Machine Learning with Applications 20 (2025) 100646

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Predicting cyberbullying victimisation in emerging markets and developing countries using the Global School-Based Health Survey

Paulo Ricardo Vieira Braga *, Katie Rose Tyrrell

University of Suffolk, UK

ARTICLE INFO

Keywords:
Cyberbullying
Cyberbullying victimisation
Predicting cyberbullying
Developing countries
Global school based health survey

ABSTRACT

Objectives: This study aimed to identify predictors of cyberbullying victimisation among adolescents and develop predictive models to support early intervention strategies.

Methods: Data from the Global School-based Health Surveys (2017–2021) were analysed, focusing on emerging markets and developing countries. A simple random sampling strategy was used to ensure equal representation across countries. A multivariable logistic regression model was applied to 26 variables to identify significant predictors of cyberbullying victimisation. Subsequently, machine learning techniques were used to develop predictive models.

Results: This logistic regression model was statistically significant ($\chi 2(26)=507.96$, $p < 0.001$), explaining 19.3 % of the variance with an AUROC of 0*.758 (95 % CI, 0.739 to 0.778)*. Twelve variables, including being bullied on school property, female gender, peer victimisation, early sexual debut, alcohol consumption, and suicidal ideation, were identified as significant predictors. The best-performing predictive model, a randomly over-sampled random forest classifier, achieved 82 % accuracy and an AUROC of 0*.83 (95 % CI, 0.81 to 0.85)*.

Conclusions: The study highlights key predictors of cyberbullying victimisation and demonstrates the potential of machine learning in developing accurate predictive models. However, reliance on self-reported data may introduce biases. Future research could integrate diverse data sources to enhance model accuracy and reliability.

1. Introduction

Cyberbullying can be defined as an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend themselves (Smith et al., 2008). Due to the ubiquity of digital technologies, young people increasingly experience ‘perpetual contact’ (Katz & Aakhus, 2002), creating environments in which it may become difficult to avoid and negate abusive behaviours. Unlike traditional forms of bullying, cyberbullying perpetrators may utilise affordances of online platforms to facilitate anonymity, making it more difficult to identify and stop (Smith, 2012). Similarly, to bullying via traditional, face-to-face methods, cyberbullying can have significant negative impacts upon emotional health and wellbeing, as well as feelings of insecurity both at home and in education settings (Cowie, 2013). Patchin and Hinduja (2012) highlight that cyberbullying can have a significant impact on the mental health of victims, including increased levels of anxiety, depression, and even suicide. These findings are corroborated by Kowalski and Limber (2013) and Van Geel et al. (2014) who found victims of cyberbullying to experience the most negative scores relating to measures of psychological health, physical health and academic performance.

The association of cyberbullying with suicidal behaviours is of particular concern, as in 2021, WHO (2023a) reported suicide as the third leading cause of death among adolescents and young adults (aged between 10–24 years old), which is corroborated by wider literature (CDC, 2023; Royal College of Paediatrics & Child Health, 2020). Providing safe supportive online and offline environments for young people could contribute to the reduction of mental illness, especially considering adolescents are more susceptible to cyberbullying, with an estimated 14 % to 57 % experiencing cyberbullying, and increasing prevalence rates (Zhu et al., 2021).

Therefore, grounded in an analysis of data sourced from the Global School-based Health Survey (GSHS), this study seeks to discover predictors of cyberbullying victimisation and to determine if the GSHS could be used to develop an effective cyberbullying victimisation predictive model. To do so, this investigation has two primary research questions that guide the trajectory of inquiry.

* Corresponding author.
E-mail addresses: p.vieirabraga@uos.ac.uk (P.R.V. Braga), k.tyrrell@uos.ac.uk (K.R. Tyrrell).

https://doi.org/10.1016/j.mlwa.2025.100646
Received 7 October 2024; Received in revised form 11 February 2025; Accepted 24 March 2025
Available online 26 March 2025

Table 1
Summary of participants by country.

Country	Country's overall response rate	Original sample		Random sample	
		Number of participants	Reported cyberbullying victimisation	Number of participants	Reported cyberbullying victimisation
Argentina	63 %	56 981	21 %	1 048	21 %
Bolivia	79 %	7 931	20 %	1 048	19 %
Panama	71 %	2 948	15 %	1 048	12 %
Saint Vincent and the Grenadines	78 %	1 877	16 %	1 048	14 %

RQ1. *What variables in the Global School-based Health Survey are predictors of cyberbullying victimisation?*

This research question aims to build upon existing literature to identify predictors of cyberbullying victimisation, drawing from the rich repository of the GSHS. Data from surveys conducted between 2017 and 2021 will be analysed, with only those countries that included the recently added cyberbullying question in their survey being considered. Countries who participated in the GSHS will be given equal representation in the analysed data, such that the predictors identified would be shared across countries, enabling a more concise understanding of cyberbullying victimisation among adolescents.

As there is emerging evidence around predicting cyberbullying, research tends to focus on predicting perpetration (Al-Garadi et al., 2019; Barlett et al., 2016) with less research on predicting victimisation (Navarro & Jasinski, 2012). While predicting perpetration is valuable to supporting prevention strategies, predicting victimisation has the potential to be valuable in victim-support strategies and could benefit from further research. Furthermore, using an established instrument, such as the GSHS, to predict victimisation could reduce implementation barriers, supporting adoption, and compliment current efforts aimed at addressing cyberbullying. Therefore, this study will address this gap in the literature by answering the second research question.

