



“AI-Assisted Suicide and Self-Harm Risk Detection: Implications for Mental Health Nursing Practice”

Jeena Scaria¹

¹Senior Lecturer

Mental Health Nursing Department

Al-Azhar College of Nursing

Thodupuzha, Kerala

¹Correspondence Email: jeenascaria78@gmail.com

Date of Publication: 06/10/2026

DOI [10.5281/zenodo.23183218](https://doi.org/10.5281/zenodo.23183218)

Abstract: Suicide and self-harm represent critical public health challenges requiring innovative early detection and intervention strategies. Recent advances in artificial intelligence, particularly machine learning and natural language processing, have demonstrated promising potential for identifying individuals at elevated risk for suicidal behavior and self-injury. This narrative review synthesizes evidence regarding AI-assisted risk detection systems in mental health contexts, examining their clinical applications, empirical evidence, limitations, and implications for psychiatric and mental health nursing practice. Key findings indicate that machine learning models trained on electronic health record data, combined with natural language processing analysis of clinical narratives, can identify suicide risk with sensitivity ranging from 0.56 to 0.88 and specificity from 0.57 to 0.99, depending on model architecture and population characteristics. Hierarchical machine learning approaches integrating clinical history with repeated nursing assessments demonstrate superior performance for detecting self-harm risk in hospitalized psychiatric patients. However, successful clinical implementation requires addressing critical challenges including algorithmic bias, data privacy concerns, clinical validation across diverse populations, and preservation of the therapeutic nurse-patient relationship. Mental health nurses are positioned as essential stakeholders in AI implementation, requiring education regarding AI capabilities and limitations, meaningful participation in system design and governance, and clear accountability frameworks. This review concludes that responsible AI integration in suicide and self-harm risk detection must prioritize ethical implementation, maintain human clinical judgment as primary, and ensure that technology augments rather than replaces nursing assessment and therapeutic intervention.

Keywords: artificial intelligence, machine learning, suicide risk, self-harm, mental health nursing, risk prediction, natural language processing, clinical decision support

Introduction

Suicide and non-suicidal self-injury represent significant and growing global mental health concerns, causing immense suffering for affected individuals, families, and communities. Globally, suicide accounts for over one million deaths annually, with millions more engaging in self-harm behaviors. [1] Mental health nurses, working across diverse settings from inpatient psychiatric units to emergency departments to community mental health centers, occupy critical positions for identifying individuals at elevated risk and implementing timely interventions. Traditional approaches to suicide and self-harm risk assessment, while essential, face substantial limitations including reliance on clinical interview accuracy, variable inter-rater reliability of standardized assessment tools, constrained observation capacity given nursing

workload demands, and inability to identify subtle behavioral or linguistic changes that may precede crisis. [2]

Recent technological advances in artificial intelligence, particularly machine learning and natural language processing, have created unprecedented opportunities for early identification of individuals at risk for suicidal behavior and self-harm. [3] Machine learning algorithms can analyze vast datasets comprising structured clinical information, unstructured narrative documentation, behavioral observations, and temporal patterns to identify risk signatures that might escape human detection. Natural language processing can extract clinically relevant information from free-text clinical narratives, discharge summaries, and patient communications to identify linguistic markers associated with suicidal ideation and



planning. [4] These technologies promise to enhance the sensitivity and specificity of risk

detection while potentially reducing burden on nursing staff, yet their implementation introduces complex clinical, ethical, and professional considerations that mental health nurses must understand and navigate.

This narrative review examines current evidence regarding artificial intelligence applications in suicide and self-harm risk detection, with specific focus on implications for mental health nursing practice. Through synthesis of peer-reviewed literature, this review explores AI capabilities and limitations, identifies evidence supporting clinical effectiveness, articulates ethical and professional concerns, and proposes frameworks for responsible implementation that preserve nursing judgment and maintain patient-centered care principles.

Machine Learning and Natural Language Processing for Suicide Risk Prediction

Machine learning applications to suicide risk prediction represent one of the most extensively researched domains of AI in mental health. [5] Multiple research groups have developed and validated machine learning models capable of analyzing electronic health record data to predict risk of suicide attempts or death. Systematic reviews examining natural language processing applications for suicide detection identified twenty peer-reviewed studies utilizing NLP-based algorithms to predict suicidal ideation and self-harm from free-text clinical documentation. [6] Research demonstrates feasibility of applying NLP-based machine learning prediction algorithms to identify suicide risk from diverse data sources including discharge summaries, clinical notes, mobile phone text responses, and open-ended interview transcripts. [7]

Studies combining structured and unstructured data in machine learning models have yielded more accurate risk prediction than either data type alone. [8] For example, research examining natural language processing for suicide risk prediction among adults discharged from psychiatric inpatient or emergency room settings compared NLP-based models analyzing unstructured responses to open-ended questions with traditional logistic regression models using structured clinical data. [9] NLP-based models demonstrated positive predictive value of 0.61, sensitivity of 0.56, and specificity of 0.57 for identifying suicidal ideation, while structured data-based models achieved higher positive predictive value of 0.73, sensitivity of 0.76, and specificity of 0.62. These findings suggest that while NLP approaches show promise, optimal predictive performance

may require integrating both structured and unstructured data sources.

