Current Psychology (2024) 43:10117–10121
https://doi.org/10.1007/s12144-023-05174-z

Reconsidering False Positives in Machine Learning Binary Classification Models of Suicidal Behavior

E. F. Haghish[1] · Nikolai Czajkowski[1,2]

Accepted: 23 August 2023 / Published online: 31 August 2023

Abstract
We posit the hypothesis that False Positive cases (FP) in machine learning classification models of suicidal behavior are at risk of suicidal behavior and should not be seen as sheer classification error. We trained an XGBoost classification model using survey data from 173,663 Norwegian adolescents and compared the classification groups for several suicide-related mental health indicators, such as depression, anxiety, psychological distress, and non-suicidal self-harm. The results showed that as the classification is made at higher risk thresholds - corresponding to higher specificity levels - the severity of anxiety and depression symptoms of the FP and True Positive cases (TP) become significantly more similar. In addition, psychological distress and non-suicidal self-harm were found to be highly prevalent among the FP group, indicating that they are indeed at risk. These findings demonstrate that FP are a relevant risk group for potential suicide prevention programs and should not be dismissed. Although our findings support the hypothesis, we account for limitations that should be examined in future longitudinal studies. Furthermore, we elaborate on the rationale of the hypothesis, potential implications, and its applicability to other mental health outcomes.

Keywords Suicide risk assessment · Adolescence suicide prevention · Supervised machine learning · False positive · Classification error · Psychometrics

Introduction

Several recent attempts have been made to classify suicidal behavior using machine learning (Burke et al., 2020; Miché et al., 2020; Shen et al., 2020; van Vuuren et al., 2021). In this paper, we point out a critical issue that has not been addressed in the literature and contrasts the common understanding of the False Positive cases (FP), which are considered as non-informative classification error (see for example Linthicum et al., 2019; van Vuuren et al., 2021). In a nutshell, when evaluating machine learning binary classification models of suicidal behavior, a closer look should be given to FP. Our argument is that FP may exhibit similar psycho-socio-behavioral response patterns to True Positive cases (TP) and may therefore include individuals with a high risk of developing suicidal tendencies. Further, it is known that individuals with suicidal tendencies may avoid reporting their suicide attempts or ideations, which is particularly common among adolescents (Brahmbhatt & Grupp-Phelan, 2019; Christl et al., 2006; Hart et al., 2013; Jones et al., 2019). Thus, it is plausible that a well-trained model can identify high-risk individuals who have not yet attempted suicide or who have refused to self-report their previous attempts. This group may be a clinically relevant target[1] for prevention programs.

Consider an imperfect binary classifier that categorizes individuals into two groups of positives and negatives, of which TP and True Negative cases (TN) are correct classifications, and FP and False Negative cases (FN) constitute misclassifications. Machine learning classification models of suicidal behavior are typically trained using multiple mental health indicators

E. F. Haghish
haghish@uio.no

[1] Department of Psychology, University of Oslo, P.O. Box 1094, 0317 Oslo, Blindern, Norway

[2] Department of Mental Disorders, Norwegian Institute of Public Health, Oslo, Norway

[1] Here we consider both adolescents who are at immediate risk of attempting suicide as well as those who are at risk of attempting suicide in the more distant future as *relevant target group* for clinical suicide intervention or prevention (see Carter & Spittal, 2018; Granello, 2010).