RQ2. *How could the Global School-based Health Survey be used with machine learning algorithms to predict cyberbullying victimisation?*

Using machine learning algorithms, this research question entails the training and testing of different predictive models to determine if it is possible to use GSHS variables to predict cyberbullying victimisation, thereby offering a pathway for early intervention.

In the pursuit of understanding and mitigating cyberbullying victimisation, this study resonates with the evolving digital landscape and aims to contribute not only to the academic discourse but also to the lives of adolescents who navigate the intricate tapestry of digital spaces.

2. Material and methods

Overall, this study used a quantitative, experimental research design to answer the research questions. Specifically, the first research question used a multivariable logistic regression to identify the predictors of cyberbullying victimisation. The second research question involved training machine learning algorithms to predict cyberbullying victimisation, which is an experimental design that followed the methodology suggested by Fernandez-Lozano et al. (2016), whereby the dataset is defined, data pre-processing occurs, followed by model learning, and finally the best model is selected and evaluated. By following this method, each stage of the experiment can be reproduced, enhancing the likelihood that the same results would be obtained if the study were repeated under similar conditions. In terms of a theoretical framework, this study modified the proposed conceptual framework by Hasan et al. (2021) to concentrate on cyberbullying victimisation and included literature supported variables of the GSHS that were associated with traditional bullying victimisation, assuming that there may be similar predictors for traditional bullying victimisation and cyberbullying victimisation.

2.1. Participants

The participants of this study included school-going adolescents between 13–17 years of age who completed the Global School-based Health Survey between 2017 and 2021. The GSHS was funded by the World Health Organisation (WHO) and the United States Centers for Disease Control and Prevention (CDC) and primarily investigated by each country's ministry for education or health. It was a self-administered questionnaire that followed a two-stage cluster design to produce data representative of all school-going students in each country. The first stage selected schools with probability proportional to enrolment size, and the second stage randomly selected classes where all students were eligible to participate (WHO, 2023b).

15 countries completed the GSHS during the period. However, only four of those countries included questions relating to cyberbullying, namely: Argentina, Panama, Bolivia, and Saint Vincent and the Grenadines. Therefore, participants in those countries were included in this study. As the GSHS sampling strategy covers the necessary approach for the results to represent the wider population in each country, this study used a simple random sampling strategy that enabled the selection of a sample that was consistent with the original responses but allowed for equal representation among countries, addressing the heterogeneity between countries and enhancing the reliability of the analysis. To achieve this, the sample size for the country with the largest number of participants was determined using the formula suggested by Daniel and Cross (2019). The country with the most participants was Argentina (56, 981), and therefore, the determined sample size needed for representation was 1048, with a confidence level of 95 % that the real value is within ± 3 % of the measured value. As all other countries exceeded this number of participants, a simple random sample of 1048 was selected from each country and used to answer the research questions of this study. Table 1 provides a summary comparing the countries original responses and rate of cyberbullying to the random samples selected.

2.2. Instrument, measures, and the theoretical framework

Participants completed a self-administered questionnaire designed to provide accurate data on health behaviours and protective factors. Countries developed their country-specific questionnaire by selecting all or some of the standardised set of questions (WHO, 2023b). The use of standardised questions with fewer response options has been evidenced to improve the reliability of the instrument (Ruel et al., 2015), which is important for comparing different participants as is conducted in this study.

The conceptual framework proposed by Hasan et al. (2021) was modified to answer the research questions, and Fig. 1 illustrates the modified, preliminary framework, concentrating on variables with possible associations to cyberbullying victimisation. The modifications included additional variables into some of the themes, the creation of a new theme, and a sole dependent variable: participants reporting cyberbullying victimisation. The new theme created was "Demographic factors", as age and sex were found to be predictors of bullying victimisation (Biswas et al., 2020; Peltzer & Pengpid, 2020). Additionally, sitting activities and sexual debut were included in the theme "Lifestyle

