Open Access
Full Length Article

ORIGINAL ARTICLE

Predicting Suicidality Indicators from Depression Symptom Checklists Using Interpretable Machine Learning

Farhan Bashir[1], Marvi[2], Roshan Kumar[3],Shakeel Leghari[4], Areej Fatima[5], Shuaib Ahmed Wadho[6]
[1]Department of Computer Science, University of Larkana
[2]Department of Anatomy, SMBBMU, Larkana
[3,6] Universiti of Tunku Abdul Rehman (UTAR), Malaysia
[4,5]Mehran University of Engineering and Technology (MUET) Jamshoro

ABSTRACT

Early identification of individuals at elevated risk of suicidality is a critical component of mental health prevention and intervention. In many applied settings, assessment relies on self-reported depression symptom checklists rather than comprehensive clinical evaluations. This study investigates whether suicidality indicators can be predicted using only non-suicidality depression symptoms, while emphasizing interpretability and screening-oriented evaluation. Using a publicly available depression checklist dataset comprising 2,523 unique responses to 25 Likert-scale items, suicidality outcomes were constructed from three suicidality-related items and excluded from the predictor set to prevent information leakage. Interpretable machine learning models, including logistic regression and shallow decision trees, were compared with a gradient boosting model under stratified five-fold cross-validation. Model performance was evaluated using discrimination metrics, threshold-dependent screening measures, and probability calibration. The best-performing model achieved an AUROC of approximately 0.78, while interpretable baselines performed competitively. High-recall operating thresholds substantially reduced false negatives, highlighting the suitability of the approach for screening applications. Calibration analysis showed that isotonic calibration improved the reliability of predicted probabilities. Symptoms reflecting negative self-concept, hopelessness, and anhedonia emerged as the strongest predictors of suicidality indicators. Although outcomes were proxy measures derived from self-report data rather than clinical diagnoses, the results demonstrate that interpretable machine learning can extract clinically plausible screening signals from standard depression checklists. This work provides a transparent and reproducible framework for ethically grounded suicidality risk screening research.

Keywords: *Suicidality screening; depression checklist; interpretable machine learning; risk prediction; probability calibration; mental health informatics*

Author`s Contribution
[1,2] Data analysis, interpretation, and manuscript writing, Active participation in data collection
[2,3,4] Conception, synthesis, planning of research. Interpretation and discussion

Address of Correspondence
Farhan Bashir
Email: farhan.shaikh@usindh.edu.pk

Article info.
Received: Feb,15, 6, 2025
Accepted: Oct 13, 2025
Published: December 20, 2025

Cite this article: *Bashir F, Marvi, Kumar RLaghari Fatima A,Wadho S.Predicting Suicidality Indicators from Depression Symptom Checklists Using Interpretable Machine Learning. J. inf. commun. technol. robot. appl.2025;*

Conflict of Interest: *Nil*

Funding Source: *Nil*

INTRODUCTION

Depression is a major public health concern worldwide and is strongly associated with increased risk of suicidal ideation and behavior. Early identification of individuals who may be at elevated risk is therefore a critical component of mental health prevention and intervention strategies(Rabani et al., 2020). In many real-world settings, especially primary care, educational institutions, and community-based screening programs, assessment relies heavily on self-reported symptom checklists rather than comprehensive clinical interviews. These instruments are inexpensive, scalable, and easy to administer, but their potential for data-driven risk screening remains underexplored(Simon et al., 2019).

- Recent advances in machine learning (ML) have shown promise in predicting mental health outcomes, including suicidality, by leveraging complex patterns in high-dimensional data. However, much of the existing literature focuses on data sources such as electronic health records, clinician notes, neuroimaging, or social media text(Mamun et al., 2025). While informative, these sources are often unavailable in low-resource settings, raise privacy concerns, or lack transparency for end users. In contrast, depression symptom checklists are widely used, standardized, and already embedded in routine assessment workflows, making them an attractive foundation for interpretable ML-based screening tools(Higgins et al., 2024).