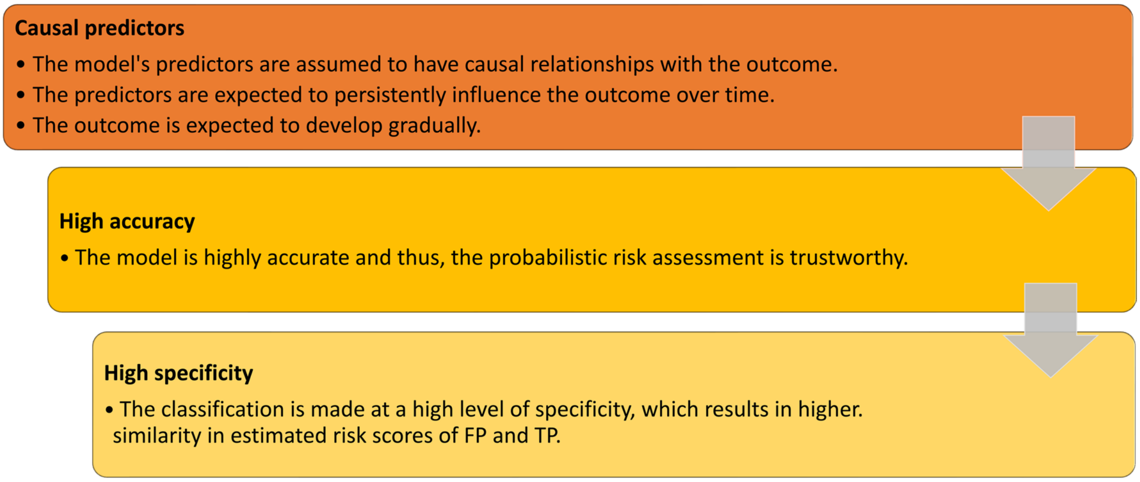
Causal predictors
• The model's predictors are assumed to have causal relationships with the outcome.
• The predictors are expected to persistently influence the outcome over time.
• The outcome is expected to develop gradually.
High accuracy
• The model is highly accurate and thus, the probabilistic risk assessment is trustworthy.
High specificity
• The classification is made at a high level of specificity, which results in higher.
similarity in estimated risk scores of FP and TP.

Fig. 1 The rationale for considering FP as a risk group in suicide attempt classification models

and risk factors and can detect patterns in the data that lead to accurate classifications (Healy, 2021; Ley et al., 2022; Walsh et al., 2017). Additionally, some of the common predictors and risk factors of suicidal behavior, such as depressive symptoms, non-suicidal self-harm, and substance use, might mediate or causally relate to the development of suicidal behavior, which are also expected to persist over time. Persistence of risk factors and lack of protective factors will increase the risk for developing suicidal behavior in the future. Furthermore, suicidal behavior is expected to develop over time, as there is evidence for pathways that lead to development of suicidality among adolescents (Haghish et al., 2023; Van Orden et al., 2010). Therefore, if the classification model demonstrates high accuracy and the classification is made for high cutoff in the estimated probabilistic risk scores (high specificity), the response patterns and symptoms of FP and TP are expected to be similar (see Fig. 1). Consequently, FP are expected to have worse mental health conditions compared to TN, providing evidence that FP are a risk group and potentially relevant to a suicide prevention program.

Drawing on this rationale, we tested two hypotheses using cross-sectional data. First, we hypothesized that at higher specificity thresholds, the severity of depression and anxiety symptoms would be more similar between the FP and TP groups. Second, we hypothesized that at a high specificity threshold, the prevalence of psychological distress and non-suicidal self-harm would be significantly higher among the FP group compared to the TN group.

Methods

We utilized data from the Ungdata project (www.ungdata.no), which included 173,663 adolescents from all municipalities in Norway, who participated in the period between 2014 and 2019. The participants completed a battery of questionnaires that covered socio-demographic information, internalizing and externalizing problems, traumatic experiences, interpersonal relationships, and suicidal behavior. Following the approach proposed by Strand et al., (2003), we computed the psychological distress score as the average of the depression and anxiety sum scores, and classified participants who scored at least 3 out of 4 as distressed. A comprehensive description of the instruments and their items can be accessed on the Open Science repository of this paper via https://osf.io/a7fgb/.