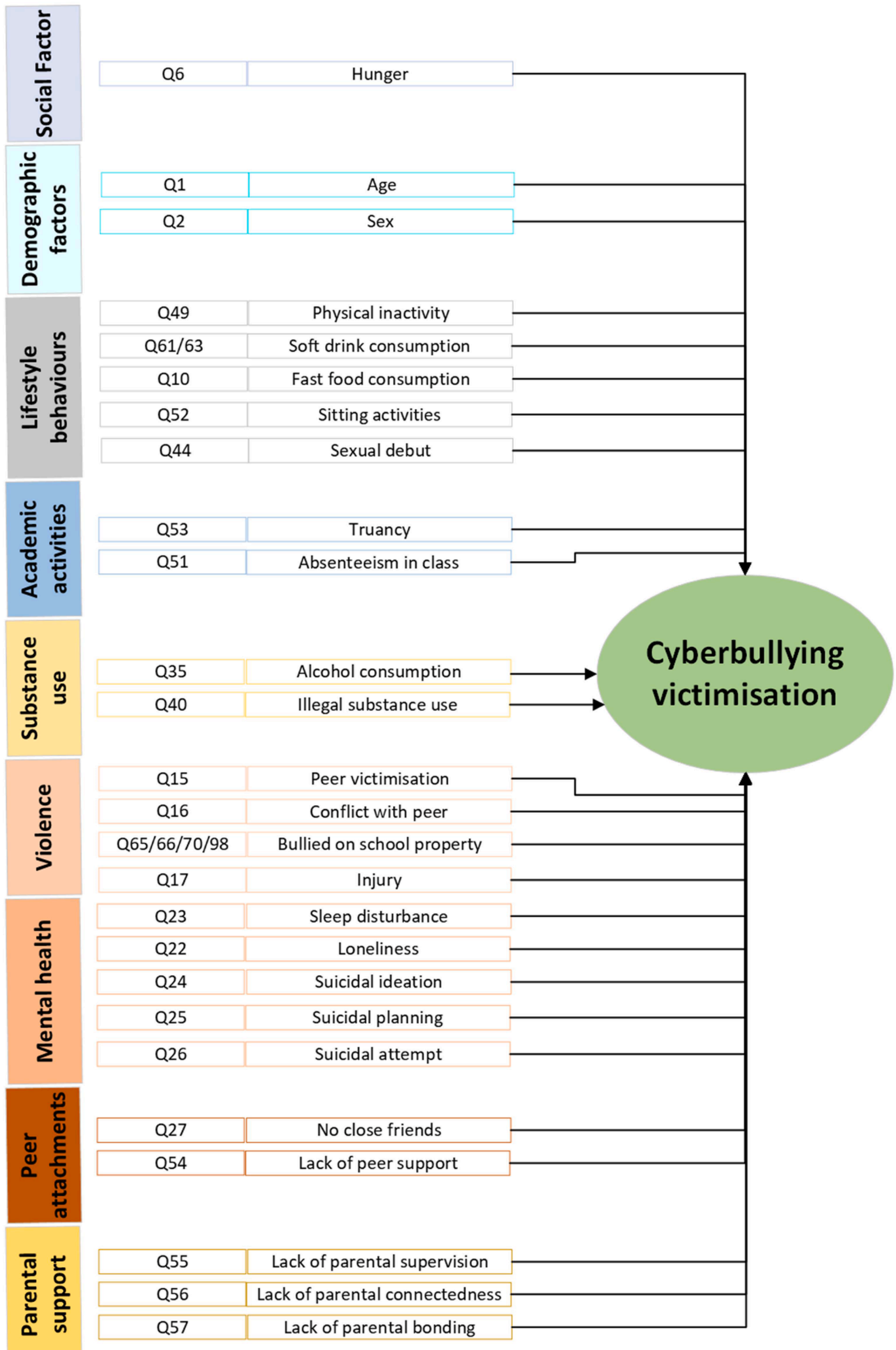

Fig. 1. Preliminary framework for cyberbullying victimisation inspired by Hasan et al. (2021) conceptual framework.

behaviours" (Peltzer & Pengpid, 2020; Vismara et al., 2022), illegal substance use in "Substance use" (Pengpi & Peltzer, 2021), bullying victimisation and injury in "Violence" (Peltzer & Pengpid, 2020; Pengpi & Peltzer, 2021), and suicidal ideation, planning, and attempt in "Mental health" (Abdirahman et al., 2012; Cheng et al., 2010; Hasan et al., 2021; Owusu et al., 2011; Peltzer & Pengpid, 2020; Pengpi & Peltzer, 2021). However, the variable relating to tobacco consumption was excluded from "Substance use", as not all countries in this study asked this question to participants. Further information about the literature and each measure is provided in the supplementary information, including the standard variable name, the questions asked, and the response options.

2.3. Data analysis

The data for each country was collected from the WHO repository. Following this, all variables used in this study were selected, any missing responses were removed, and a random sample of 1048 responses were drawn from each country using a computerised random sampling software. Once all the data was collated into a single dataset, the data was

Table 2
Variables predicting likelihood of reporting cyberbullying victimisation.

