KMAN PUBLISHING INSTITUTE
Journal Website

Article history:
Received 25 September 2025
Revised 15 December 2025
Accepted 19 December 2025
Published Online 10 February 2026

Journal of Adolescent and Youth Psychological Studies

Volume 7, Issue 2, pp 1-11

E-ISSN: 2981-2526

# Predicting Cyberbullying Perpetration via Random Forest Modeling of Moral Disengagement and Empathy Deficits

Tomas. Jankauskas[1], Kabelo. Radebe[2*], Amira. Chennoufi[3]

[1] Department of Social Psychology, Vilnius University, Vilnius, Lithuania
[2] Department of Health Psychology, North-West University, Potchefstroom, South Africa
[3] Department of Clinical Psychology, University of Sfax, Sfax, Tunisia

*** Corresponding author email address**: kabelo.radebe@nwu.ac.za

A r t i c l e  I n f o

**Article type:**
*Original Research*

**How to cite this article:**
Jankauskas, T., Radebe, K., & Chennoufi, A. (2026). Predicting Cyberbullying Perpetration via Random Forest Modeling of Moral Disengagement and Empathy Deficits. *Journal of Adolescent and Youth Psychological Studies, 7*(2), 1-11.
http://dx.doi.org/10.61838/kman.jayps.5086



A B S T R A C T

**Objective:** The present study aimed to predict cyberbullying perpetration among adolescents using Random Forest modeling of moral disengagement mechanisms and empathy deficits.

**Methods and Materials:** This cross-sectional quantitative study was conducted among 742 secondary school students (ages 13–18 years) from three provinces in South Africa using multi-stage cluster sampling. Participants completed validated self-report instruments measuring cyberbullying perpetration, moral disengagement, and empathy deficits, alongside demographic indicators and daily internet usage. Data were screened, cleaned, and randomly divided into training (70%) and testing (30%) datasets. A Random Forest classifier with 500 trees was trained to distinguish high versus low cyberbullying perpetration. Hyperparameters were optimized using cross-validation. Model performance was evaluated through accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). A logistic regression model was estimated as a baseline comparator. Variable importance indices and partial dependence plots were generated to examine predictor contributions and non-linear interaction patterns.

**Findings:** Inferential analyses indicated that moral disengagement and empathy deficits were significant predictors of cyberbullying perpetration ($p < .001$). Male students reported higher levels of cyberbullying and moral disengagement ($p < .001$). The Random Forest model outperformed logistic regression, achieving superior classification accuracy (0.86 vs. 0.74) and AUC-ROC (0.91 vs. 0.78). Variable importance metrics identified overall moral disengagement, dehumanization, and attribution of blame as the strongest predictors, followed by empathy deficits. Partial dependence analysis revealed non-linear threshold effects, with sharp increases in predicted cyberbullying probability at higher levels of moral disengagement, particularly when combined with elevated empathy deficits.

**Conclusion:** The findings demonstrate that cyberbullying perpetration is most strongly predicted by moral disengagement mechanisms and empathy deficits.

***Keywords:*** *Cyberbullying perpetration; Moral disengagement; Empathy deficits; Random Forest; Machine learning; Adolescents.*

## 1. Introduction

The rapid digitalization of adolescent social life has transformed peer interaction into a hybrid ecology in which face-to-face and online dynamics are deeply intertwined. While digital platforms offer unprecedented opportunities for connection, identity exploration, and civic engagement, they also facilitate new forms of aggression, among which cyberbullying has emerged as a critical public health and educational concern. Cyberbullying perpetration—defined as intentional and repeated harm inflicted through electronic communication technologies—has been consistently associated with psychological distress, academic maladjustment, conduct problems, and long-term psychosocial impairment. Recent empirical work underscores that cyberbullying is not merely a technological extension of traditional bullying but rather a phenomenon shaped by unique contextual affordances such as anonymity, audience scalability, permanence of content, and reduced immediate social feedback (Wachs et al., 2022; Wang et al., 2020). These characteristics alter moral appraisal processes and emotional regulation, thereby reshaping how adolescents interpret and justify aggressive online behavior.

Contemporary research increasingly conceptualizes cyberbullying perpetration within a socio-cognitive framework grounded in Bandura's moral disengagement theory. Moral disengagement refers to cognitive mechanisms that allow individuals to disengage internal moral standards from harmful conduct, thereby minimizing guilt and self-sanctions. Mechanisms such as moral justification, displacement of responsibility, diffusion of responsibility, dehumanization, and distortion of consequences have been shown to facilitate aggression in both offline and online contexts. Empirical evidence consistently demonstrates that moral disengagement is a robust predictor of cyberbullying perpetration across cultural settings (Bakioğlu et al., 2024; Cheng, 2024; Fissel et al., 2024). Longitudinal analyses further suggest that moral disengagement not only precedes cyber-aggressive behavior but may also be reinforced through repeated perpetration, forming a reciprocal amplification cycle (Falla et al., 2023; Gao et al., 2023). The stability of this association has been replicated in diverse adolescent samples, including those exposed to deviant peer affiliations and violent media influences (Wang & Zhou, 2023; Wang et al., 2023).

Parallel to cognitive moral processes, empathy has been identified as a central protective factor against cyber-aggression. Empathy—comprising affective responsiveness to others' emotional states and cognitive perspective-taking—functions as an internal inhibitory mechanism that constrains aggressive impulses. Reduced empathy or empathy deficits have been consistently linked to greater likelihood of engaging in cyberbullying (Francisco et al., 2023; Günal & Ayaz-Alkaya, 2025). The interplay between empathy and moral disengagement is particularly salient: adolescents with lower empathic sensitivity may more readily dehumanize victims or minimize perceived harm, thereby activating moral disengagement mechanisms (Castellanos et al., 2023; Francisco et al., 2024). Studies examining hate speech, racist discourse, and stigmatization-based cyber aggression indicate that diminished empathy significantly interacts with moral cognitive distortions to predict online hostility (Rodríguez-Hidalgo et al., 2025; Wachs et al., 2023).