The Extreme Gradient Boosting algorithm (XGBoost; Chen & Guestrin, 2016) was used to develop the classification model, as it is expected to outperform decision tree ensemble algorithms as well as generalized linear models (Haghish et al., 2023; Sahin, 2020). We trained the algorithm on 80% of the data (n = 138,931), with the remaining 20% (n = 34,732) reserved for testing. To optimize performance on the imbalanced outcome variable, we fine-tuned the model using 10-fold cross-validation to maximize the Area Under Precision-Recall Curve (AUCPR), which is the preferred performance metric for imbalanced (low prevalent) outcomes (Davis & Goadrich, 2006). The adjROC R package (Haghish, 2022) was employed to compute the classification cutoffs for specificity values ranging from 0.4 to 1.0, and for each threshold, we computed the severity of depression and anxiety symptoms for all classification groups. To test our first hypothesis, we subtracted the depression and anxiety scores of TP from FP at different thresholds to calculate their differences and fitted linear regression models on the results. For the second hypothesis, we compared the prevalence of psychological distress and non-suicidal self-harm between FP and TN using Fisher's exact test. Note that both the

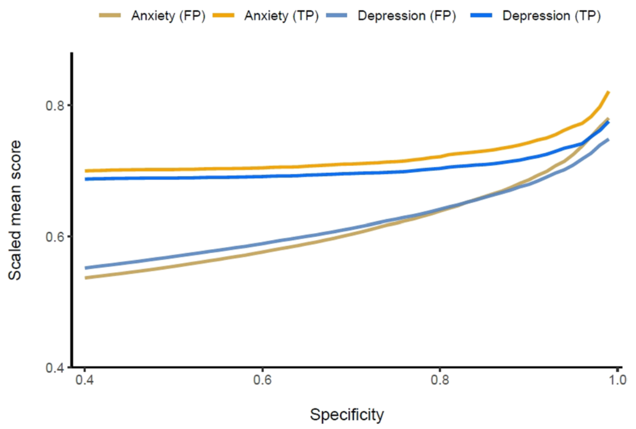

Fig. 2 Scaled mean score of anxiety and depression sum scores evaluated for specificity thresholds ranging from 0.4 to 1.0

binary psychological distress item was computed after the model training. Moreover, the non-suicidal self-harm item was excluded from the dataset in the process of model training.

Results

The trained XGBoost model[2] achieved an AUC of 92.8% and an AUPRC of 48.8%. Additionally, the model exhibited a sensitivity of 51.7%, specificity of 96.9%, and a Cohen's Kappa of 0.466. These results suggest that the model's accuracy and inter-rater agreement in classifying suicidal behavior are high.

Figure 2 illustrates that as specificity increased, the average sum scores of depression and anxiety for TP and FP groups became more similar. Linear regression analysis confirmed a significant negative slope for the distance between the TP and FP scores for depression (β = - 0.69, Adjusted R2 = 0.991, F (1, 97) = 11240, $p < 0.0001$) and anxiety (β = - 0.71, Adjusted R2 = 0.972, F (1, 14) = 3382, $p < 0.0001$). These findings indicate the difference between the mean sum scores of the two groups decreased as a function of increase in specificity, thereby supporting our first hypothesis.

In the testing sample, 7.2% of adolescents were found to experience psychological distress, with varying rates across the different classification groups. Specifically, the prevalence was highest in the TP group at 63.4%, followed by FP at 55.3%, FN at 20.0%, and TN at 4.2%. Fisher's exact test revealed a statistically significant difference in psychological distress levels between the FP and FN groups (Odds ratio = 0.036, 95% CI = 0.031 – 0.041, $p < 0.0001$). Indeed, the rate of psychological distress of the FP was also significantly higher than the TN group (Odds ratio = 4.958, 95% CI = 3.976 – 6.202, $p < 0.0001$). Additionally, the prevalence of non-suicidal self-harm was highest in the TP group at 95.2%, followed by FN at 80.8%, FP at 65.2%, and TN at 9.9%. Consistent with expectations, Fisher's exact test also revealed a statistically significant difference in non-suicidal self-harm prevalence between the FP and TN groups (Odds ratio = 0.059, 95% CI = 0.050 – 0.068, $p < 0.0001$), supporting the second hypothesis.