Variables	Total adolescents	Number of adolescents	%	p	Odds ratio	95 % CI for odds ratio	
						Lower	Upper
Social factor							
Hunger	4192	127	3.0 %	0.743	0.92	0.57	1.50
Demographic factors							
Age	4192	1722	41.1 %	0.604	1.05	0.87	1.27
Sex	4192	1891	45.1 %	0.000***	0.57	0.47	0.69
Lifestyle behaviours							
Physical inactivity	4192	3087	73.6 %	0.306	0.90	0.73	1.10
Soft drink consumption	4192	2082	49.7 %	0.672	0.96	0.80	1.15
Fast food consumption	4192	2211	52.7 %	0.030*	1.22	1.02	1.47
Sitting activities	4192	1061	25.3 %	0.072	1.20	0.98	1.46
Sexual debut	4192	1580	37.7 %	0.000***	1.50	1.22	1.83
Academic activities							
Truancy	4192	1198	28.6 %	0.379	1.09	0.90	1.33
Absenteeism in class	4192	2961	70.6 %	0.058	0.83	0.68	1.01
Substance use							
Alcohol consumption	4192	1523	36.3 %	0.001**	1.40	1.15	1.71
Illegal substance use	4192	622	14.8 %	0.019*	0.73	0.56	0.95
Violence							
Peer victimisation	4192	788	18.8 %	0.000***	1.55	1.25	1.94
Conflict with peer	4192	984	23.5 %	0.025*	1.28	1.03	1.60
Bullied on school property	4192	869	20.7 %	0.000***	3.58	2.96	4.34
Injury	4192	1767	42.2 %	0.316	1.10	0.91	1.33
Mental health							
Sleep disturbance	4192	513	12.2 %	0.118	1.22	0.95	1.58
Loneliness	4192	751	17.9 %	0.033*	1.29	1.02	1.62
Suicidal ideation	4192	877	20.9 %	0.004**	1.49	1.14	1.94
Suicidal planning	4192	730	17.4 %	0.048*	1.33	1.00	1.75
Suicidal attempt	4192	608	14.5 %	0.160	0.81	0.61	1.08
Peer attachments							
No close friend	4192	273	6.5 %	0.026*	0.65	0.44	0.95
Lack of peer support	4192	2505	59.8 %	0.373	1.09	0.90	1.32
Parental support							
Lack of parental supervision	4192	2720	64.9 %	0.417	1.09	0.88	1.35
Lack of parental connectedness	4192	2640	63.0 %	0.509	1.08	0.86	1.34
Lack of parental bonding	4192	2148	51.2 %	0.508	1.07	0.88	1.30
Constant				0.000	0.07		

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$
Note: sex is for males compared to females, and age is for younger compared to older.

re-coded. For consistency, the coding was in line with the approach of Hasan et al. (2021), and the additional variables followed the same binary coding system. For a comprehensive view of the themes, measures, and coding, please refer to the supplementary information in the supplementary information.

2.3.1. RQ1: what variables in the global school-based health survey are predictors of cyberbullying victimisation?

Using the SPSS statistical software package, a multivariable logistic regression was applied to the data to answer the research question. This involved ensuring all the assumptions of the regression were met, testing the model's statistical significance, and identifying the amount of variance explained in the model by the independent variables. To evaluate the model, metrics, including specificity, sensitivity, negative predictive value (NPV), positive predictive value (PPV), and accuracy were analysed, and the area under the receiver operating characteristic curve (AUROC) was used to determine the model's overall measure of discrimination (Royston & Altman, 2010). Following the evaluation of the model, the variables in the equation were assessed by their statistical significance and odds ratio. Finally, the preliminary framework was adjusted, confirmed, and visualised based on the significant predictors identified.

2.3.2. RQ2: how could the global school-based health survey be used with machine learning algorithms to predict cyberbullying victimisation?

To answer this research question, Python and related data science libraries were used for statistical analyses, guided by the approach suggested by Fernandez-Lozano et al. (2016). Therefore, the predictors identified in RQ1 satisfied the feature reduction requirement and were used to train multiple classification algorithms, namely: decision tree classifier, random forest classifier, and XGBoost classifier. These algorithms were selected for their interpretability and transparency (Quinlan, 1986), strong performance in classification tasks (Breiman, 2001), and efficiency and robustness to counter overfitting through regularisation (Chen & Guestrin, 2016). As part of the training, all algorithms underwent a 10-fold cross validation hyperparameter tuning exercise to find the best parameters for each algorithm, and due to the imbalanced nature of the dataset, each algorithm was trained on both random over-sampling and random under-sampling (Hasanin & Khoshgoftaar, 2018; Hayaty et al., 2020). The training was performed on 75 % of the sample, and the testing on the remaining 25 % of the sample. After training, all models were assessed and evaluated with the entire sample according to their accuracy, weighted average f1-score, AUROC, and Matthews correlation coefficient (MCC). Lastly, the best model's parameters were mentioned, and the feature importance was visualised.

