-critical or self-hating cognitions, has difficulty regulating painful affect, and enters an online context where anonymity or disinhibition reduces behavioral restraint. This interpretation is aligned with transdiagnostic perspectives suggesting that maladaptive behaviors often arise from interacting psychological vulnerabilities rather than from single isolated symptoms (Carmichael et al., 2023; Nazari et al., 2025). It is also consistent with cyberbehavioral studies showing that online aggression, trolling, cyberbullying, and other forms of online deviance are best understood through the combined effects of personality, affective states, self-control, relational context, and digital affordances (Gan et al., 2024; Giannakopoulos & Prassou, 2025; Maftai et al., 2023; Soares et al., 2023).

The most important predictor in the final model was toxic online disinhibition. This finding suggests that digital self-harm is strongly connected to the online tendency to express hostility, impulsivity, aggression, or norm-violating behavior under conditions of reduced restraint. Although toxic online disinhibition has typically been examined in relation to harm directed toward others, the present findings indicate that the same psychological mechanism may also facilitate harm directed toward the self. In this regard, digital self-harm can be interpreted as an inwardly directed version of harmful online expression: the hostile language,

anonymity, and performative features commonly seen in cyber aggression are applied to one's own identity. This explanation is supported by studies showing that online disinhibition contributes to cyberbullying perpetration, reactive trolling, online gaming aggression, cyberdating abuse, and toxic digital behavior (Ding et al., 2025; Mao & Hu, 2024; Ramadhani & Merida, 2024; Sánchez-Hernández et al., 2024). The findings also align with the adaptation and measurement literature on online disinhibition, which emphasizes that disinhibition is not a unitary phenomenon but may include both benign and toxic components (Mantara et al., 2023). In the present study, benign online disinhibition had a weaker predictive contribution, whereas toxic online disinhibition emerged as the dominant online-contextual risk factor. This distinction is important because it shows that digital self-harm is not primarily explained by openness or online self-disclosure, but by the hostile and impulsive form of disinhibited digital behavior.

Hated self was the second strongest predictor of digital self-harm severity, indicating that explicitly punitive, hostile, and self-attacking attitudes toward the self were central to the model. This finding supports the conceptualization of digital self-harm as a digital enactment of self-directed aggression. Individuals high in hated self may be more likely to believe they deserve criticism, humiliation, or punishment, and online platforms may provide an accessible medium through which these internal attacks are externalized. In this sense, anonymous or pseudonymous self-abusive posting may function as a behavioral extension of self-critical cognition. The stronger role of hated self compared with inadequate self suggests that digital self-harm is especially associated with hostile self-rejection rather than general feelings of inferiority alone. This distinction is compatible with evidence that severe maladaptive outcomes are often linked to more antagonistic, dysregulated, or self-destructive psychological traits (Bant et al., 2025; Hudon et al., 2025; Rose et al., 2025). It also resonates with work on dark personality, anger, and cyber aggression, which indicates that hostility and punitive affect can be translated into harmful digital behavior when regulatory or moral restraints are weakened (Nocera et al., 2022). In the present context, however, the hostility appears to be directed inward, making digital self-harm a distinctive phenomenon at the intersection of self-criticism and online aggression.

Emotion dysregulation also played a major role in the prediction of digital self-harm. Among its components, limited access to emotion regulation strategies, impulse

control difficulties, and nonacceptance of emotional responses were especially important. This pattern suggests that digital self-harm may be most likely when individuals experience distress as intolerable, lack confidence in their ability to regulate emotional states, and have difficulty inhibiting behavior during emotional arousal. Such individuals may turn to digital self-harm as an immediate, maladaptive strategy for expressing distress, punishing the self, seeking social response, or externalizing internal emotional chaos. The prominence of limited access to regulation strategies is particularly meaningful because it indicates that digital self-harm may not only reflect the presence of negative affect, but also the perceived absence of adaptive alternatives. This interpretation is consistent with evidence that emotion dysregulation mediates the relationship between pathological traits and maladaptive functioning (Nazari et al., 2025). It is also consistent with findings from gaming disorder research showing that self-regulation capacities and pursued rewards are central to understanding maladaptive digital behavior (Achab et al., 2022). Accordingly, digital self-harm may be understood as a dysregulated digital response to affective distress, particularly when individuals do not possess or cannot access more adaptive coping strategies.