Recent research examining machine learning models for first-time suicide attempt prediction demonstrated the value of integrating machine learning with natural language processing of unstructured electronic health record data. [10] Large-scale data-driven approaches analyzing both structured clinical variables and unstructured narrative documentation achieved area under the receiver operating characteristic curve values of 0.83 to 0.86 for predicting suicide attempts within ninety days of clinical contact. Importantly, these studies highlight that unstructured clinical text from patient narratives, discharge summaries, and clinical encounters provides substantial predictive information not captured through structured data alone. [11] This finding has significant implications for mental health nursing, as detailed clinical documentation of patient statements, emotional presentation, and observed behaviors may contain vital risk indicators. [12]

AI-Assisted Detection of Self-Harm Risk in Hospitalized Psychiatric Patients

Self-harm risk detection in hospitalized psychiatric settings presents distinct challenges from suicide risk prediction, as self-injury patterns are often covert, modified by environmental constraints, and influenced by dynamic psychological factors. [2] Recent research has developed hierarchical machine learning models that integrate baseline admission clinical information with longitudinal nursing observations to identify hospitalized psychiatric patients at elevated risk for self-harm and suicidal behavior. [13] A comprehensive study of hospitalized patients with schizophrenia diagnosis developed a two-layered machine learning framework combining admission clinical history with repeated nursing assessments of behavioral and emotional indicators.

The hierarchical machine learning model identified dynamic indicators significantly predictive of self-harm risk. Negative self-evaluation demonstrated the highest odds ratio at 2.303, followed by self-reported depression (odds ratio 1.812), insomnia (odds ratio 1.768), talking to oneself (odds ratio 1.733), crying (odds ratio 1.700), and reduced conversation with others (odds ratio 1.422). [13] The optimized model combining baseline clinical variables with dynamic nursing observations achieved area under the curve of 0.9048, with sensitivity of 0.8542 and specificity of 0.7789, representing superior performance compared to either data component alone. Notably, this hierarchical approach



demonstrated higher sensitivity than traditional flat machine learning models, indicating more clinically appropriate detection of high-risk patients. These findings underscore the importance of systematic nursing observation documentation and suggest that machine learning models can effectively integrate static clinical information with dynamic behavioral observations to enhance risk detection. [13]

Deep learning approaches utilizing recurrent neural networks with Long Short-Term Memory architecture have been applied to ecological momentary assessment data to predict near-term suicidal and self-harm ideation. [14] Ecological momentary assessment involves brief, repeated assessments of current mood and symptoms via smartphone or mobile device, capturing real-time data regarding emotional states and ideation. Deep learning models trained on two-week sequences of ecological momentary assessment data demonstrated feasibility for predicting suicidal and self-harm ideation in general population samples. Such approaches offer potential for continuous, unobtrusive monitoring of risk trajectories, enabling identification of individuals whose risk is escalating in real-time, thereby supporting timely intervention. [14]

Machine learning models for predicting self-harm risk among young adults using national mental health data have demonstrated modest but clinically meaningful predictive performance. [15] Studies analyzing demographic and clinical commissioning data for hospitalized young adults identified ensemble machine learning models combining multiple algorithms as most effective, achieving area under the receiver operating characteristic curve of 0.77 for predicting self-harm episodes during hospitalization. [16] While these predictive values are lower than ideal, the large scale of these studies and use of readily available demographic and clinical data suggest potential for population-level risk stratification supporting nursing triage and resource allocation.

Video and Behavioral Analysis Technologies for Self-Harm Detection

Emerging technologies utilizing deep learning-based video analysis represent a frontier in self-harm detection technology. [17] These systems analyze real-time video data from monitoring cameras or mobile devices to identify behavioral patterns associated with self-injury or other self-harm behaviors. Researchers have developed deep learning models trained on action recognition datasets to classify behaviors indicative of potential self-harm. The clinical utility of such technologies lies in their potential to augment continuous monitoring by nursing staff,

identifying self-harm attempts that might occur during times of constrained supervision such as nighttime hours or periods of reduced staffing. [17]

However, implementation of video-based behavioral monitoring presents significant ethical and practical challenges.