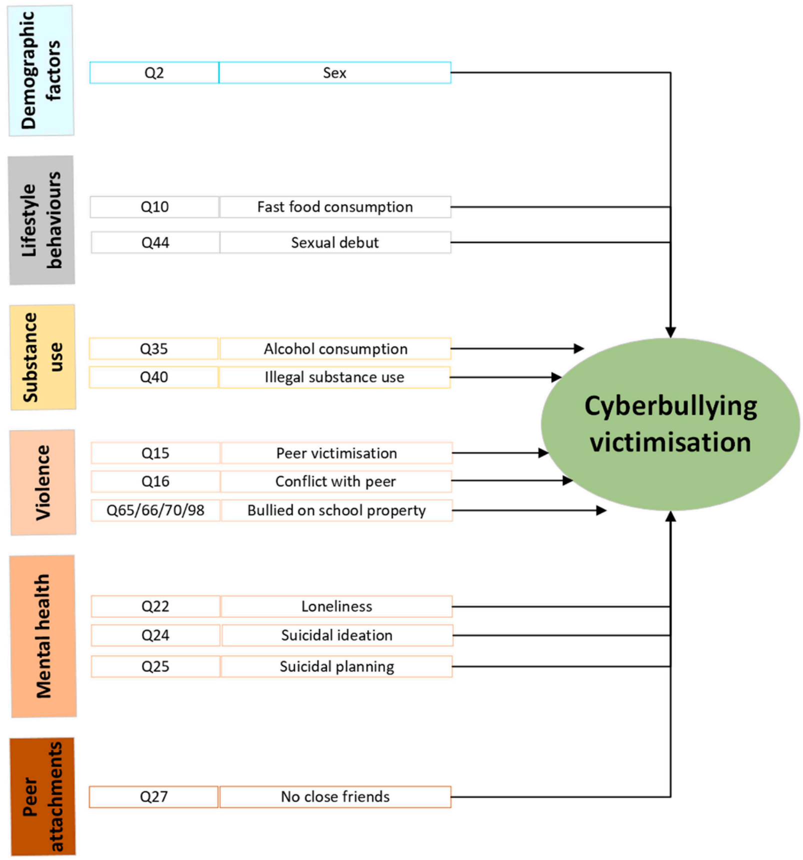

Fig. 2. The final framework with the statistically significant predictors.

3. Results

3.1. *RQ1: what variables in the Global School-Based Health Survey are predictors of cyberbullying victimisation?*

A multivariable logistic regression was performed to ascertain the effects of the GSHS variables on the likelihood that participants reported cyberbullying victimisation. All the required assumptions of the regression were met, and in terms of assessing the goodness of fit, a Hosmer and Lemeshow test was conducted, which was not statistically significant ($p = 0.129$), indicating that the model is not a poor fit. When compared to a baseline model that assumed all participants reported no cyberbullying victimisation, the logistic regression model was statistically significant ($\chi^2(26)=507.96,\ p < 0.001$) and correctly classified 84.1 % of the cases compared to the baseline model's 83.6 %. Sensitivity was 15.0 %, specificity was 97.6 %, PPV was 55.1 % and NPV was 85.4 %. The model explained 19.3 % (Nagelkerke R^2) of the variance in reported cyberbullying victimisation, and the AUROC was 0.*758* (*95 % CI, 0.739 to 0.778*), which is an acceptable level of discrimination according to Hosmer Jr et al. (2013). Of the 26 predictor variables, 12 were statistically significant, as shown in Table 2. Adolescents who reported being bullied on school property had 3.58 times higher odds of reporting cyberbullying victimisation than those who did not report being bullied on school property, the highest odds ratio of all the variables. In terms of the other variables, females, those who reported peer victimisation, sexual debut, consuming alcohol, and suicidal ideation were ±1.5 times more likely to report being cyberbullied.

3.1.1. *Final framework based on results*

In total, this study found 12 variables in the preliminary framework that were statistically significant predictors of cyberbullying victimisation, which is across four countries with equal representation in the sample. Fig. 2 illustrates the final framework with only the significant predictors.

Table 3
Model comparisons based on evaluation metrics.

		Evaluation metrics for cyberbullying victimisation			
Model	Sampling strategy	Accuracy	Weighted Avg. F1-score	AUROC	MCC
Random Forest Classifier	Random under-sampling	69 %	0.73	0.76	0.31
	Random over-sampling	82 %	0.83	0.83	0.44
Decision Tree Classifier	Random under-sampling	68 %	0.72	0.75	0.29
	Random over-sampling	77 %	0.79	0.79	0.37
XGBoost Classifier	Random under-sampling	71 %	0.75	0.77	0.32
	Random over-sampling	79 %	0.81	0.82	0.41

3.2. *RQ2: how could the Global School-Based Health Survey be used with machine learning algorithms to predict cyberbullying victimisation?*

Multiple machine learning algorithms were trained on the predictors identified in RQ1, 12 in total, and they were trained on both random under-sampling and random over-sampling due to the imbalanced nature of the dataset. The training was performed on 75 % of the sample, and the testing on the remaining 25 % of the sample. Of the models

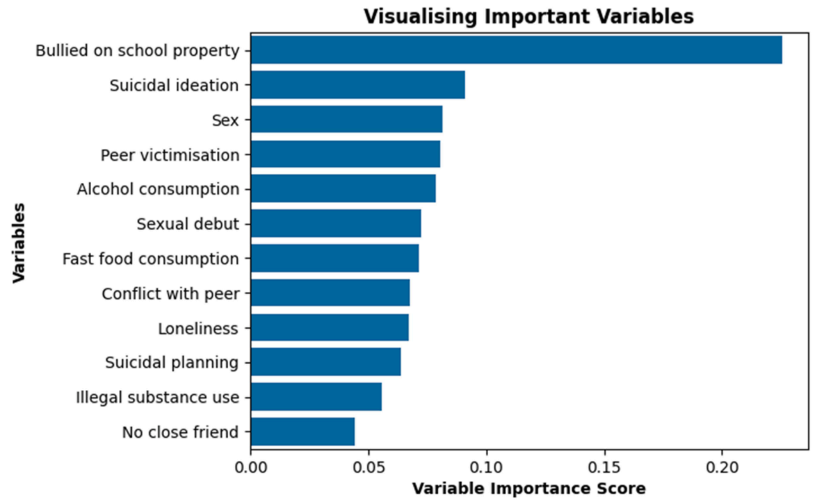

Fig. 3. Visualising important variables of the over-sampled random forest classifier.

trained, the over-sampled random forest classifier scored the highest in all of the evaluation metrics (see Table 3 for a comprehensive view of all models and their evaluation metrics). The model predicted classes with 82 % accuracy, and it had an MCC of *0.44* and an AUROC of *0.83 (95 % CI, 0.81 to 0.85)*, indicating the model has moderate to strong predictive power and an excellent level of discrimination between classes according to Hosmer Jr et al. (2013). Furthermore, the best parameters for this model were using the square root of the features when splitting (without bootstrapping), 10 for the maximum depth of the tree, one as the minimum number of samples at a leaf node, two as the minimum number of samples to split an internal node, and 50 decision trees in the random forest.