Longitudinal evidence strengthens the argument for a dynamic interplay between empathy erosion and moral disengagement in predicting cyberbullying. For instance, cross-lagged models reveal that early aggression predicts subsequent increases in moral disengagement and decreases in empathy, suggesting bidirectional causality (Falla et al., 2021). Similarly, mediation analyses demonstrate that empathy indirectly influences cyberbullying through moral disengagement pathways, indicating that cognitive justifications partially account for empathy-related risk (Ling et al., 2023; Luo & Bussey, 2022). The mediating role of moral disengagement has also been observed in contexts involving authoritarian parenting, emotional intelligence deficits, and online anonymity (Lubis et al., 2022; Rahmawati & Virlia, 2023). Such findings highlight that empathy and moral disengagement are not isolated constructs but operate within broader ecological systems shaped by family, peers, and digital affordances.

Beyond individual moral-cognitive traits, contextual and relational influences significantly shape cyberbullying perpetration. Parental psychological control, parental phubbing, and parental rejection have been identified as predictors of adolescents' moral disengagement and cyber-aggressive conduct (Lan et al., 2025; Wang et al., 2020; Xu, 2025). Peer selection and influence processes further intensify aggression trajectories by normalizing deviant norms and diffusing responsibility (Tu et al., 2025; Wang & Zhou, 2023). Exposure to violent video games and trait anger have similarly been associated with cyberbullying through moderated mediation pathways involving moral disengagement (Wang et al., 2023; Yang et al., 2020). These findings converge on the notion that moral disengagement



JAYPS Adolescent and Youth Psychological Studies
E-ISSN: 2981-2526

functions as a cognitive bridge linking environmental risk factors to behavioral outcomes.

The moral dimension of cyberbullying is further illuminated by research on moral identity, internet moral judgment, and online authenticity. Adolescents with stronger moral identity internalization exhibit lower moral disengagement and higher empathy, which in turn reduces cyber-aggressive behavior (Morgan & Fowers, 2021; Yang et al., 2023). Conversely, individuals exhibiting callous-unemotional traits and dark tetrad personality features demonstrate heightened moral disengagement and reduced empathic responsiveness, thereby elevating cyberbullying risk (Gajda et al., 2022; Gómez & Durán, 2024). Cross-national comparisons in Spain and Poland underscore that social, emotional, and moral competencies collectively buffer against both bullying and cyberbullying perpetration (Llorent et al., 2021).

Recent scholarship has extended these insights to diverse populations, including juvenile offenders and emerging adults, revealing that moral disengagement retains predictive salience across developmental stages (Sylvain & Talpade, 2024; Zhang & Konishi, 2024). Moreover, research on onset risk factors identifies early deficits in socio-emotional regulation and peer relations as precursors to later cyberbullying involvement (Sorrentino et al., 2023). Studies focusing on social stigma, racism, and online hate speech indicate that moral disengagement mechanisms amplify prejudicial aggression when empathy is attenuated (Rodríguez-Hidalgo et al., 2025; Wachs et al., 2022).

Despite the robust body of evidence linking moral disengagement and empathy to cyberbullying, much of the existing literature relies on traditional linear statistical techniques, primarily regression-based mediation and moderation models. While these approaches are valuable for hypothesis testing and inferential clarity, they may inadequately capture complex non-linear interactions, threshold effects, and hierarchical predictor relationships inherent in cyber-aggressive behavior. Emerging methodological discussions advocate for integrating machine learning approaches to enhance predictive accuracy and identify high-risk profiles (Eden & Landau, 2025; Xiao et al., 2025). Machine learning algorithms such as Random Forest modeling allow for the detection of intricate interaction patterns among socio-cognitive variables without imposing strict parametric assumptions.

Furthermore, although several studies validate measurement scales for moral disengagement and empathy in online contexts (Bakioğlu et al., 2024; Concha-Salgado et al., 2022), there remains limited research applying ensemble learning techniques to simultaneously model multidimensional moral disengagement mechanisms and empathy deficits as predictors of cyberbullying perpetration. Recent work exploring cross-sectional and longitudinal associations underscores the necessity of predictive analytics capable of distinguishing adolescents at elevated risk for intervention targeting (Abdelaliem, 2024; Günal & Ayaz-Alkaya, 2025). In addition, social cognition research suggests that emotional responses to bullying scenarios vary across developmental stages, which may introduce non-linear age-related patterns not adequately captured by traditional regression (Arató et al., 2020; Tao, 2023).

Integrating these theoretical and empirical strands, a predictive modeling framework grounded in moral disengagement theory and empathy research offers a promising avenue for advancing cyberbullying prevention science. By leveraging Random Forest algorithms, it becomes possible to quantify the relative importance of distinct moral disengagement mechanisms (e.g., dehumanization versus diffusion of responsibility), examine interaction surfaces between empathy deficits and cognitive distortions, and evaluate classification accuracy beyond conventional statistical indices. Such methodological innovation aligns with calls for data-driven identification of psychosocial risk markers in digital aggression research (Falla et al., 2023; Wachs et al., 2023).