[18] Privacy concerns are paramount, as video surveillance of psychiatric patients raises substantial questions regarding patient dignity, autonomy, and fundamental rights. Additionally, training data collection for such systems is inherently challenging, requiring either simulation of self-harm behaviors or use of sensitive historical video data from previous incidents. [17] Model performance must be rigorously validated to prevent false positives that would generate excessive alerts and false negatives that would miss actual incidents. Furthermore, such systems must be integrated with meaningful clinical response mechanisms, ensuring that detected behaviors trigger appropriate nursing interventions rather than simply recording incidents.

Clinical Challenges and Implementation Barriers

Despite promising research evidence, multiple substantial barriers impede implementation of AI-assisted suicide and self-harm risk detection systems in real-world clinical settings. [8] Data quality and completeness present fundamental challenges, as electronic health record data often contains missing values, inconsistent documentation practices, and incomplete longitudinal information. Machine learning models trained on incomplete or biased data may perform poorly when applied to new patient populations or different clinical contexts. [19] Clinical validation of AI systems remains inadequate in many cases, with insufficient prospective evaluation of model performance in actual clinical settings with real patients. Algorithmic bias represents a critical concern in suicide and self-harm risk prediction. [20] Machine learning models trained on historical data inherently reflect biases present in that data, potentially leading to biased risk predictions. For example, if training data comes predominantly from psychiatric inpatient populations, models may perform poorly for individuals first presenting to primary care or emergency departments. Racial and ethnic disparities in mental health diagnosis, treatment access, and documentation practices may result in models that systematically misclassify risk for certain populations. [21] Mental health nurses implementing AI systems have responsibility to advocate for equitable model development, demand evidence of equitable performance across populations, and resist adoption of systems lacking such validation.



Clinical workflow integration presents practical implementation challenges. [22] Effective AI systems must provide risk assessments and alerts at appropriate times with sufficient specificity to avoid overwhelming nursing staff with false alarms while maintaining sufficient sensitivity to identify genuinely high-risk individuals. Many systems generate excessive alerts, leading to alert fatigue where nursing staff begin ignoring or overriding system recommendations. Successful implementation requires careful calibration of alert thresholds, integration with existing clinical workflows, staff training, and ongoing feedback mechanisms to refine system performance based on clinical experience.

Ethical Framework for AI-Assisted Risk Detection in Mental Health Nursing

Implementation of artificial intelligence systems for suicide and self-harm risk detection must be grounded in a comprehensive ethical framework addressing core principles applicable to mental health nursing. [23] Ethical considerations include autonomy and informed consent, requiring that patients understand how their data will be used in AI systems and have meaningful opportunity to decline participation. [24] Beneficence and non-maleficence obligate that AI systems are implemented only when evidence supports improved outcomes and risks are actively managed. Confidentiality and privacy protection require robust data governance, secure storage, and compliance with regulatory standards protecting sensitive mental health information. [25] Justice and fairness demand that AI systems perform equitably across all patient populations without perpetuating or amplifying existing disparities.

Transparency and explainability are essential ethical requirements. [26] Mental health nurses must understand how AI systems arrive at risk assessments and recommendations. Black-box systems that provide predictions without clear justification undermine professional accountability and clinical judgment. [24] Systems should provide interpretable explanations for risk assessments, identifying specific clinical or linguistic factors contributing to elevated risk predictions. Additionally, documentation of algorithmic decision-making processes, validation evidence, performance limitations, and known biases should be readily accessible to clinical users.

Professional accountability and human oversight remain fundamental principles. [25] Psychiatric and mental health nurses retain ultimate responsibility for patient assessment and clinical decision-making, regardless of AI system recommendations.

Systems should be designed to support and augment human clinical judgment rather than replace it. Clear accountability frameworks must establish responsibility for monitoring system performance, responding to adverse events, and correcting system errors. When AI-assisted assessments result in missed risk identification or inappropriate intervention decisions, responsibility lies with the clinical team and institution, not the system. [24]

Implications and Recommendations for Mental Health Nursing Practice

Mental health nurses are positioned as essential stakeholders in the responsible development, implementation, and governance of AI systems for suicide and self-harm risk detection. [22] Nursing leadership should prioritize education initiatives to develop AI literacy among nursing professionals, enabling critical evaluation of AI systems and informed participation in implementation decisions. Mental health nursing organizations should advocate for nursing representation in healthcare technology governance structures, ensuring that nursing perspectives shape development priorities and procurement decisions.

Healthcare organizations considering implementation of AI-assisted risk detection systems should adopt value-based approaches, implementing only systems with strong evidence supporting clinical efficacy, conducting rigorous validation studies prior to widespread deployment, and maintaining transparent communication with nursing staff and patients. [8] Implementation should include comprehensive change management support, dedicated education programs, clear protocols for responding to AI-generated alerts, and mechanisms for ongoing feedback enabling system refinement based on clinical experience. Importantly, implementation should be preceded by engagement with mental health nurses regarding feasibility, workflow impact, and clinical concerns.