In terms of feature importance, the model used all 12 variables for the predictions, with bullied on school property being the most important feature (importance score: 0.*23*), more than double the second most important feature, suicidal ideation (importance score: 0.*09*). See Fig. 3 for an illustration of the important variables. Conclusively, a predictive model with 83 % accuracy and an excellent level of discrimination was developed using 12 variables from the GSHS across four countries that were equally represented in the sample.

4. Discussion and conclusion

This study explored the intricate landscape of cyberbullying victimisation among adolescents, using data from the GSHS and employing both traditional statistical analysis and advanced machine learning techniques. The investigation shed light on several key aspects of cyberbullying victimisation, uncovering predictors and potential predictive models for interventions.

Similar to the work relating to bullying victimisation of Hasan et al. (2021), this study found peer victimisation, alcohol consumption, suicidal ideation and planning, conflict with peers, no close friends, and loneliness to be predictors of cyberbullying victimisation, which is consistent with other studies on cyberbullying (Brewer & Kerslake, 2015; Şahin, 2012; Van Geel et al., 2014). Furthermore, being bullied on school property was the predictor with the highest odds ratio (OR=3.58), meaning adolescents who were bullied on school property were more than three times more likely to report cyberbullying victimisation. This corroborates with the findings of Beran and Li (2007) and Baldry et al. (2015), however, another study indicated little overlap between cyberbullying victims and school bullying victims (Kubiszewski et al., 2015), highlighting the nuances related to bullying predictors. Furthermore, females were more likely to report cyberbullying victimisation and age was not a statistically significant predictor, which is consistent with the findings of a meta-analysis on the predictors of cyberbullying by Guo (2016). However, this further emphasises the differences between school bullying and cyberbullying, as males and younger students were found to be more likely bullied in school (Biswas et al., 2020; Peltzer & Pengpid, 2020). Conversely, similarities are also prevalent, especially in terms of the GSHS variables, as factors such as sexual debut, fast food consumption, and illegal substance use were found to be predictors of cyberbullying victimisation in this study, which corroborate with findings relating to bullying victimisation (Hasan et al., 2021; Peltzer & Pengpid, 2020; Pengpi & Peltzer, 2021).

As for predicting cyberbullying victimisation using the GSHS, the presented models, particularly the random forest classifier, offer a pathway toward early intervention and support for victims, bridging the gap between adolescents' hesitance in reporting cyberbullying and the need for timely assistance. The best performing model achieved 82 % accuracy and had an excellent level of discrimination. Interestingly, the model used all 12 of the statistically significant predictors identified in RQ1. Hypothetically, this allows for a scenario where adolescents could be asked 12 questions, and using the model, one would be able to predict if they were to report cyberbullying victimisation with 82 % accuracy. This has the potential to provide support to victims proactively, especially considering adolescents tend to resist reporting cyberbullying and prefer to confide in their peers, or no one at all (Connolly, 2018; Slonje & Smith, 2008).

However, the ethical implications arising from of incorrect predictions need to be considered. In instances where the model incorrectly identifies an adolescent as at risk, there is a possibility of unwarranted interventions. Such misclassifications may lead to unnecessary stigmatisation or anxiety, potentially affecting the individual's psychological wellbeing (Strindberg et al., 2020). Conversely, failing to flag those who are genuinely at risk could result in missed opportunities for early support and intervention, thereby leaving vulnerable individuals without the necessary assistance. Given these implications, it is imperative that any deployment of predictive models, such as the one developed in this study, is accompanied by robust ethical oversight. This could include integrating the model within an established support framework, such as the Positive Behavioural Interventions and Supports framework (Bradshaw, 2013; Lawrence, 2017). This integration enables clear ethical guidelines for practitioners and policymakers regarding the responsible and effective use of predictive analytics in the context of cyberbullying victimisation.

In terms of this study's limitations, there are two significant areas that need to be emphasised. Firstly, the GSHS is a self-administered questionnaire, and therefore has self-reported measures which could be prone to biases or inaccuracies (Campbell & Fiske, 1959). Secondly, self-reported measures relating to sensitive topics, which could arguably include cyberbullying, have been found to experience common misreporting (Tourangeau & Yan, 2007). Both need to be considered carefully when interpreting the results of this study and when identifying ways to further research in this area. As for possible improvements to the predictive models, modifying the way variables are coded may improve performance. In this study, all variables were coded in a dichotomous manner to remain consistent with previous research, however, this might limit the predictive power of the variables. For example, loneliness was originally a five-point scale response, where one was "never" and five was "always", keeping the original scale could provide more variation for the algorithms to adjust, potentially enhancing its calculations.