Discussion

The results supported our claims that for an accurate suicide classification model and at a high specificity threshold, adolescents in the FP group might be comparable to TP, showing severe symptoms of psychological distress and non-suicidal self-harm at rates much higher than those in the TN group. In our analysis, FP showed more severe signs of psychological distress than the FN group that reflects on why the model had evaluated their suicide attempt risk to be higher than the FN. Depression, anxiety, and non-suicidal self-harm are well-established risk factors for suicidal behavior and thus, the study's findings support our hypotheses and arguments (Carballo et al., 2020; Darke et al., 2010; Greening et al., 2008; Lewis et al., 2014; Lohner & Konrad, 2006; Toprak et al., 2011).

Well-trained machine learning suicide classification models are expected to identify important predictors and interactions between the predictors. In estimating suicide attempt risk, machine learning models are also expected to take the severity of these predictors into account. Therefore, it is to be expected that the severity of important suicide-related indicators such as depression, anxiety, psychological distress, and the prevalence of non-suicidal self-harm are reflected in the suicide risk estimations of the model. What is noteworthy here, however, is that the identified high-risk non-suicidal adolescents should be conceptualized as a risk group relevant to a suicide prevention program rather than mere classification error that should be dismissed. Although they do not report any suicide attempts, yet, compared to true negative cases, they are likely to be at higher risk of developing suicidal tendencies or attempt suicide in near future.

It is important to note that this study relied on cross-sectional data and only identified similarities and differences between the FP, TP, and TN groups as evidence for suicide attempt risk, which is a limitation. Future longitudinal studies should investigate whether FP adolescents are significantly more likely to attempt suicide or develop suicidal behavior than those in the TN group. As we used a

[2] Note that AUC and Cohen's Kappa can be biased under severe class imbalance and in this regard, AUPRC provides a more reliable model performance assessment.

representative dataset of Norwegian adolescents, we anticipate the findings could be applicable to other age groups. However, this should be confirmed in future research. Nevertheless, the study had several strengths, in particular the use of a comprehensive dataset. This dataset not only included a large number of participants, but also risk and protective survey items from a broad range of psycho-socio-environmental domains, thereby enabling an accurate estimation of suicide attempt risk.

Research has found low effectiveness for suicide intervention programs (Fox et al., 2020; Large, 2018), highlighting the importance of prevention rather than addressing suicidal tendencies after they emerge (Carter & Spittal, 2018). Therefore, identifying adolescents vulnerable to developing suicidal behavior is pivotal. Our study suggests that the FP group may be a relevant target for a suicide prevention program. If future longitudinal research confirms that FP adolescents (or other age groups) are clinically relevant and at high risk of attempting suicide, it will be necessary to devise new measures to evaluate the performance of machine learning classifiers for suicidal behavior. Such measures should consider the clinical relevance of the FP group. This information could lead to a redefinition of clinical relevance and the development of optimized cutoff values that maximize clinical relevance rather than overemphasizing sensitivity and specificity (Brown & Barlow, 2016; Hayes & Bell, 2014). This approach may also be applicable to other health-related binary outcomes and is not limited to suicide research. While this idea is novel, its generalization requires further investigation, particularly by using data from longitudinal studies.

Authors' contribution E. F. Haghish developed the research idea, carried out the statistical analysis, and wrote the first draft. Nikolai Czajkowski actively participated in discussions and revisions of the paper.

Data availability The Ungdata dataset (https://www.ungdata.no/) is not publicly available. However, researchers can apply for access to the dataset via https://www.nsd.no/.

Declarations

Ethical considerations The study was approved by the internal ethical committee at the Department of Psychology, University of Oslo. The respondents had given an informed consent to participate in the surveys and the data collection was in compliance with the 1964 Declaration of Helsinki and its later addenda. The authors also confirm that there is no conflict of interest nor any funding for this research.

Conflict of interest Authors confirm that they have no conflicts of interest to disclose.