In light of accumulating evidence demonstrating the centrality of moral disengagement and empathy deficits in cyberbullying perpetration, alongside the need for advanced predictive methodologies capable of capturing complex non-linear interactions, the present study aims to predict cyberbullying perpetration among adolescents using Random Forest modeling of moral disengagement mechanisms and empathy deficits.

## 2. Methods and Materials

### *2.1. Study Design and Participants*

This study was designed as a cross-sectional, predictive quantitative investigation aimed at modeling cyberbullying perpetration among adolescents using machine learning techniques, specifically the Random Forest algorithm. The target population comprised secondary school students enrolled in public schools across three provinces of South Africa: Gauteng, KwaZulu-Natal, and the Western Cape. A multi-stage cluster sampling strategy was employed to ensure geographical and socio-demographic



JAYPS Adolescent and Youth Psychological Studies E-ISSN: 2981-2526

representativeness. In the first stage, districts were randomly selected within each province. In the second stage, schools within selected districts were randomly chosen from official provincial education department lists. In the third stage, intact classrooms were randomly selected and all students within those classrooms were invited to participate. Inclusion criteria required participants to be between 13 and 18 years of age, enrolled in Grades 8 to 12, and able to complete a self-report questionnaire in English. Students with identified severe cognitive impairments or those absent on the day of data collection were excluded. The final sample consisted of 742 adolescents (n = 742), of whom 381 were female (51.3%) and 361 were male (48.7%). The mean age of participants was 15.67 years (SD = 1.42).

### 2.2. *Measures*

Data were collected using a structured self-report questionnaire composed of standardized psychometric instruments validated for adolescent populations. Cyberbullying perpetration was assessed using the Cyberbullying Offending Scale, adapted for the South African context. The instrument measures the frequency of engaging in behaviors such as sending threatening messages, spreading rumors online, impersonation, and posting humiliating content over the past six months. Responses were recorded on a five-point Likert scale ranging from 1 (never) to 5 (very frequently). Higher scores indicated greater involvement in cyberbullying perpetration. Internal consistency reliability in the present study was satisfactory (Cronbach's alpha = 0.89). Moral disengagement was measured using the Moral Disengagement Scale for Adolescents, which assesses cognitive mechanisms such as moral justification, euphemistic labeling, advantageous comparison, displacement of responsibility, diffusion of responsibility, distortion of consequences, dehumanization, and attribution of blame. The scale consists of 24 items rated on a five-point Likert continuum from strongly disagree to strongly agree. In the present sample, the overall reliability coefficient was 0.91. Empathy deficits were assessed using the Basic Empathy Scale, which captures both cognitive and affective components of empathy. For the purpose of modeling empathy deficits, items were reverse-coded where necessary so that higher scores reflected lower empathic responsiveness. The scale demonstrated acceptable internal consistency (Cronbach's alpha = 0.87). Demographic variables including age, gender, grade level, and daily internet usage time were also collected and considered as potential covariates. Prior to administration, instruments were piloted with a small group of 30 students to ensure clarity and cultural appropriateness; minor wording adjustments were made accordingly. All questionnaires were administered in classroom settings under the supervision of trained research assistants to standardize administration procedures and minimize response bias.

### 2.3. *Data Analysis*

Data analysis was conducted in several sequential stages using R (version 4.x) and Python (scikit-learn library). Initially, data screening procedures were performed, including inspection for missing values, outliers, and normality assumptions. Missing data constituted less than 3% of the dataset and were handled using multiple imputation via chained equations to preserve statistical power and reduce bias. Descriptive statistics and Pearson correlations were calculated to examine preliminary relationships among variables. For predictive modeling, the dataset was randomly partitioned into a training set (70%, n = 519) and a testing set (30%, n = 223) to evaluate out-of-sample performance. The primary analytic approach involved training a Random Forest classifier to predict levels of cyberbullying perpetration. Cyberbullying perpetration scores were dichotomized into low and high categories using a theoretically informed percentile cutoff to facilitate classification modeling. The Random Forest model was constructed with 500 decision trees, and hyperparameters including the number of variables randomly sampled at each split (mtry) and maximum tree depth were optimized using five-fold cross-validation within the training set. Model performance was evaluated using accuracy, precision, recall, F1-score, and area under the Receiver Operating Characteristic curve (AUC-ROC). Variable importance was assessed using mean decrease in Gini impurity and permutation importance methods to identify the relative contribution of moral disengagement dimensions and empathy deficits to cyberbullying perpetration prediction. Additionally, partial dependence plots were generated to examine non-linear relationships between key predictors and the probability of high cyberbullying perpetration. To compare model performance, a logistic regression baseline model was also estimated, and comparative metrics were reported. Statistical significance for traditional analyses was set at $p < .05$, while machine learning evaluation emphasized predictive accuracy and generalization performance rather than solely inferential significance.



JAYPS Adolescent and Youth Psychological Studies
E-ISSN: 2981-2526

## 3. Findings and Results

Prior to conducting predictive modeling, descriptive statistics and zero-order correlations among the principal study variables were examined to provide an overview of central tendencies, dispersion indices, and preliminary associations. Table 1 presents the means, standard deviations, and Pearson correlation coefficients among cyberbullying perpetration, overall moral disengagement, empathy deficits, and daily internet usage time.