In conclusion, this study contributes to our understanding of cyberbullying victimisation among adolescents by identifying predictors and using machine learning for prediction. The findings advocate for a comprehensive approach to intervention, aiming to foster a safer online environment for the younger generation through victim support. Acknowledging the limitations of the study, further research is encouraged to build upon this work, refining models, investigating contextual nuances, including more countries, and ultimately striving for a world where adolescents can receive timely and effective victim-centred support after experiencing online harms.

CRediT authorship contribution statement

Paulo Ricardo Vieira Braga: Conceptualization, Data curation, Methodology, Formal analysis, Investigation, Writing – original draft, Visualization. **Katie Rose Tyrrell:** Conceptualization, Methodology, Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.mlwa.2025.100646.

Data availability

To access the data, a WHO registration process needs to be followed.

References

Abdirahman, H. A., Bah, T., Shrestha, H., & Jacobsen, K. H. (2012). Bullying, mental health, and parental involvement among adolescents in the Caribbean. *The West Indian Medical Journal, 61*(5), 504–508. https://doi.org/10.7727/wimj.2012.212

Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., … Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges. *IEEE Access : Practical Innovations, Open Solutions, 7*, 70701–70718. https://doi.org/10.1109/ACCESS.2019.2918354

Baldry, A. C., Farrington, D. P., & Sorrentino, A. (2015). Am I at risk of cyberbullying"? A narrative review and conceptual framework for research on risk of cyberbullying and cybervictimization: The risk and needs assessment approach. *Aggression and Violent Behavior, 23*, 36–51. https://doi.org/10.1016/j.avb.2015.05.014

Barlett, C. P., Gentile, D. A., & Chew, C. (2016). Predicting cyberbullying from anonymity. *Psychology of Popular Media Culture, 5*(2), 171. https://doi.org/10.1037/ppm0000055

Beran, T., & Li, Q. (2007). The relationship between cyberbullying and school bullying. *The Journal of Student Wellbeing, 1*(2), 16–33. https://doi.org/10.21913/JSW.v1i2.172

Biswas, T., Scott, J. G., Munir, K., Thomas, H. J., Huda, M. M., Hasan, M. M., … Mamun, A. A. (2020). Global variation in the prevalence of bullying victimisation amongst adolescents: Role of peer and parental supports. *EClinicalMedicine*, 20. https://doi.org/10.1016/j.eclinm.2020.100276

Bradshaw, C. P. (2013). Preventing bullying through positive behavioral interventions and supports (PBIS): A multitiered approach to prevention and integration. *Theory into Practice, 52*(4), 288–295. https://doi.org/10.1080/00405841.2013.829732

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Brewer, G., & Kerslake, J. (2015). Cyberbullying, self-esteem, empathy and loneliness. *Computers in Human Behavior, 48*, 255–260. https://doi.org/10.1016/j.chb.2015.01.073

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81. https://doi.org/10.1037/h0046016

CDC. (2023). *National vital statistics system, mortality 2018-2021 on cdc wonder online database, released in 2021* (2021). National Center for Health Statistics.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD, 16*, 785–794. https://doi.org/10.1145/2939672.2939785

Cheng, Y., Newman, I. M., Qu, M., Mbulo, L., Chai, Y., Chen, Y., & Shell, D. F. (2010). Being bullied and psychosocial adjustment among middle school students in China. *Journal of School Health, 80*(4), 193–199. https://doi.org/10.1111/j.1746-1561.2009.00486.x

Connolly, J. P. (2018). Exploring the factors influencing gifted adolescents' resistance to report experiences of cyberbullying behavior: Toward an improved understanding. *Journal for the Education of the Gifted, 41*(2), 136–159. https://doi.org/10.1177/0162353218763869

Cowie, H. (2013). Cyberbullying and its impact on young people's emotional health and well-being. *The Psychiatrist, 37*(5), 167–170. https://doi.org/10.1192/pb.bp.112.040840

Daniel, W. W., & Cross, C. L. (2019). *Biostatistics: A foundation for analysis in the health sciences* (11th edition). Wiley.

Fernandez-Lozano, C., Gestal, M., Munteanu, C. R., Dorado, J., & Pazos, A. (2016). A methodology for the design of experiments in computational intelligence with multiple regression models. *PeerJ, 4*, e2721. https://doi.org/10.7717/peerj.2721

Guo, S. (2016). A meta-analysis of the predictors of cyberbullying perpertration and victimization. *Psychology in the Schools, 53*(4), 432–453. https://doi.org/10.1002/pits.21914

Hasan, M. M., Fatima, Y., Pandey, S., Tariqujjaman, M., Cleary, A., Baxter, J., & Mamun, A. A. (2021). Pathways linking bullying victimisation and suicidal behaviours among adolescents. *Psychiatry Research, 302*. https://doi.org/10.1016/j.psychres.2021.113992

Hasanin, T., & Khoshgoftaar, T. (2018). The Effects of Random Undersampling with Simulated Class Imbalance for Big Data. In *2018 IEEE International Conference on Information Reuse and Integration (IRI), July*.