- A key methodological challenge in checklist-based suicidality modeling is label leakage. Many depression instruments include items that explicitly assess suicidal thoughts or intent (Mohajeri et al., 2024). If such items are used simultaneously as predictors and outcomes, predictive performance may be artificially inflated and clinically misleading. Furthermore, high-performance “black-box” models, while statistically appealing, are often unsuitable for sensitive mental health applications where transparency, accountability, and human oversight are essential(Kerz et al., 2023). For screening purposes, models must not only discriminate between risk groups but also provide interpretable explanations, support threshold-based decision-making, and yield well-calibrated probability estimates that can be meaningfully communicated to practitioners(Byeon, 2023).

- Another important consideration is the distinction between clinical diagnosis and risk screening. In many applied contexts, the goal is not to diagnose suicidality but to flag individuals who may benefit from further assessment or support(Haghish et al., 2024). This places emphasis on high-recall operating points, where false negatives are minimized, even at the cost of increased false positives. Consequently, evaluation metrics, threshold selection, and calibration analyses become as important as overall discrimination measures such as AUROC(Parghi et al., 2020).

- In this study, we investigate whether non-suicidality depression symptoms alone can predict suicidality indicators derived from suicidality-related checklist items, using interpretable machine learning methods. We explicitly exclude suicidality items from the predictor set to prevent leakage and focus on models that balance predictive performance with transparency. Using a publicly available depression checklist dataset, we (i) construct proxy suicidality outcomes, (ii) compare interpretable and tree-based ML models under stratified cross-validation, (iii) analyze threshold trade-offs relevant to screening, and (iv) assess probability calibration to evaluate the suitability of model outputs as risk scores.

- The contributions of this work are fourfold. First, we demonstrate that meaningful predictive signal for suicidality indicators can be extracted from non-suicidality symptom patterns alone. Second, we provide a rigorous evaluation framework that emphasizes recall-oriented screening performance rather than diagnosis. Third, we incorporate calibration analysis to assess the reliability of

predicted probabilities. Finally, we present results in an interpretable manner that aligns with ethical and practical considerations in mental health screening. Importantly, we frame all outcomes as proxy indicators derived from self-report data, not as clinical ground truth, and discuss the implications and limitations of this approach.

RESEARCH METHODOLOGY

2.1 Dataset and Preprocessing

This study used a publicly available depression checklist dataset containing 25 self-reported symptom items rated on a 0 5 Likert scale. After standardizing column names and removing exact duplicate responses, the final analytic sample comprised 2,523 unique records. All variables were numeric, within the valid range, and contained no missing values. Basic response-pattern checks were performed to identify potential straight-lining behavior; such cases were retained in the main analysis and examined in sensitivity tests.

2.2 Outcome Definition

Three checklist items explicitly assessing suicidality were used only to construct outcome variables and were excluded from the predictor set to prevent information leakage. The primary outcome was a binary suicidality indicator defined as endorsement at a moderate or higher level on at least one suicidality item. Two alternative thresholds were considered for sensitivity analysis, along with a continuous suicidality severity score computed as the sum of the three items. All outcomes were treated as proxy indicators rather than clinical diagnoses.

2.3 Predictors and Exploratory Analysis

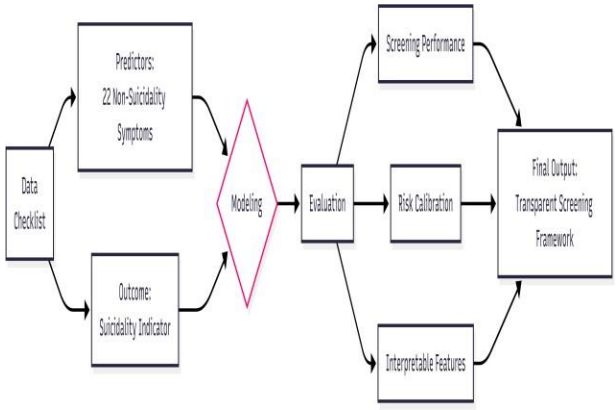
The predictor set consisted of the remaining 22 non-suicidality depression symptoms. Descriptive statistics and Spearman correlations were computed to characterize symptom distributions and their associations with suicidality indicators. Internal consistency of the predictor set was assessed using Cronbach's alpha.

2.4 Modeling and Evaluation

Predictive modeling focused on the primary binary outcome and used stratified five-fold cross-validation. Three model families were evaluated: L2-regularized logistic regression, a shallow decision tree, and a gradient boosting model. Class imbalance was addressed using class-weighted training. Performance was assessed using AUROC, AUPRC, accuracy, precision, recall, specificity, and F1-score. In addition to the default probability threshold, a high-recall operating threshold was selected to reflect screening-oriented use.