References

Brahmbhatt, K., & Grupp-Phelan, J. (2019). Parent-adolescent agreement about adolescent's suicidal thoughts: A divergence. *Pediatrics, 143*(2), e20183071. https://doi.org/10.1542/peds.2018-3071

Brown, T. A., & Barlow, D. H. (2016). A proposal for a dimensional classification system based on the shared features of the DSM-IV anxiety and mood disorders: Implications for assessment and treatment. *Psychological Assessment, 21*(3), 256–271. https://doi.org/10.1037/a0016608

Burke, T. A., Jacobucci, R., Ammerman, B. A., Alloy, L. B., & Diamond, G. (2020). Using machine learning to classify suicide attempt history among youth in medical care settings. *Journal of Affective Disorders, 268*, 206–214. https://doi.org/10.1016/j.jad.2020.02.048

Carballo, J., Llorente, C., Kehrmann, L., Flamarique, I., Zuddas, A., Purper-Ouakil, D., Hoekstra, P., Coghill, D., Schulze, U., & Dittmann, R. (2020). Psychosocial risk factors for suicidality in children and adolescents. *European Child & Adolescent Psychiatry, 29*(6), 759–776. https://doi.org/10.1007/s00787-018-01270-9

Carter, G., & Spittal, M. J. (2018). Suicide risk assessment: Risk stratification is not accurate enough to be clinically useful and alternative approaches are needed. *Crisis: The Journal of Crisis Intervention and Suicide Prevention, 39*(4), 229–234. https://doi.org/10.1027/0227-5910/a000558

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*, 785–794. https://dl.acm.org/doi/10.1145/2939672.2939785

Christl, B., Wittchen, H.-U., Pfister, H., Lieb, R., & Bronisch, T. (2006). The accuracy of prevalence estimations for suicide attempts. How reliably do adolescents and young adults report their suicide attempts? *Archives of Suicide Research, 10*(3), 253–263. https://doi.org/10.1080/13811110600582539

Darke, S., Torok, M., Kaye, S., & Ross, J. (2010). Attempted suicide, self-harm, and violent victimization among regular illicit drug users. *Suicide and Life-Threatening Behavior, 40*(6), 587–596. https://doi.org/10.1521/suli.2010.40.6.587

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240. https://doi.org/10.1145/1143844.1143874

Fox, K. R., Huang, X., Guzmán, E. M., Funsch, K. M., Cha, C. B., Ribeiro, J. D., & Franklin, J. C. (2020). Interventions for suicide and self-injury: A meta-analysis of randomized controlled trials across nearly 50 years of research. *Psychological Bulletin, 146*(12), 1117–1145. https://doi.org/10.1037/bul0000305

Granello, D. H. (2010). The process of suicide risk assessment: Twelve core principles. *Journal of Counseling & Development, 88*(3), 363–370. https://doi.org/10.1002/j.1556-6678.2010.tb00034.x

Greening, L., Stoppelbein, L., Fite, P., Dhossche, D., Erath, S., Brown, J., Cramer, R., & Young, L. (2008). Pathways to suicidal behaviors in childhood. *Suicide and Life-Threatening Behavior, 38*(1), 35–45. https://doi.org/10.1521/suli.2008.38.1.35

Haghish, E. F. (2022). *AdjROC: Computing Sensitivity at a Fix Value of Specificity and Vice Versa* (0.2.0) [Computer software]. https://CRAN.R-project.org/package=adjROC

Haghish, E. F., Bang Nes, R., Obaidi, M., Qin, P., Stänicke, L. I., Bekkhus, M., Laeng, B., & Czajkowski, N. (2023). *Unveiling Adolescent Suicidality: Holistic Analysis of Protective and Risk Factors Using Multiple Machine Learning Algorithms [Manuscript submitted for publication]*.

Hart, S. R., Musci, R. J., Ialongo, N., Ballard, E. D., & Wilcox, H. C. (2013). Demographic and clinical characteristics of consistent and inconsistent longitudinal reporters of lifetime suicide attempts in adolescence through young adulthood. *Depression and Anxiety, 30*(10), 997–1004. https://doi.org/10.1002/da.22135

Hayes, J., & Bell, V. (2014). Diagnosis: One useful method among many. *The Lancet Psychiatry, 1*(6), 412–413. https://doi.org/10.1016/S2215-0366(14)70399-2