**Table 1**

*Descriptive Statistics and Pearson Correlations Among Study Variables (N = 742)*

| Variable | M | SD | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| 1. Cyberbullying Perpetration | 2.31 | 0.87 | — | | | |
| 2. Moral Disengagement | 2.78 | 0.74 | 0.52** | — | | |
| 3. Empathy Deficits | 2.64 | 0.69 | 0.47** | 0.41** | — | |
| 4. Daily Internet Usage (hours) | 3.92 | 1.58 | 0.29** | 0.18** | 0.22** | — |

**p < 0.01

As shown in Table 1, the mean level of cyberbullying perpetration was 2.31 (SD = 0.87), indicating that, on average, students reported low to moderate engagement in cyberbullying behaviors during the past six months. Moral disengagement demonstrated a mean of 2.78 (SD = 0.74), while empathy deficits showed a mean of 2.64 (SD = 0.69), suggesting moderate variability across participants in cognitive–moral justifications and empathic responsiveness. Correlational analyses revealed a strong positive association between moral disengagement and cyberbullying perpetration ($r = 0.52$, $p < 0.01$), indicating that adolescents who endorsed higher levels of moral disengagement mechanisms were significantly more likely to report cyberbullying behaviors. Empathy deficits were also positively and significantly associated with cyberbullying perpetration ($r = 0.47$, $p < 0.01$), suggesting that reduced empathic concern and perspective-taking were linked to higher perpetration. Daily internet usage demonstrated a moderate positive relationship with cyberbullying ($r = 0.29$, $p < 0.01$), though the magnitude was notably smaller than that of moral disengagement and empathy deficits. The intercorrelation between moral disengagement and empathy deficits ($r = 0.41$, $p < 0.01$) indicated partial conceptual overlap while preserving discriminant validity.

To further explore demographic differences in cyberbullying perpetration and key predictors, group comparisons were conducted across gender and grade level. The results of independent samples t-tests and one-way ANOVA analyses are summarized in Table 2.

**Table 2**

*Group Differences in Cyberbullying Perpetration and Predictors by Gender and Grade Level*

| Variable | Group | M | SD | Test Statistic | p |
|---|---|---|---|---|---|
| Cyberbullying Perpetration | Male (n = 361) | 2.46 | 0.90 | t = 3.84 | < 0.001 |
| | Female (n = 381) | 2.17 | 0.83 | | |
| Moral Disengagement | Male | 2.91 | 0.76 | t = 4.12 | < 0.001 |
| | Female | 2.66 | 0.71 | | |
| Empathy Deficits | Male | 2.78 | 0.68 | t = 3.97 | < 0.001 |
| | Female | 2.51 | 0.67 | | |
| Cyberbullying Perpetration | Grade 8–9 | 2.18 | 0.81 | F = 5.62 | 0.001 |
| | Grade 10–11 | 2.34 | 0.88 | | |
| | Grade 12 | 2.47 | 0.92 | | |

The findings in Table 2 indicate statistically significant gender differences across all primary constructs. Male students reported higher levels of cyberbullying perpetration (M = 2.46) compared to female students (M = 2.17), with the difference reaching statistical significance ($t = 3.84$, $p < 0.001$). Similarly, males demonstrated significantly higher moral disengagement and empathy deficits relative to females. Grade-level differences were also significant ($F = 5.62$, $p = 0.001$), with older students (Grade 12) exhibiting the highest mean cyberbullying perpetration scores. Post-



JAYPS Adolescent and Youth Psychological Studies
E-ISSN: 2981-2526

hoc comparisons using Tukey's test indicated that Grade 12 students differed significantly from Grade 8–9 students, whereas differences between intermediate grades were smaller and non-significant.

The core objective of this study was to evaluate the predictive performance of the Random Forest model in classifying high versus low cyberbullying perpetration. Model evaluation metrics for both the Random Forest classifier and the logistic regression baseline model are presented in Table 3.

**Table 3**

*Comparative Model Performance Metrics on Test Set (n = 223)*

| Model | Accuracy | Precision | Recall | F1-Score | AUC-ROC |
|---|---|---|---|---|---|
| Logistic Regression | 0.74 | 0.71 | 0.69 | 0.70 | 0.78 |
| Random Forest | 0.86 | 0.84 | 0.83 | 0.83 | 0.91 |

As shown in Table 3, the Random Forest classifier substantially outperformed the logistic regression baseline across all performance indices. The Random Forest model achieved an accuracy of 0.86, indicating that 86% of cases in the test dataset were correctly classified. Precision and recall values of 0.84 and 0.83 respectively suggest strong balance between correctly identifying high cyberbullying perpetrators and minimizing false positives. The F1-score of 0.83 reflects robust harmonic integration of precision and recall. Most notably, the AUC-ROC value of 0.91 indicates excellent discriminatory capacity between high and low cyberbullying groups. In contrast, logistic regression yielded lower predictive performance, particularly in recall and AUC values. These findings demonstrate the superior capacity of non-linear ensemble modeling to capture complex interactions and hierarchical relationships among moral disengagement mechanisms and empathy deficits.

**Table 4**

*Random Forest Variable Importance Rankings*

| Predictor | Mean Decrease in Gini | Permutation Importance |
|---|---|---|
| Overall Moral Disengagement | 0.142 | 0.118 |
| Dehumanization | 0.119 | 0.102 |
| Attribution of Blame | 0.111 | 0.096 |
| Empathy Deficits (Total) | 0.108 | 0.091 |
| Distortion of Consequences | 0.094 | 0.082 |
| Diffusion of Responsibility | 0.087 | 0.074 |
| Daily Internet Usage | 0.061 | 0.049 |
| Age | 0.039 | 0.028 |
| Gender | 0.031 | 0.022 |