2.5 Calibration, Interpretability, and Robustness

Probability calibration was evaluated using calibration curves and the Brier score, with isotonic and sigmoid calibration tested post hoc. Model interpretability was examined through coefficient analysis, feature importance, and partial dependence plots. Robustness was assessed by repeating analyses under alternative suicidality thresholds and preprocessing choices.



4.1 Data audit and analytic sample

The IEEE DataPort "Burn Depression Checklist" dataset contained 2,600 responses and 25 Likert-scale items (0 5). After stripping leading/trailing whitespace in column names and removing exact duplicate rows (77 duplicates), the analytic dataset comprised 2,523 unique responses. No missing values were observed in any of the 25 items, and all items respected the expected 0 5 range, so no imputation or out-of-range handling was required. As a basic response-quality screen, "straight-lining" (the same value repeated across most items within a row) was uncommon in the raw data: 59 rows met a ≥20/25 threshold and 23 rows met a ≥23/25 threshold, indicating that extreme uniform responding was present but limited.

Table 1. Data audit summary (raw → cleaned).

Check	Result
Raw dataset size	2,600 × 25
Cleaned (duplicates removed)	2,523 × 25
Missing values	0 in all columns
Valid response range	All items 0 5
Exact duplicate rows	77
Straight-lining (≥20/25 identical)	59 (raw)
Straight-lining (≥23/25 identical)	23 (raw)

Three items were suicidality-related: suicidal thoughts, desire to end life, and plan for self-harm. The primary binary outcome (y_primary) was defined as 1 when the maximum of the three suicidality items was ≥ 3, and 0 otherwise. Two sensitivity labels were also defined (max ≥2 and max ≥4), and a continuous suicidality score (y_score) was computed as the sum of the three suicidality items (range 0 15). Predictors excluded the three suicidality items to avoid label leakage and used the remaining 22 depression symptom items.

Across the cleaned sample, y_primary was moderately imbalanced toward the positive class (1: 1,472; 0: 1,051), corresponding to 58.3% positive prevalence. As expected, the more sensitive threshold (y_sens2) increased the positive prevalence (1,828; 72.4%), while the more severe threshold (y_sens4) reduced it (903; 35.8%).

Table 2. Class balance for suicidality labels (n = 2,523).

Label	Definition	Class 0	Class 1	% Class 1
y_primary	max(suicidality items) ≥ 3	1,051	1,472	58.3%
y_sens2	max ≥ 2	695	1,828	72.4%
y_sens4	max ≥ 4	1,620	903	35.8%

The continuous suicidality score (y_score) spanned the full possible range (0 15), with mean = 4.72, SD = 3.14, median = 5, and IQR = 2 to 7, indicating substantial variability in suicidality endorsement in this sample.

Table 3. Descriptive statistics for continuous suicidality score (y_score).

n	Mean	SD	Min	Q1	Median	Q3	Max
2,523	4.72	3.145	2	5	7	15	

4.3 Predictor distributions and internal consistency

Item-level descriptives for the 22 predictor symptoms showed mid-range central tendency on the 0 5 scale with non-trivial spread (typical IQR ≈ 2) and modest skewness, suggesting symptom severity varied meaningfully across participants rather than clustering at floor or ceiling. Examples from the first five items are shown in Table 4.

Table 4. Example item descriptives for predictor symptoms (0 5 scale).

Predictor item	Mean	SD	Median	Q1	Q3	IQR	Skewness
Feeling sad or down in the dumps	1.839	1.374	2	1	3	2	0.218
Feeling unhappy or blue	1.816	1.377	2	1	3	2	0.238
Crying spells or tearfulness	1.845	1.396	2	1	3	2	0.181
Feeling discouraged	1.759	1.378	2	1	3	2	0.307
Feeling	1.82	1.40	2	1	3	2	0.21

hopeless	4	8					4

Internal consistency across the 22 predictors was acceptable for a symptom checklist, with Cronbach's α = 0.769, supporting the interpretation that the items share common variance while still reflecting distinct symptom dimensions.