Healy, B. C. (2021). Machine and deep learning in MS research are just powerful statistics–No. *Multiple Sclerosis Journal, 27*(5), 663–664. https://doi.org/10.1177/1352458520978648

Jones, J. D., Boyd, R. C., Calkins, M. E., Ahmed, A., Moore, T. M., Barzilay, R., Benton, T. D., & Gur, R. E. (2019). Parent-adolescent agreement about adolescents' suicidal thoughts. *Pediatrics, 143*(2), e20181771. https://doi.org/10.1542/peds.2018-1771

Large, M. M. (2018). The role of prediction in suicide prevention. *Dialogues in Clinical Neuroscience*, *20*(3), 197–205. 10.31887/DCNS.2018.20.3/mlarge

Lewis, A. J., Bertino, M. D., Bailey, C. M., Skewes, J., Lubman, D. I., & Toumbourou, J. W. (2014). Depression and suicidal behavior in adolescents: A multi-informant and multi-methods approach to diagnostic classification. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00766

Ley, C., Martin, R. K., Pareek, A., Groll, A., Seil, R., & Tischer, T. (2022). Machine learning and conventional statistics: Making sense of the differences. *Knee Surgery, Sports Traumatology, Arthroscopy, 30*, 753–757. https://doi.org/10.1007/s00167-022-06896-6

Linthicum, K. P., Schafer, K. M., & Ribeiro, J. D. (2019). Machine learning in suicide science: Applications and ethics. *Behavioral Sciences & the Law, 37*(3), 214–222. https://doi.org/10.1002/bsl.2392

Lohner, J., & Konrad, N. (2006). Deliberate self-harm and suicide attempt in custody: Distinguishing features in male inmates' self-injurious behavior. *International Journal of Law and Psychiatry, 29*(5), 370–385. https://doi.org/10.1016/j.ijlp.2006.03.004

Miché, M., Studerus, E., Meyer, A. H., Gloster, A. T., Beesdo-Baum, K., Wittchen, H.-U., & Lieb, R. (2020). Prospective prediction of suicide attempts in community adolescents and young adults, using regression methods and machine learning. *Journal of Affective Disorders, 265*, 570–578. https://doi.org/10.1016/j.jad.2019.11.093

Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences, 2*(7), 1308. https://doi.org/10.1007/s42452-020-3060-1

Shen, Y., Zhang, W., Chan, B. S. M., Zhang, Y., Meng, F., Kennon, E. A., Wu, H. E., Luo, X., & Zhang, X. (2020). Detecting risk of suicide attempts among Chinese medical college students using a machine learning algorithm. *Journal of Affective Disorders, 273*, 18–23. https://doi.org/10.1016/j.jad.2020.04.057

Strand, B. H., Dalgard, O. S., Tambs, K., & Rognerud, M. (2003). Measuring the mental health status of the Norwegian population: A comparison of the instruments SCL-25, SCL-10, SCL-5 and MHI-5 (SF-36). *Nordic Journal of Psychiatry, 57*(2), 113–118. https://doi.org/10.1080/08039480310000932

Toprak, S., Cetin, I., Guven, T., Can, G., & Demircan, C. (2011). Self-harm, suicidal ideation and suicide attempts among college students. *Psychiatry Research, 187*(1–2), 140–144. https://doi.org/10.1016/j.psychres.2010.09.009

Van Orden, K. A., Witte, T. K., Cukrowicz, K. C., Braithwaite, S. R., Selby, E. A., & Joiner, T. E., Jr. (2010). The interpersonal theory of suicide. *Psychological Review, 117*(2), 575–600. https://doi.org/10.1037/a0018697

van Vuuren, C., van Mens, K., de Beurs, D., Lokkerbol, J., van der Wal, M., Cuijpers, P., & Chinapaw, M. (2021). Comparing machine learning to a rule-based approach for predicting suicidal behavior among adolescents: Results from a longitudinal population-based survey. *Journal of Affective Disorders, 295*, 1415–1420. https://doi.org/10.1016/j.jad.2021.09.018

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science, 5*(3), 457–469. https://doi.org/10.1177/2167702617691560