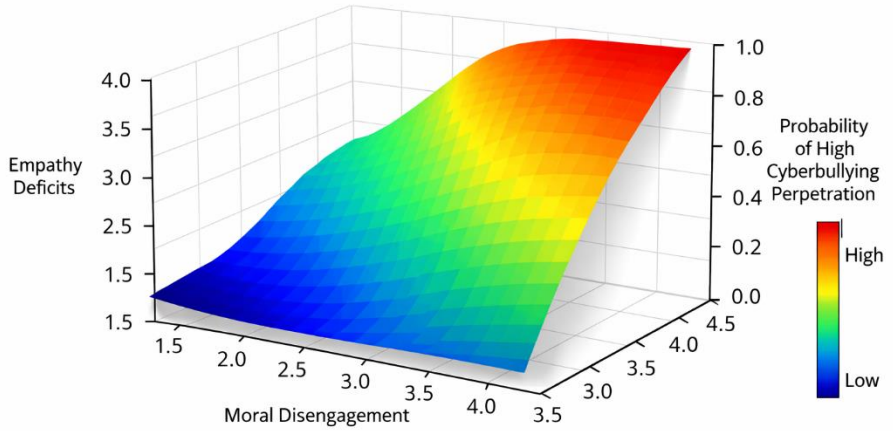
The results in Table 4 reveal that overall moral disengagement emerged as the most influential predictor in classifying cyberbullying perpetration, followed by specific mechanisms such as dehumanization and attribution of blame. Empathy deficits ranked fourth in importance, indicating that while empathic impairment plays a significant role, moral cognitive restructuring mechanisms exert even stronger predictive influence. Notably, demographic variables such as age and gender contributed less substantially to model performance compared to psychological constructs. These findings underscore the primacy of moral cognitive distortions in predicting cyber-aggressive behavior within the sampled adolescent population.



JAYPS Adolescent and Youth Psychological Studies
E-ISSN: 2981-2526

**Figure 1**

*Partial Dependence Plot of Moral Disengagement and Empathy Deficits on Probability of High Cyberbullying Perpetration*




The partial dependence analysis demonstrated a clear non-linear relationship between moral disengagement and the predicted probability of high cyberbullying perpetration. Specifically, the probability curve exhibited a relatively gradual increase at lower levels of moral disengagement, followed by a sharp escalation beyond a threshold score of approximately 3.00. A similar but slightly attenuated pattern was observed for empathy deficits, with higher deficits corresponding to progressively increased predicted probability. The interaction surface indicated that adolescents simultaneously scoring high on moral disengagement and empathy deficits exhibited the greatest predicted risk, suggesting synergistic amplification effects between moral cognitive distortions and reduced empathic concern. These patterns illustrate the advantage of Random Forest modeling in capturing threshold effects and interaction dynamics that may not be detectable using linear parametric approaches.

## 4. Discussion

The primary objective of the present study was to predict cyberbullying perpetration among adolescents using Random Forest modeling of moral disengagement mechanisms and empathy deficits. The findings provide robust empirical support for the central role of moral disengagement in explaining cyber-aggressive behavior. Descriptive and correlational analyses demonstrated strong positive associations between overall moral disengagement and cyberbullying perpetration, and the Random Forest variable importance metrics identified moral disengagement—particularly dehumanization and attribution of blame—as the most influential predictors. These results are consistent with accumulating cross-sectional and longitudinal evidence showing that moral disengagement functions as a critical cognitive enabler of online aggression (Cheng, 2024; Fissel et al., 2024). In alignment with findings among adolescents and emerging adults, our data reinforce the assertion that disengaging moral self-sanctions allows perpetrators to reinterpret harmful digital actions as acceptable or trivial (Abdelaliem, 2024; Zhang & Konishi, 2024).

The prominence of dehumanization in the importance ranking is particularly noteworthy. Dehumanization involves cognitively stripping victims of full moral status, which in digital contexts may be facilitated by physical distance and reduced visibility of emotional reactions. Similar mechanisms have been identified in research examining racist hate speech and stigma-based cyberbullying, where moral disengagement mediates aggression through diminished moral concern (Rodríguez-Hidalgo et al., 2025; Wachs et al., 2022). Our findings further extend prior mediation models demonstrating that deviant peer affiliation and violent media exposure increase cyberbullying via moral disengagement pathways (Wang & Zhou, 2023; Wang et al., 2023). By applying a non-linear ensemble model, the present study reveals that these moral mechanisms may exert threshold effects, where risk sharply escalates beyond moderate levels of disengagement.

Empathy deficits also emerged as a significant predictor of cyberbullying perpetration, both in correlational analyses and in the Random Forest importance ranking. This finding corroborates extensive literature indicating that reduced affective empathy and impaired perspective-taking heighten vulnerability to online aggression (Francisco et al., 2023;



JAYPS Adolescent and Youth Psychological Studies E-ISSN: 2981-2526

Günal & Ayaz-Alkaya, 2025). The partial dependence analysis revealed that high empathy deficits amplified the predictive probability of cyberbullying, particularly when combined with elevated moral disengagement. This interaction aligns with longitudinal research suggesting reciprocal dynamics between aggression, declining empathy, and increased moral disengagement (Falla et al., 2021; Falla et al., 2023). The synergy observed in the interaction surface supports theoretical models positing that empathy operates as an inhibitory control mechanism whose erosion facilitates cognitive moral restructuring (Castellanos et al., 2023; Ling et al., 2023).

Importantly, the Random Forest classifier substantially outperformed the logistic regression baseline model, achieving higher accuracy, F1-score, and AUC values. This methodological outcome highlights the value of machine learning techniques in capturing non-linear patterns that may remain obscured in traditional parametric approaches. Emerging discussions emphasize that digital aggression is shaped by complex socio-cognitive configurations rather than simple linear effects (Eden & Landau, 2025; Xiao et al., 2025). The improved classification performance observed in this study underscores the relevance of adopting advanced predictive analytics to enhance early identification of at-risk adolescents.