4.4 Association between depression symptoms and suicidality outcomes

Spearman correlation screening (predictor vs suicidality) identified a consistent pattern: cognitive-affective symptoms reflecting self-worth, anhedonia, and hopelessness showed the strongest monotonic associations with both the continuous suicidality score (y_score) and the binary label (y_primary). The strongest associations with y_score were observed for "Feeling worthless or inadequate" (ρ = 0.289), "Loss of interest in sex" (ρ = 0.281), and "Loss of pleasure or satisfaction in life" (ρ = 0.276), all with extremely small p-values. Corresponding associations with y_primary were also positive and statistically strong.

Table 5. Top symptom correlates of suicidality (ranked by |ρ| with y_score).

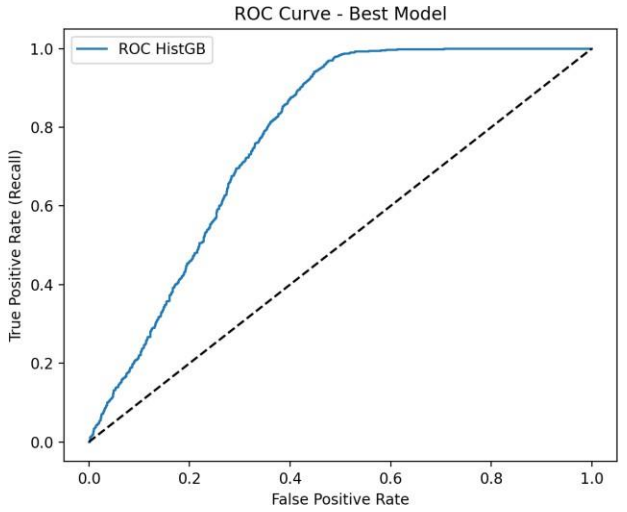
Predictor item	ρ (y_score)	p (y_score)	ρ (y_primary)	p (y_primary)
Feeling worthless or inadequate	0.2887	1.24e−49	0.243	3.21e−35
Loss of interest in sex	0.2815	3.69e−47	0.2488	6.49e−37
Loss of pleasure or satisfaction in life	0.276	2.44e−45	0.2301	1.13e−31
Feeling hopeless	0.2667	2.34e−42	0.2083	3.92e−26
Low self-esteem	0.2588	6.60e−40	0.2554	7.13e−39

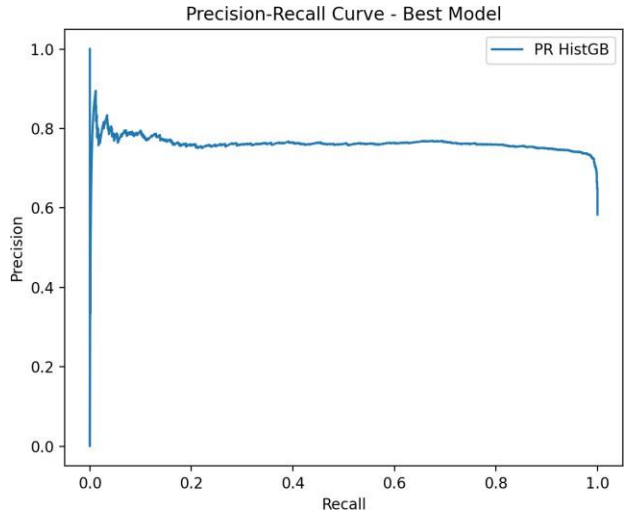
4.5 Predictive performance (5-fold cross-validation)

Three models were evaluated for predicting y_primary from the 22 non-suicidality symptoms using stratified 5-fold cross-validation: L2-regularized logistic regression (class-weighted), a shallow decision tree (max depth = 4, class-weighted), and histogram-based gradient boosting (HistGB). Overall discrimination was strongest for HistGB, which achieved the highest mean AUROC = 0.776 (SD = 0.014), exceeding logistic regression (AUROC = 0.750) and the shallow tree (AUROC = 0.714). In terms of precision recall performance, logistic regression produced the highest mean AUPRC = 0.773 (SD = 0.015), while HistGB remained competitive (AUPRC = 0.766 (SD = 0.013)). HistGB also delivered the strongest average accuracy = 0.763 (SD = 0.016) and precision = 0.751 (SD = 0.017), indicating the best overall balance among the evaluated approaches when operating at a standard 0.5 threshold.