The role of anonymous account use further supports the importance of digital affordances. Participants who reported anonymous or pseudonymous account use had higher predicted digital self-harm severity, although this variable was less influential than toxic online disinhibition and self-critical or emotional variables. This suggests that anonymity may function as an enabling condition rather than a sufficient cause. In other words, anonymity may reduce accountability and increase behavioral distance from the act, but it is most likely to result in digital self-harm when combined with self-directed hostility and emotional dysregulation. This interpretation is consistent with studies on online deviance and cyber violence, which show that the structure of online communication can facilitate harmful behavior by reducing inhibition, increasing distance from consequences, and allowing rapid performance of aggression (Pan et al., 2022; Scott & Barlett, 2023; Vale & Matos, 2023). It is also aligned with literature on cyber-racism, cyberhate, trolling, and online abuse, where anonymity and perceived distance may strengthen the expression of hostile content (Lai, 2024; Soares et al., 2023; Zhang et al., 2024). In digital self-harm, these mechanisms may operate paradoxically: the individual becomes both the producer and target of harmful content, using the online

environment to stage, amplify, or validate self-directed negativity.

The negative association between reassured self and digital self-harm provides an important protective interpretation. Participants who reported a greater capacity to respond to themselves with warmth, reassurance, and emotional support were less likely to receive high predicted digital self-harm scores. This finding supports the idea that self-reassurance may buffer the transition from emotional distress to self-directed online harm. When individuals can respond to mistakes, rejection, shame, or distress with a supportive internal voice, they may be less likely to convert negative emotions into punitive digital behaviors. This is important because the study did not only identify risk factors; it also identified a protective psychological dimension that may be relevant for prevention. The protective role of self-reassurance is compatible with wider evidence that online aggression and harmful digital behaviors are shaped by self-control, internal states, and contextual influences (Ramadhani & Merida, 2024; Wang & Ngai, 2021). It also complements research on harmful relational and sexualized digital behaviors, such as cyber intimate partner aggression, cyberdating abuse, aggravated sexting, and revenge pornography, where poor regulation, insecure relational patterns, and disinhibition contribute to online harm (Booth et al., 2023; Morelli et al., 2025; Sánchez-Hernández et al., 2024; Wanjiku, 2022). In the present study, self-reassurance may represent an internal regulatory resource that reduces the need to enact distress through digital self-attack.

The findings also have implications for understanding digital self-harm as part of a broader ecology of online risk. Prior research has shown that cyberbullying, cyberhate, online aggression, bystander inaction, virtual sexual assault, and workplace cyberbullying are shaped by both individual and social mechanisms (Giannakopoulos & Prassou, 2025; Lee, 2026; Rehman, 2024; You, 2023). The present study extends this logic to self-directed online harm by showing that digital self-harm is embedded in the same digital ecology but follows a distinctive psychological pathway. Rather than treating digital self-harm as merely attention-seeking or as a technological anomaly, the findings suggest that it should be understood as a meaningful expression of self-directed distress within online environments that permit anonymity, rapid emotional expression, and social feedback. The study also connects with work on online counseling and psychiatric participation in digital forums, which emphasizes that online spaces are not only sites of risk but

also potential contexts for identification, support, and intervention (Agathokleous & Taiwo, 2022; Looi et al., 2023). Therefore, identifying digital self-harm risk through psychological and behavioral predictors may support earlier detection and more responsive digital mental health strategies.

5. Conclusion

Overall, the results support a multidimensional model in which digital self-harm is predicted by the interaction of self-critical cognition, affective dysregulation, and toxic online behavioral disinhibition. Toxic online disinhibition was the strongest contextual-behavioral predictor, hated self was the strongest self-evaluative predictor, and limited access to emotion regulation strategies and impulse control difficulties were the strongest emotional-regulatory predictors. These findings suggest that digital self-harm among young adults in Chile may be most severe when hostile self-attitudes are combined with poor emotional regulation and a disinhibited online style. The findings are consistent with previous research on online aggression, cyberbullying, cyber violence, trolling, and harmful digital behavior, while also extending this literature by focusing on self-directed rather than other-directed harm (Ding et al., 2025; Gan et al., 2024; Maftei et al., 2023; Mao & Hu, 2024; Zheng et al., 2025). The use of a LightGBM model added further value by demonstrating that machine learning can capture complex risk patterns and identify the relative contribution of psychological predictors in a way that is useful for both research and prevention.