The demographic findings further contextualize the predictive results. Male students reported significantly higher levels of cyberbullying perpetration, moral disengagement, and empathy deficits, consistent with cross-national studies linking gender to differential aggression trajectories (Gajda et al., 2022; Llorent et al., 2021). Older adolescents demonstrated slightly elevated perpetration rates, paralleling findings indicating developmental increases in exposure to online disinhibition and peer norm internalization (Sorrentino et al., 2023; Yang et al., 2020). However, demographic variables contributed less substantially to predictive performance compared to psychological constructs, suggesting that moral-cognitive variables may transcend simple demographic categorizations.

The strong association between daily internet usage and cyberbullying, though weaker than moral disengagement and empathy deficits, aligns with prior research demonstrating that digital immersion amplifies exposure to risk contexts (Wang et al., 2020). Furthermore, parental psychological control and rejection have been shown to increase moral disengagement, which in turn predicts cyber-aggression (Lan et al., 2025; Xu, 2025). Although parental factors were not directly modeled in the present study, the findings resonate with ecological frameworks positing that family dynamics shape adolescents' moral reasoning and empathy development.

The results also converge with evidence that online moral judgment and moral identity influence cyberbullying behavior (Morgan & Fowers, 2021; Yang et al., 2023). Adolescents exhibiting weaker moral identity internalization may be more susceptible to disengagement strategies that legitimize harm. Similarly, trait callous-unemotional features and dark personality characteristics are associated with higher moral disengagement and reduced empathy, thereby heightening cyberbullying risk (Gajda et al., 2022; Gómez & Durán, 2024). The convergence of these findings strengthens the theoretical proposition that cyberbullying perpetration is best understood as the behavioral outcome of interacting moral-cognitive and socio-emotional vulnerabilities.

The mediation-oriented literature provides additional support for the interpretation of our findings. Moral disengagement has been repeatedly identified as a mediating mechanism linking emotional intelligence deficits, anonymity perceptions, and authoritarian parenting styles to cyberbullying (Lubis et al., 2022; Rahmawati & Virlia, 2023). The Random Forest modeling employed in the present study complements these inferential approaches by emphasizing predictive salience rather than solely pathway significance. In doing so, it identifies which components of moral disengagement most substantially differentiate high-risk adolescents, thereby informing targeted intervention strategies.

Cross-cultural validation studies of moral disengagement scales confirm the construct's stability across contexts (Bakioğlu et al., 2024; Concha-Salgado et al., 2022). The present findings extend this cross-cultural robustness to a South African adolescent sample, supporting the universality of moral disengagement processes in digital aggression. Moreover, the observed empathy–moral disengagement interaction is consistent with research on bystander responses and hate speech, which demonstrates that empathy moderates the moral reasoning underlying online harm (Wachs et al., 2023).

## 5. Conclusion

Taken together, the findings substantiate a socio-cognitive risk model in which moral disengagement functions as a proximal cognitive mechanism enabling



JAYPS Adolescent and Youth Psychological Studies
E-ISSN: 2981-2526

cyberbullying, while empathy deficits weaken emotional constraints against harm. The superiority of the Random Forest model indicates that predictive frameworks capable of integrating multi-dimensional moral disengagement components offer enhanced explanatory precision relative to conventional linear models. These results contribute methodologically and theoretically to the cyberbullying literature by demonstrating that ensemble learning approaches can operationalize moral-cognitive theory within high-accuracy predictive systems.

## 6. Limitations & Suggestions

Despite its contributions, this study possesses several limitations. First, the cross-sectional design precludes causal inference regarding the directional relationships between moral disengagement, empathy deficits, and cyberbullying perpetration. Although longitudinal research suggests reciprocal influences, the present data cannot determine temporal precedence. Second, reliance on self-report measures may introduce social desirability bias and shared method variance, potentially inflating observed associations. Third, the dichotomization of cyberbullying for classification modeling, while necessary for machine learning evaluation, may reduce sensitivity to gradations of perpetration severity. Fourth, although the sample was geographically diverse within South Africa, it may not fully represent adolescents from rural or socioeconomically marginalized contexts. Finally, other relevant variables such as peer norms, parental mediation, and online disinhibition were not directly incorporated into the predictive model.

Future research should adopt longitudinal and multi-wave designs to examine dynamic reciprocal processes between empathy erosion and moral disengagement over time. Integrating ecological variables such as parental practices, peer network structures, and exposure to online hate content would provide a more comprehensive predictive model. Researchers may also explore explainable artificial intelligence techniques to enhance interpretability of ensemble models. Cross-cultural comparative studies could assess whether threshold effects observed in moral disengagement replicate across diverse sociocultural environments. Additionally, incorporating behavioral or peer-report indicators may mitigate self-report bias and strengthen ecological validity.

From a practical standpoint, the findings underscore the importance of intervention programs targeting moral disengagement mechanisms and empathy enhancement simultaneously. School-based prevention initiatives should incorporate reflective activities that challenge dehumanization and blame attribution while fostering perspective-taking and emotional sensitivity. Digital literacy curricula should address the cognitive distortions that enable online harm and promote ethical decision-making in virtual contexts. Early identification systems utilizing predictive analytics may assist educators and counselors in detecting adolescents at elevated risk for cyberbullying involvement. Ultimately, integrating socio-emotional skill development with data-informed monitoring may contribute to reducing cyber-aggressive behavior and fostering safer digital school climates.