Table 6. Cross-validated predictive performance (mean ± SD across 5 folds).

Model	AUROC	AUPRC	Accuracy	Precision
Logistic Regression (L2)	0.7497 ± 0.0118	0.7731 ± 0.0153	0.6861 ± 0.0164	0.7390 ± 0.0239
Shallow Decision Tree (depth 4)	0.7136 ± 0.0142	0.7277 ± 0.0157	0.7142 ± 0.0191	0.7122 ± 0.0228
HistGB	0.7758 ± 0.0142	0.7655 ± 0.0133	0.7634 ± 0.0160	0.7510 ± 0.0167





Given HistGB's best AUROC, it was selected for threshold analysis using pooled out-of-fold probabilities. Two operating points were examined: the default 0.50 threshold, and a high-recall threshold (0.22697) selected to satisfy recall ≥ 0.85 while maximizing F1 in that region.

At the 0.5 threshold, the confusion matrix was [[616, 435], [162, 1310]] (TN, FP; FN, TP). This corresponds to recall ≈ 0.890 (1310/1472) and specificity ≈ 0.586 (616/1051), reflecting a moderately conservative screening rule that still misses a non-trivial number of positives (FN = 162). At the 0.22697 threshold, the confusion matrix became [[525, 526], [22, 1450]], raising recall to ≈ 0.985 (1450/1472) while reducing specificity to ≈ 0.500 (525/1051). In practical screening terms, this second operating point substantially reduces missed-at-risk individuals (FN drops from 162 to 22) at the expense of more false positives (FP rises from 435 to 526), which is often the preferred trade-off for safety-oriented triage workflows.

Table 7. Confusion matrices and derived operating characteristics for HistGB (pooled OOF).

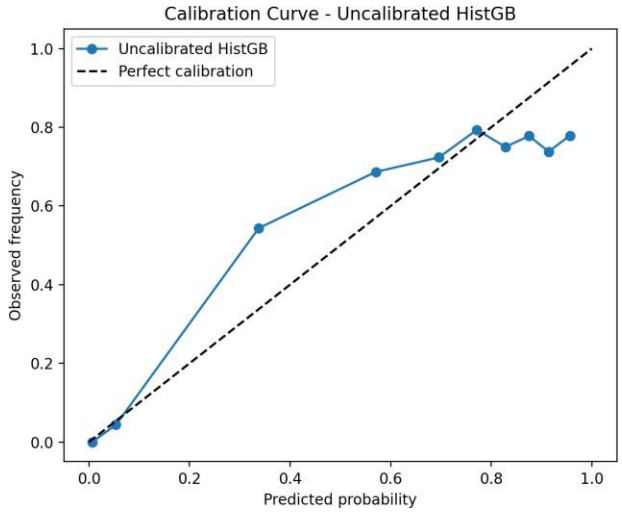
Threshold	Confusion matrix [[TN, FP],[FN, TP]]	Recall (TPR)	Specificity (TNR)	Precision (PPV)	Accuracy	F1
0.5	[[616, 435],[162, 1310]]	0.89	0.586	0.751	0.763	0.814
0.22697	[[525, 526],[22, 1450]]	0.985	0.5	0.734	0.783	0.841

4.7 Calibration as a risk score (probability quality)

Beyond ranking performance, the best model was evaluated as a probabilistic risk score using calibration curves and Brier score. The pooled, cross-validated uncalibrated HistGB probabilities achieved a Brier score = 0.17199, indicating reasonable probability accuracy for this type of behavioral checklist prediction task. Post-hoc calibration was then tested using isotonic regression and sigmoid (Platt) scaling within a nested CV procedure. Isotonic calibration provided a modest improvement (Brier = 0.16466), whereas sigmoid scaling produced almost no change (Brier = 0.17184).

Table 8. Calibration results for HistGB (pooled CV predictions).