Hayaty, M., Muthmainah, S., & Ghufran, S. M. (2020). Random and synthetic over-sampling approach to resolve data imbalance in classification. *International Journal of Artificial Intelligence Research, 4*(2), 86–94. https://doi.org/10.29099/ijair.v4i2.152

Hosmer, D. W., Jr, Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression (Vol. 398)*. John Wiley & Sons.

Katz, J.E., & Aakhus, M. (2002). *Perpetual contact: Mobile communication, private talk, public performance* (J. E. Katz & M. Aakhus, Eds.). Cambridge University Press. https://doi.org/10.1017/CBO9780511489471.

Kowalski, R. M., & Limber, S. P. (2013). Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health, 53*(1), S13–S20. https://doi.org/10.1016/j.jadohealth.2012.09.018

Kubiszewski, V., Fontaine, R., Potard, C., & Auzoult, L. (2015). Does cyberbullying overlap with school bullying when taking modality of involvement into account? *Computers in Human Behavior, 43*, 49–57. https://doi.org/10.1016/j.chb.2014.10.049

Lawrence, S. (2017). Bullying in secondary schools: Action planning using a positive behavior intervention and support framework. *American Secondary Education, 45*(2), 85–92. http://www.jstor.org/stable/45147897.

Navarro, J. N., & Jasinski, J. L. (2012). Going cyber: Using routine activities theory to predict cyberbullying experiences. *Sociological Spectrum, 32*(1), 81–94. https://doi.org/10.1080/02732173.2012.628560

Owusu, A., Hart, P., Oliver, B., & Kang, M. (2011). The Association Between Bullying and Psychological Health Among Senior High School Students in Ghana, West Africa. *Journal of School Health, 81*(5), 231–238. https://doi.org/10.1111/j.1746-1561.2011.00590.x

Patchin, J. W., & Hinduja, S. (2012). Cyberbullying: An update and synthesis of the research. *Cyberbullying prevention and response* (pp. 13–35). Routledge.

Peltzer, K., & Pengpid, S. (2020). Prevalence of bullying victimisation and associated factors among in-school adolescents in Mozambique. *Journal of Psychology in Africa, 30*(1), 64–68. https://doi.org/10.1080/14330237.2020.1712809

Pengpi, S., & Peltzer, K. (2021). *Prevalence and correlates of frequent and infrequent bullying victimization among school adolescents from five Southeast Asian countries*. https://doi.org/10.7454/msk.v25i2.1282

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106. https://doi.org/10.1007/bf00116251

Royal College of Paediatrics and Child Health. (2020). *State of child health*. RCPCH. Retrieved Aug 24 from https://stateofchildhealth.rcpch.ac.uk/evidence/mortality/adolescent-mortality/.

Royston, P., & Altman, D. G. (2010). Visualizing and assessing discrimination in the logistic regression model. *Statistics in Medicine, 29*(24), 2508–2520. https://doi.org/10.1002/sim.3994

Ruel, E., Wagner, W. E., & Gillespie, B. J. (2015). *The practice of survey research*. SAGE Publications. https://books.google.co.uk/books?id=JdY5DQAAQBAJ.

Şahin, M. (2012). The relationship between the cyberbullying/cybervictmization and loneliness among adolescents. *Children and Youth Services Review, 34*(4), 834–837. https://doi.org/10.1016/j.childyouth.2012.01.010

Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology, 49*(2), 147–154. https://doi.org/10.1111/j.1467-9450.2007.00611.x

Smith, P. K. (2012). Cyberbullying and cyber aggression. *Handbook of school violence and school safety* (pp. 93–103). Routledge.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry, 49*(4), 376–385. https://doi.org/10.1111/j.1469-7610.2007.01846.x

Strindberg, J., Horton, P., & Thornberg, R. (2020). The fear of being singled out: Pupils' perspectives on victimisation and bystanding in bullying situations. *British Journal of Sociology of Education, 41*(7), 942–957. https://doi.org/10.1080/01425692.2020.1789846

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859. https://doi.org/10.1037/0033-2909.133.5.859

Van Geel, M., Vedder, P., & Tanilon, J. (2014). Relationship between peer victimization, cyberbullying, and suicide in children and adolescents: A meta-analysis. *JAMA Pediatrics, 168*(5), 435–442. https://doi.org/10.1001/jamapediatrics.2013.4143

Vismara, M., Girone, N., Conti, D., Nicolini, G., & Dell'Osso, B. (2022). The current status of cyberbullying research: A short review of the literature. *Current Opinion in Behavioral Sciences, 46*, Article 101152. https://doi.org/10.1016/j.cobeha.2022.101152

WHO. (2023a). *Adolescent and young adult health*. World Health Organisation. Retrieved 9 Aug 2023 from https://www.who.int/news-room/fact-sheets/detail/adolescents-health-risks-and-solutions.

WHO. (2023b). *GSHS methodology*. Retrieved 16 March 2023 from https://www.who.int/teams/noncommunicable-diseases/surveillance/systems-tools/global-school-based-student-health-survey/methodology.

Zhu, C., Huang, S., Evans, R., & Zhang, W. (2021). Cyberbullying among adolescents and children: A comprehensive review of the global situation, risk factors, and preventive measures. *Frontiers in Public Health, 9*, Article 634909. https://doi.org/10.3389/fpubh.2021.634909