## 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 to this article.

## References

Abdelaliem, A. (2024). Cyberbullying Motivations and Moral Disengagement Among Adolescent Cyberbullies: Exploring the Mediating Roles. *Cyprus Turkish Journal of Psychiatry and Psychology*, *6*(1), 3. https://doi.org/10.35365/ctjpp.24.1.01

Arató, N., Zsidó, A. N., Lénárd, K., & Lábadi, B. (2020). Cybervictimization and Cyberbullying: The Role of Socio-Emotional Skills. *Frontiers in Psychiatry*, *11*. https://doi.org/10.3389/fpsyt.2020.00248

Bakioğlu, F., Çapan, B. E., Kırteke, S., & Pakpour, A. H. (2024). Adaptation of the Online Moral Disengagement Scale in Turkish: Its Association With Empathetic Tendency and Cyberbullying. https://doi.org/10.21203/rs.3.rs-4359069/v1

Castellanos, M., Wettstein, A., Wachs, S., & Bilz, L. (2023). Direct and Indirect Effects of Social Dominance Orientation on Hate Speech Perpetration via Empathy and Moral Disengagement Among Adolescents: A Multilevel Mediation Model. *Aggressive Behavior*, *50*(1). https://doi.org/10.1002/ab.22100

Cheng, C. (2024). Moral Disengagement and the Effect on Cyberbullying Among Adolescents and Young Adults. *Journal of Education Humanities and Social Sciences*, *40*, 60-63. https://doi.org/10.54097/zttypd29

Concha-Salgado, A., Ramírez, A. M., Pérez, B., Pérez-Luco, R., & García-Cueto, E. (2022). Moral Disengagement as a Self-Regulatory Cognitive Process of Transgressions: Psychometric Evidence of the Bandura Scale in Chilean Adolescents. *International journal of environmental research and public health*, *19*(19), 12249. https://doi.org/10.3390/ijerph191912249

Eden, S., & Landau, O. (2025). Cyberbullying, Moral Disengagement, and Empathy: Exploring Relationship in Children With Behavioral Disorder. *Psychology of violence*. https://doi.org/10.1037/vio0000631

Falla, D., Romera, E. M., & Ruiz, R. O. (2021). Aggression, Moral Disengagement and Empathy. A Longitudinal Study Within the Interpersonal Dynamics of Bullying. *Frontiers in psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.703468

Falla, D., Ruiz, R. O., Ferreira, P. C., Simão, A. M. V., & Romera, E. M. (2023). The Effect of Cyberbullying Perpetration on Empathy and Moral Disengagement: Testing a Mediation Model in a Three-Wave Longitudinal Study. *Psychology of violence*, *13*(5), 436-446. https://doi.org/10.1037/vio0000472

Fissel, E. R., Bryson, S. L., & Lee, J. R. (2024). Minimizing Responsibility: The Impact of Moral Disengagement on Cyberbullying Perpetration Among Adults. *Crime & Delinquency*, *71*(10), 3244-3268. https://doi.org/10.1177/00111287241237979

Francisco, S., Ferreira, P. C., Simão, A. M. V., & Pereira, N. (2023). Measuring Empathy Online and Moral Disengagement in Cyberbullying. *Frontiers in psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1061482

Francisco, S., Ferreira, P. C., Simão, A. M. V., & Pereira, N. (2024). Moral Disengagement and Empathy in Cyberbullying: How They Are Related in Reflection Activities About a Serious Game. *BMC psychology*, *12*(1). https://doi.org/10.1186/s40359-024-01582-3

Gajda, A., Moroń, M., Królik, M., Małuch, M., & Mraczek, M. (2022). The Dark Tetrad, Cybervictimization, and Cyberbullying: The Role of Moral Disengagement. *Current Psychology*, *42*(27), 23413-23421. https://doi.org/10.1007/s12144-022-03456-6

Gao, L., Li, X., Wu, X., & Wang, X. (2023). Longitudinal Associations Among Student–student Relationship, Moral Disengagement, and Adolescents' Bullying Perpetration. *School Psychology*, *38*(5), 337-347. https://doi.org/10.1037/spq0000534

Gómez, A. S., & Durán, N. (2024). Association Between Callous-Unemotional Traits, Empathy, and Moral Disengagement Mechanisms in Juvenile Offenders. *Anuario de Psicología Jurídica*, *34*(2), 85-95. https://doi.org/10.5093/apj2023a7

Günal, B. D., & Ayaz-Alkaya, S. (2025). Cyberbullying and Empathy Levels in Adolescents and Predictive Factors: A Cross-Sectional Study. *Public Health Nursing*. https://doi.org/10.1111/phn.70016

Lan, S., Wang, Y., Zhao, J., Hou, X., & Li, C. (2025). From Home to the Screen: How Parental Rejection Fuels Cyberbullying in College Students. *PLoS One*, *20*(5), e0323124. https://doi.org/10.1371/journal.pone.0323124

Ling, G., Li, X., & Wang, X. (2023). Agreeableness and Adolescents' Cyberbullying Perpetration: A Longitudinal Moderated Mediation Model of Moral Disengagement and Empathy. *Journal of personality*, *91*(6), 1461-1477. https://doi.org/10.1111/jopy.12823

Llorent, V. J., Diaz-Chaves, A., Zych, I., Twardowska-Staszek, E., & Marín-López, I. (2021). Bullying and Cyberbullying in Spain and Poland, and Their Relation to Social, Emotional and Moral Competencies. *School Mental Health*, *13*(3), 535-547. https://doi.org/10.1007/s12310-021-09473-3