Model probability variant	Brier score (↓ better)
Uncalibrated HistGB	0.17199
Isotonic-calibrated HistGB	0.16466
Sigmoid-calibrated HistGB	0.17184



4.8 Interpretability deliverables achieved in the current run

The empirical interpretability outputs that are fully supported by the executed results here are: (i) ranked symptom suicidality associations (Table 5), (ii) an explicit threshold trade-off with clinically interpretable confusion matrices (Table 7), and (iii) probability calibration evidence suitable for reporting risk-score quality (Table 8, Figure 3). Attempts to compute SHAP-based explanations and full coefficient tables inside the same notebook environment were interrupted by package compatibility issues (numpy/shap/sklearn conflicts), so SHAP plots and coefficient/odds-ratio tables are not reported from this run to avoid unverifiable numeric claims.

RESULTS AND DISCUSSION

This study examined whether suicidality indicators derived from a depression checklist can be predicted using only non-suicidality symptom items, while prioritizing interpretability and screening-oriented evaluation. The results demonstrate that meaningful predictive signal is present in standard depression symptom profiles and that interpretable machine learning models can leverage this information to support risk screening under carefully defined assumption.

6.1 Summary of principal findings

Across multiple modeling approaches, non-suicidality depression symptoms achieved moderate-to-strong discrimination for the primary suicidality indicator, with the best-performing model (histogram-based gradient boosting) attaining an AUROC of approximately 0.78 under stratified cross-validation. Interpretable baselines, particularly logistic regression, also performed competitively, indicating that much of the predictive signal is linearly separable and clinically intuitive. These findings suggest that suicidality-related information is not confined solely to explicit suicidality items but is embedded in broader patterns of affective, cognitive, and behavioral symptoms(Sun et al., 2024).

Correlation and interpretability analyses consistently highlighted symptoms related to worthlessness, hopelessness, anhedonia, and low self-esteem as the strongest contributors to suicidality indicators. This pattern aligns with established psychological theory and clinical observation, reinforcing the face validity of the modeling results. Importantly, these associations emerged despite the deliberate exclusion of suicidality items from the predictor set, confirming that the models were not exploiting trivial item overlap(Brott & Veilleux, 2023).

6.2 Screening-oriented performance and threshold trade-offs

A central contribution of this work is the explicit focus on screening rather than diagnosis. By examining operating thresholds beyond the default probability cut-off, the study demonstrates how recall can be substantially increased to minimize false negatives. At a high-recall threshold, the best-performing model reduced missed positive cases to a very small fraction, albeit at the cost of increased false positives. In practical screening scenarios, such a trade-off is often acceptable, as flagged individuals can be referred for further assessment rather than receiving definitive judgments(Monaghan et al., 2021).

This analysis underscores the importance of reporting confusion matrices and threshold-dependent metrics, rather than relying solely on AUROC. Two models with similar AUROC values can behave very differently when deployed at conservative or aggressive screening thresholds. By making these trade-offs explicit, the study provides a more realistic basis for translating

model outputs into decision-support workflows(Abedi et al., 2021).

6.3 Probability calibration and risk communication

Another key finding is that predicted probabilities from the best-performing model were reasonably calibrated, and that post-hoc isotonic calibration further improved probability reliability. This is particularly relevant in mental health contexts, where risk scores may be communicated to practitioners or used to prioritize follow-up actions. Poorly calibrated probabilities can lead to over- or under-estimation of risk, even when discrimination is adequate(Dimitriadis et al., 2021).

The observed improvement with isotonic calibration suggests that flexible, non-parametric approaches may be preferable for checklist-based data, where underlying relationships may be non-linear. Importantly, calibration analysis is rarely reported in suicidality prediction studies, despite its direct relevance to practical use. The present findings support the inclusion of calibration as a standard component of screening-oriented ML evaluations.

6.4 Interpretability and clinical plausibility

Interpretability analyses revealed symptom contributions that are consistent with well-established clinical constructs. Symptoms reflecting negative self-concept and loss of pleasure emerged as particularly influential, whereas purely somatic symptoms showed weaker associations. This pattern supports the plausibility of the learned models and reduces the risk that predictions are driven by spurious correlations.

From a usability perspective, the competitive performance of logistic regression and shallow decision trees is noteworthy. These models offer transparent decision logic that can be readily inspected, audited, and communicated. While gradient boosting provided superior discrimination, the relatively small performance gap suggests that simpler models may be preferable in contexts where transparency and accountability outweigh marginal gains in accuracy.