6. Limitations & Suggestions

The present study had several limitations. First, the cross-sectional design does not allow causal conclusions about whether self-criticism, emotion dysregulation, and online disinhibition lead to digital self-harm or whether digital self-harm intensifies these psychological vulnerabilities over time. Second, all variables were assessed through self-report instruments, which may be affected by recall bias, social desirability, underreporting, or misunderstanding of sensitive online behaviors. Third, although the sample included young adults from several regions of Chile, it may not fully represent all Chilean young adults, particularly those with limited internet access, lower digital literacy, or different cultural patterns of online communication. Fourth, the use of an adapted digital self-harm checklist may require further psychometric validation in Chilean Spanish,

including factorial validity, criterion validity, and measurement invariance across gender and age groups. Fifth, although the LightGBM model showed strong predictive performance, machine learning models are dependent on the quality and scope of the predictors included; therefore, important variables such as depression, shame, loneliness, trauma history, peer victimization, family conflict, and offline self-harm may have been omitted.

Future studies should use longitudinal designs to clarify temporal and causal pathways between self-criticism, emotion dysregulation, online disinhibition, and digital self-harm. Research should also examine whether digital self-harm predicts later offline self-harm, suicidal ideation, cybervictimization, social withdrawal, or help-seeking behavior. Future studies could expand the predictor set by including depressive symptoms, anxiety, shame, perceived burdensomeness, loneliness, rejection sensitivity, bullying history, trauma exposure, peer relationships, family functioning, and platform-specific behaviors. It would also be valuable to compare multiple machine learning models across different cultural contexts and age groups to determine whether the same predictors remain important in adolescents, university students, clinical populations, and non-student young adults. In addition, future research should combine self-report data with digital trace data, ecological momentary assessment, or qualitative interviews to better understand the motives, emotional states, and social meanings associated with digital self-harm episodes. Cross-cultural validation is also necessary to determine whether the meaning and expression of digital self-harm differ across Latin American, European, Asian, and North American contexts.

From a practical perspective, the findings suggest that prevention and intervention efforts should not treat digital self-harm simply as problematic internet use. Screening programs in universities, counseling centers, and digital mental health services should assess self-critical thinking, hostile self-attitudes, emotion regulation difficulties, and toxic online disinhibition when evaluating risk for digital self-harm. Interventions should focus on strengthening self-reassurance, improving adaptive emotion regulation strategies, reducing impulsive online responding, and helping young adults recognize how anonymity and disinhibition may intensify self-directed harm. Digital literacy programs should include content on anonymous self-directed abuse, not only cyberbullying toward others. Clinicians and counselors should ask sensitively about whether clients have ever posted harmful content about

themselves anonymously or encouraged others to attack them online, because many individuals may not spontaneously disclose this behavior. Online platforms, educational institutions, and mental health professionals should also collaborate to create confidential reporting and support pathways for users who show signs of self-directed digital harm.

Acknowledgments

We would like to express our appreciation and gratitude to all those who cooperated in carrying out this study.

Declaration of Interest

The authors of this article declared no conflict of interest.

Ethical Considerations

The study protocol adhered to the principles outlined in the Helsinki Declaration, which provides guidelines for ethical research involving human participants.

Transparency of Data

In accordance with the principles of transparency and open research, we declare that all data and materials used in this study are available upon request.

Funding

This research was carried out independently with personal funding and without the financial support of any governmental or private institution or organization.

Authors' Contributions

All authors equally contributed in this article.

Declaration

In order to correct and improve the academic writing of our paper, we have used the language model ChatGPT.