Lubis, A. Y., Mikarsa, H. L., & Andriani, I. (2022). Mediation of Moral Disengagement on Cyberbullying Perpetration Influenced by Emotional Intelligence and Anonymity of Indonesian Adolescents on Social Media. *Российский Психологический Журнал*, *19*(4), 231-242. https://doi.org/10.21702/rpj.2022.4.15

Luo, A., & Bussey, K. (2022). Mediating Role of Moral Disengagement in the Perpetration of Cyberbullying by Victims and Bystanders. *Journal of adolescence*, *94*(8), 1142-1149. https://doi.org/10.1002/jad.12092

Morgan, B., & Fowers, B. J. (2021). Empathy and Authenticity Online: The Roles of Moral Identity, Moral Disengagement, and Parenting Style. *Journal of personality*, *90*(2), 183-202. https://doi.org/10.1111/jopy.12661

Rahmawati, N. P., & Virlia, S. (2023). The Role of Moral Disengagement and Authoritarian Parenting Style Towards Cyberbullying Attitude Among Social Media Users. *Jurnal Ilmiah Psikologi Terapan*, *11*(2), 105-111. https://doi.org/10.22219/jipt.v11i2.25550

Rodríguez-Hidalgo, A. J., Camargo, V. S., & Hurtado-Mellado, A. (2025). Cyberbullying Based on Social Stigmas and Social, Emotional and Moral Competencies. *Behavioral Sciences*, *15*(5), 646. https://doi.org/10.3390/bs15050646

Sorrentino, A., Esposito, A., Acunzo, D., Santamato, M., & Aquino, A. (2023). Onset Risk Factors for Youth Involvement in Cyberbullying and Cybervictimization: A Longitudinal Study. *Frontiers in psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1090047

Sylvain, E., & Talpade, M. (2024). Exploring the Characteristics of Cyberbullied TikTokers Based on Their Ethnicity. *International Journal of Arts Humanities & Social Science*, *05*(06), 49-53. https://doi.org/10.56734/ijahss.v5n6a8

Tao, R. (2023). Social Cognition and Emotional Response to Bullying Among College Students. *Lecture Notes in Education Psychology and Public Media*, *27*(1), 206-215. https://doi.org/10.54254/2753-7048/27/20231188

Tu, Z. C., Cui, Y., Zhang, W., & Luo, F. (2025). Peer Influence and Selection Impact on Adolescent Aggression: Exploring Nonaggressive Delinquency, Peer Victimization, and Moral Disengagement. *Journal of adolescence*, *97*(5), 1344-1359. https://doi.org/10.1002/jad.12501

Wachs, S., Bilz, L., Wettstein, A., & Espelage, D. L. (2023). Validation of the Multidimensional Bystander Responses to Racist Hate Speech Scale and Its Association With Empathy and Moral Disengagement Among Adolescents. *Aggressive Behavior*, *50*(1). https://doi.org/10.1002/ab.22105

Wachs, S., Bilz, L., Wettstein, A., Wright, M. F., Kansok-Dusche, J., Krause, N., & Ballaschk, C. (2022). Associations Between

Witnessing and Perpetrating Online Hate Speech Among Adolescents: Testing Moderation Effects of Moral Disengagement and Empathy. *Psychology of violence*, *12*(6), 371-381. https://doi.org/10.1037/vio0000422

Wang, L., & Zhou, J. (2023). Violent Video Game Exposure and Cyberbullying in Early Adolescents: A Latent Moderated Mediation Model. *Cyberpsychology Behavior and Social Networking*, *26*(6), 417-424. https://doi.org/10.1089/cyber.2022.0335

Wang, X., Wang, S., & Zeng, X. (2023). Does Deviant Peer Affiliation Accelerate Adolescents' Cyberbullying Perpetration? Roles of Moral Disengagement and Self-control. *Psychology in the Schools*, *60*(12), 5025-5040. https://doi.org/10.1002/pits.23037

Wang, X., Wang, W., Qiao, Y., Ling, G., Yang, J., & Wang, P. (2020). Parental Phubbing and Adolescents' Cyberbullying Perpetration: A Moderated Mediation Model of Moral Disengagement and Online Disinhibition. *Journal of interpersonal violence*, *37*(7-8), NP5344-NP5366. https://doi.org/10.1177/0886260520961877

Xiao, Q., Li, C., Chen, C., & Ma, J. (2025). Whose Prosocial Intentions Are More Affected by Mindfulness, Young Adolescents or Young Adults? *PsyCh Journal*, *14*(6), 912-925. https://doi.org/10.1002/pchj.70036

Xu, J. (2025). Parental Psychological Control and Cyberbullying Among Adolescents: The Mediating Roles of Sleep Quality and Moral Disengagement and the Moderating Role of Grade. *Frontiers in psychology*, *16*. https://doi.org/10.3389/fpsyg.2025.1664970

Yang, H., Zhang, T., Shi, H.-f., & Fan, C. (2023). Empathy and Bystander Helping Behavior in Cyberbullying Among Adolescents: The Mediating Role of Internet Moral Judgment and the Moderating Role of Internet Self-Efficacy. *Frontiers in psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1196571

Yang, J., Li, W., Ling, G., & Wang, X. (2020). How Is Trait Anger Related to Adolescents' Cyberbullying Perpetration? A Moderated Mediation Analysis. *Journal of interpersonal violence*, *37*(9-10), NP6633-NP6654. https://doi.org/10.1177/0886260520967129

Zhang, M., & Konishi, C. (2024). Cyberbullying Among Emerging Adults: The Role of Parental Practices and Morality. *Journal of Education and Development*, *8*(1), 27. https://doi.org/10.20849/jed.v8i1.1396



JAYPS Adolescent and Youth Psychological Studies
E-ISSN: 2981-2526