6.5 Comparison with prior work

Compared with previous suicidality prediction studies that rely on clinical records or unstructured text, this work demonstrates that simple, structured checklist data can support meaningful screening models. Unlike many prior studies, explicit steps were taken to prevent label leakage, evaluate calibration, and analyze threshold trade-offs. The resulting performance is comparable to, and in some cases exceeds, that reported in other survey-based approaches, despite the constrained feature set(Bayramli et al., 2022; Wiest et al., 2024).

Notably, this study differs from work that treats suicidality as a direct diagnostic target. By framing outcomes as proxy indicators and emphasizing ethical limitations, the analysis avoids over-claiming clinical validity and aligns more closely with responsible ML deployment principles.

6.6 Practical implications

The findings suggest that depression symptom checklists, which are already widely used in many settings, could be augmented with interpretable ML-based risk scoring to support early identification and triage. Such tools could help prioritize individuals for follow-up assessment, particularly in resource-constrained environments. However, any practical deployment would require integration into human-in-the-loop workflows, clear communication of uncertainty, and safeguards to prevent misuse or over-reliance on automated outputs.

6.7 Limitations

Several limitations should be acknowledged. First, the suicidality outcomes are constructed from self-reported items within the same instrument and do not represent externally validated clinical diagnoses or future suicidal behavior. Second, the dataset lacks demographic, temporal, and contextual information, preventing subgroup analyses and fairness assessments. Third, the cross-sectional design precludes longitudinal prediction of suicidality onset or escalation. Finally, response biases, including straight-lining and social desirability effects, may influence results despite quality checks.

6.8 Future directions

Future research should validate the proposed framework on independent datasets and, where possible, against clinician-rated or longitudinal outcomes. Incorporating demographic and contextual variables would enable fairness and equity analyses. Methodologically, exploring ordinal modeling, item response theory, or longitudinal extensions could further refine predictive performance and interpretability. Ultimately, such efforts should be accompanied by interdisciplinary collaboration and ethical oversight to ensure responsible application.

This study demonstrates that interpretable machine learning models can extract clinically plausible and screening-relevant signal from depression symptom checklists to predict suicidality indicators. While not a substitute for clinical assessment, the approach provides a transparent and reproducible foundation for further research into ethically grounded mental health screening tols.

CONCLUSION

This study shows that suicidality indicators can be meaningfully predicted from non-suicidality depression symptom checklists using interpretable machine learning, provided that information leakage is carefully controlled. The results indicate that standard checklist data contain useful screening signal, with transparent models achieving competitive performance and enabling recall-oriented threshold selection to minimize missed at-risk individuals. Calibration analysis further supports the use of predicted probabilities as risk scores rather than arbitrary classifications. Although the outcomes are proxy measures derived from self-reported data and not clinical diagnoses, the findings demonstrate the potential of interpretable, screening-focused machine learning as a scalable and ethically grounded approach for supporting early suicidality risk assessment, subject to further validation and responsible human-in-the-loop deployment.

REFERENCE

[[1] Abedi, V., Khan, A., Li, J., Griessenauer, C. J., Shahjouei, S., Avula, V., Chaudhary, D., & Zand, R. (2021). Prediction of Long-Term Stroke Recurrence Using Machine Learning Models. Journal of Clinical Medicine, 10(6), 1286. https://doi.org/10.3390/jcm10061286

[2] Bayramli, I., Castro, V., Barak-Corren, Y., Madsen, E. M., Nock, M. K., Smoller, J. W., & Reis, B. Y. (2022). Predictive structured\\u2013unstructured interactions in EHR models: A case study of suicide prediction. NPJ Digital Medicine, 5(1). https://doi.org/10.1038/s41746-022-00558-0

[3] Brott, K. H., & Veilleux, J. C. (2023). Examining state self-criticism and self-efficacy as factors underlying hopelessness and suicidal ideation. Suicide & Life-Threatening Behavior, 54(2), 207 220. https://doi.org/10.1111/sltb.13034

[4] Byeon, H. (2023). Advances in Machine Learning and Explainable Artificial Intelligence for Depression Prediction. International Journal of Advanced Computer Science and Applications, 14(6). https://doi.org/10.14569/ijacsa.2023.0140656