References

- Achab, S., Rothen, S., Giustiniani, J., Nicolier, M., Franc, E., Zullino, D., Mauny, F., & Haffen, É. (2022). Predictors of Gaming Disorder or Protective From It, in a French Sample: A Symptomatic Approach to Self-Regulation and Pursued Rewards, Providing Insights for Clinical Practice. *International journal of environmental research and public health*, 19(15), 9476. <https://doi.org/10.3390/ijerph19159476>
- Agathokleous, G., & Taiwo, A. O. (2022). How Online Counselling Is Utilised, Evaluated, and Received. 251-277. <https://doi.org/10.4018/978-1-7998-7991-6.ch011>
- Bant, L., Bogaerts, S., Howard, R., Mazzeschi, C., & Garofalo, C. (2025). The Quadripartite Violence Typology: Further Validation of the Angry Aggression Scale in a Dutch Community Sample. *Psychology of violence*, 15(6), 712-729. <https://doi.org/10.1037/vio0000594>
- Booth, N., Derefaka, G., Khan, R., & Brewer, G. (2023). Cyber Intimate Partner Aggression in Adulthood: The Role of Insecure Attachment and Self-Control. *Journal of Criminal Psychology*, 13(4), 367-380. <https://doi.org/10.1108/jcp-03-2023-0015>
- Carmichael, J., Ponsford, J., Gould, K. R., Forbes, M. K., Kotov, R., Fornito, A., & Spitz, G. (2023). A Transdiagnostic, Hierarchical Taxonomy of Psychopathology Following Traumatic Brain Injury (HiTOP-TBI). <https://doi.org/10.31234/osf.io/vxtqk>
- Ding, J., Lin, Y., & Chen, I. H. (2025). Why Individuals With Trait Anger and Revenge Motivation Are More Likely to Engage in Cyberbullying Perpetration? The Online Disinhibition Effect. *Frontiers in Public Health*, 13. <https://doi.org/10.3389/fpubh.2025.1496965>
- Gan, W., Chen, Z., Wu, Z., Huang, X., & Wang, F. (2024). Aggression in Online Gaming: The Role of Online Disinhibition, Social Dominance Orientation, Moral Disengagement and Gender Traits Among Chinese University Students. *Frontiers in Public Health*, 12. <https://doi.org/10.3389/fpubh.2024.1459696>
- Giannakopoulos, G., & Prassou, A. (2025). Mediating and Moderating Mechanisms in the Relationship Between Social Media Use and Adolescent Aggression: A Scoping Review of Quantitative Evidence. *European Journal of Investigation in Health Psychology and Education*, 15(6), 98. <https://doi.org/10.3390/ejihpe15060098>
- Hudon, A., Harvey, E., Nicolas, S., Dufour, M., Guérin-Thériault, C., Bérubé-Fortin, J., Combey, I., Yue, Y. C., Perreault, A., Pagé, S. B., & MacDermott, V. (2025). Cyberpsychopathy: A Multidimensional Framework for Understanding Psychopathic Traits in Digital Environments. *European Journal of Investigation in Health Psychology and Education*, 15(6), 107. <https://doi.org/10.3390/ejihpe15060107>
- Lai, M. H. C. (2024). The Psychological Boomerang: Investigating the Mental Health Impacts of Spreading Anti-Lgbtq+ Misinformation and Hate Speech Online. *Communications in Humanities Research*, 63(1), 77-82. <https://doi.org/10.54254/2753-7064/2024.17703>
- Lee, N. (2026). Understanding Bystander Responses to Virtual Sexual Assault: Predictors of Intervention and Self-Protection in Virtual Environments. *International Journal of Police Science & Management*, 28(2), 159-170. <https://doi.org/10.1177/14613557261447745>
- Looi, J. C., Allison, S., & Bastiampillai, T. (2023). Remaining on Course, Online: Reflections and Recommendations on More Effective Psychiatrist Participation in Online Chat Forums in the Context of Social Media Dynamics. *Australasian Psychiatry*, 31(5), 652-655. <https://doi.org/10.1177/10398562231186244>
- Maftai, A., Opariuc-Dan, C., & Grigore, A.-N. (2023). Toxic Sensation Seeking? Psychological Distress, Cyberbullying, and the Moderating Effect of Online Disinhibition Among Adults. *Scandinavian journal of psychology*, 65(1), 61-69. <https://doi.org/10.1111/sjop.12956>
- Mantara, A. Y., Sa'id, M., Zahra, G. A., Rizkina, A. T., & Febriyanti, L. (2023). Adaptation of the Online Disinhibition