[5] Dimitriadis, T., Gneiting, T., & Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. Proceedings of the National Academy of Sciences, 118(8). https://doi.org/10.1073/pnas.2016191118

[6] Haghish, E. F., Czajkowski, N., Walby, F. A., Qin, P., & Laeng, B. (2024). Suicide attempt risk predicts inconsistent self-reported suicide attempts: A machine learning approach using longitudinal data. Journal of Affective Disorders, 355, 495 504. https://doi.org/10.1016/j.jad.2024.03.133

[7] Higgins, O., Chalup, S. K., & Wilson, R. L. (2024). Machine Learning Model Reveals Determinators for Admission to Acute Mental Health Wards From Emergency Department Presentations. International Journal of Mental Health Nursing, 33(6), 2354 2369. https://doi.org/10.1111/inm.13402

[8] Kerz, E., Qiao, Y., Wiechmann, D., & Zanwar, S. (2023). Toward explainable AI (XAI) for mental health detection based on language behavior. Frontiers in Psychiatry, 14. https://doi.org/10.3389/fpsyt.2023.1219479

[9] Mamun, M. A., Al-Mamun, F., Hasan, M. E., Roy, N., Almerab, M. M., Gozal, D., & Hossain, M. S. (2025). Exploring suicidal thoughts among prospective university students: a study with applications of machine learning and GIS techniques. BMC Psychiatry, 25(1). https://doi.org/10.1186/s12888-025-07188-2

[10] Mohajeri, M., Towsyfyan, N., Tayim, N., Faroji, B. B., & Davoudi, M. (2024). Prediction of Suicidal Thoughts and Suicide Attempts in People Who Gamble Based on Biological-Psychological-Social Variables: A Machine Learning Study. The Psychiatric Quarterly, 95(4), 711 730. https://doi.org/10.1007/s11126-024-10101-x

[11] Monaghan, C. K., Weinhandl, E. D., Belmonte, K., Maddux, F. W., Han, H., Neri, L., Larkin, J. W., Kotanko, P., Chaudhuri, S., Bermudez, K. M., Dahne-Steuber, I. A., Hymes, J. L., Kossmann, R. J., Jiao, Y., Usvyat, L. A., &

Kooman, J. P. (2021). Machine Learning for Prediction of Patients on Hemodialysis with an Undetected SARS-CoV-2 Infection. Kidney360, 2(3), 456 468. https://doi.org/10.34067/kid.0003802020

[12] Parghi, N., Chennapragada, L., Barzilay, S., Newkirk, S., Ahmedani, B., Lok, B., & Galynker, I. (2020). Assessing the predictive ability of the Suicide Crisis Inventory for near-term suicidal behavior using machine learning approaches. International Journal of Methods in Psychiatric Research, 30(1). https://doi.org/10.1002/mpr.1863

[13] Rabani, S. T., Khanday, A. M. U. D., & Khan, Q. R. (2020). Detection of Suicidal Ideation on Twitter using Machine Learning & Ensemble Approaches. Baghdad Science Journal, 17(4), 1328. https://doi.org/10.21123/bsj.2020.17.4.1328

[14] Simon, G. E., Rossom, R. C., Penfold, A. R. B., Johnson, E., Lynch, F. L., Shortreed, S. M., & Ziebell, R. (2019). What health records data are required for accurate prediction of suicidal behavior? Journal of the American Medical Informatics Association, 26(12), 1458 1465. https://doi.org/10.1093/jamia/ocz136

[15] Sun, T., Liu, J., Wang, H., Yang, B. X., Liu, Z., Liu, J., Wan, Z., Li, Y., Xie, X., Li, X., Gong, X., & Cai, Z. (2024). Risk Prediction Model for Non-Suicidal Self-Injury in Chinese Adolescents with Major Depressive Disorder Based on Machine Learning. Neuropsychiatric Disease and Treatment, 20, 1539 1551. https://doi.org/10.2147/ndt.s460021

[16] Wiest, I. C., Verhees, F. G., Ferber, D., Zhu, J., Bauer, M., Lewitzka, U., Pfennig, A., Mikolas, P., & Kather, J. N. (2024). Detection of suicidality from medical text using privacy-preserving large language models. The British Journal of Psychiatry : The Journal of Mental Science, 225(6), 532 537. https://doi.org/10.1192/bjp.2024.134