- Effect Scale. *Kne Social Sciences*.
<https://doi.org/10.18502/kss.v8i19.14381>
- Mao, Y., & Hu, B. (2024). From Trolling Victimization to Reactive Trolling: Moderated Mediation Effects of Online Disinhibition and Motivations. *Social Science Computer Review*, 43(1), 27-47.
<https://doi.org/10.1177/08944393241250013>
- Morelli, M., Pazzagli, C., Sciabica, G. M., Chirumbolo, A., & Andreassi, S. (2025). The Association Between Pathological Personality Domains and Aggravated and Risky Sexting in a Sample of Emerging Adults. *Scandinavian journal of psychology*, 67(1), 147-157.
<https://doi.org/10.1111/sjop.70021>
- Nazari, A., Mafakheri, N., Shafiei, R., Nargesi, F., Sharp, C., & Komasi, S. (2025). Decoding Personality Functioning: The Impact of DSM-5 Pathological Traits Mediated by Emotion Dysregulation and Metacognition. *Borderline personality disorder and emotion dysregulation*, 12(1).
<https://doi.org/10.1186/s40479-025-00315-1>
- Nocera, T. R., Dahlen, E. R., Mohn, R. S., Leuty, M. E., & Batastini, A. B. (2022). Dark Personality Traits and Anger in Cyber Aggression Perpetration: Is Moral Disengagement to Blame? *Psychology of Popular Media*, 11(1), 24-34.
<https://doi.org/10.1037/ppm0000295>
- Pan, Y., Chen, G., Liu, Q., & Shi, Y. (2022). Research on the Governance Dilemma and Countermeasures of Cyber Violence. 1409-1421. https://doi.org/10.2991/978-2-494069-89-3_162
- Ramadhani, A. Z., & Merida, S. C. (2024). Self-Control and the Phenomenon of Toxic Online Disinhibition in Teenagers Who Have Twitter. *Nusantara Journal of Behavioral and Social Sciences*, 3(2), 45-52. <https://doi.org/10.47679/202454>
- Rehman, T. U. (2024). Workplace Cyberbullying: Nature, Characteristics, and Implications. *Evaluation Review*, 49(4), 625-654. <https://doi.org/10.1177/0193841x241304293>
- Rose, L., Phillips, N. L., Lynam, D. R., Bauer, B. W., & Miller, J. D. (2025). Associations Between Psychopathy and Suicide: A Meta-Analytic Review. *Personality Disorders Theory Research and Treatment*, 16(6), 562-576.
<https://doi.org/10.1037/per0000743>
- Sánchez-Hernández, M. D., Herrera, M. C., & Expósito, F. (2024). Is Online Disinhibition Related to Cyberdating Abuse Perpetration Through Moral Disengagement? The Moderating Role of Gender, Sexism, and Cybervictimization. *Sex Roles*, 90(7), 938-959. <https://doi.org/10.1007/s11199-024-01480-3>
- Scott, J. E., & Barlett, C. P. (2023). Understanding Cyber-Racism Perpetration Within the Broader Context of Cyberbullying Theory: A Theoretical Integration. *Children*, 10(7), 1156.
<https://doi.org/10.3390/children10071156>
- Sharma, M. K., Anand, N., Srivastava, K., & Sunil, S. (2021). Internet Regression and the Meaning of Self. *Medico-Legal Journal*, 89(2), 122-124.
<https://doi.org/10.1177/0025817220984014>
- Soares, F. B., Gruzd, A., Jacobson, J., & Hodson, J. (2023). To Troll or Not to Troll: Young Adults' Anti-Social Behaviour on Social Media. *PLoS One*, 18(5), e0284374.
<https://doi.org/10.1371/journal.pone.0284374>
- Vale, M., & Matos, M. (2023). How Can Cyberhate Victimization and Perpetration Be Understood? , 130-150.
<https://doi.org/10.4018/979-8-3693-1528-6.ch008>
- Wang, L., & Ngai, S. S. (2021). Understanding the Effects of Personal Factors and Situational Factors for Adolescent Cyberbullying Perpetration: The Roles of Internal States and Parental Mediation. *Journal of adolescence*, 89(1), 28-40.
<https://doi.org/10.1016/j.adolescence.2021.03.006>
- Wanjiku, E. (2022). Revenge Pornography on the Internet. *Communicare Journal for Communication Studies in Africa*, 40(1), 151-170. <https://doi.org/10.36615/jcsa.v40i1.1515>
- You, L. (2023). The Bystander Effect: Why Do Witnesses Refrain From Acting in Cyberbullying Cases?
<https://doi.org/10.31219/osf.io/zrqp6>
- Zhang, Q., Chen, Q., & Gu, X. (2024). Trolling, Cyberstalking, Body-Shaming, Slut-Shaming a Study on Online Abuses of Social Media. *Communications in Humanities Research*, 28(1), 232-241. <https://doi.org/10.54254/2753-7064/28/20230304>
- Zhao, Z. (2023). The Causes and Measures of Online Deviance. *Journal of Education Humanities and Social Sciences*, 22, 415-420. <https://doi.org/10.54097/ehss.v22i.12491>
- Zheng, G., Huang, J., Lan, J., & Wang, Y. (2025). Parenting in Context: A Systematic Review of Family Factors and Cyberbullying Perpetration Among Adolescents. *Child & Family Social Work*, 31(1), 420-446.
<https://doi.org/10.1111/cfs.13281>