The therapeutic nurse-patient relationship must remain central to suicide and self-harm risk assessment and intervention. [22] AI systems should be implemented in ways that enhance rather than diminish direct nursing assessment, patient observation, and therapeutic engagement. Time saved through AI-assisted documentation or administrative tasks should be redirected to direct patient contact, supportive presence, and therapeutic communication. Nursing assessment incorporating empathic listening, careful observation of emotional and behavioral presentation, and collaborative exploration of patient concerns remains irreplaceable for comprehensive risk assessment and therapeutic intervention. [25]



Conclusion

Artificial intelligence technologies, particularly machine learning and natural language processing, demonstrate promise for enhancing detection of individuals at elevated risk for suicide and self-harm. [5] Evidence supports the capability of these systems to identify clinically meaningful risk patterns when applied to structured clinical data, unstructured clinical narratives, behavioral observations, and real-time monitoring data. [13] Hierarchical machine learning models integrating clinical history with longitudinal nursing observations show particular promise for identifying hospitalized psychiatric patients at heightened risk, achieving sensitivity and specificity values clinically acceptable for supporting nursing decision-making. However, realizing AI's potential while mitigating risks requires responsible implementation grounded in clear ethical principles, rigorous clinical validation, and unwavering commitment to placing human clinical judgment and therapeutic relationship at the center of care. Mental health nurses have fundamental responsibility and opportunity to shape the trajectory of AI integration in suicide and self-harm risk detection. [21] Through participation in system development, advocacy for ethical implementation standards, engagement in rigorous research validating AI applications in nursing-relevant contexts, and leadership ensuring that technology serves patients and nurses rather than constraining practice, mental health nursing can advance responsible AI integration. The ultimate measure of success will be demonstrated improvements in early identification of at-risk individuals, timely access to evidence-based interventions, and most importantly, prevention of suicide and self-harm while supporting recovery and healing. When implemented with nursing leadership, ethical commitment, and unwavering focus on therapeutic relationships, AI systems can become powerful tools supporting mental health nurses in their essential mission of caring for vulnerable populations at their greatest times of need.

References

- [1] El-Kareem, A., et al. (2023). Self-harm and suicide. In *Psychiatric-Mental Health Nursing* (6th ed.). OpenStax.
- [2] Nursing Times. (2025). Trauma nursing 5: Identifying and assessing patients who self-harm. *Nursing Times*, 121(6), 22-25.
- [3] Eker, F., & Salihu, A. (2023). Application of natural language processing (NLP) in detecting and preventing suicide ideation: A systematic review. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 120, 110637.
- [4] Wang, Y., et al. (2023). Suicide prediction with natural language processing of electronic health records. *MedRxiv*. Retrieved from <https://www.medrxiv.org/content/10.1101/2023.09.28.23296268>
- [5] Cohen, J., Wright-Berryman, J., Rohlf, L., Trocinski, D., Daniel, L., & Klatt, T. W. (2022). Integration and validation of a natural language processing machine learning suicide risk prediction model based on open-ended interview language in the emergency department. *Frontiers in Digital Health*, 4, 818705.
- [6] Eker, F., & Salihu, A. (2023). Application of natural language processing (NLP) in detecting and preventing suicide ideation: A systematic review. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 120, 110637.
- [7] Stanley, B., et al. (2018). A feasibility study using a machine learning suicide risk prediction model based on open-ended interview language in adolescent therapy sessions. *Archives of Suicide Research*, 22(3), 398-416.
- [8] Sanz-Barbero, B., et al. (2023). Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Psychiatry Research*, 221, 156-164.
- [9] Tsui, F. R., Shi, L., Ruiz, V., et al. (2021). Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. *JAMIA Open*, 4(1), oaab011.
- [10] Obeid, J. S., Dahne, J., Christensen, S., Howard, S., Crawford, T., Frey, L. J., & Stecker, T. (2020). Identifying and predicting intentional self-harm in electronic health record clinical notes: Deep learning approach. *JMIR Medical Informatics*, 8(7), e17784.
- [11] Ghosh, S., et al. (2024). A hierarchical machine learning model for predicting self-harm and suicidal behaviour in hospitalised patients with schizophrenia using clinical history and nursing observations. *Frontiers in Psychiatry*, 15, 1760380.
- [12] Kessler, R. C., et al. (2024). Predicting suicidal and self-harm ideation using ecological momentary assessment: Deep learning analysis in a general population sample. *BMC Psychiatry*, 26, 81.
- [13] Torok, M., et al. (2023). Assessment of machine learning algorithms in national data to classify the risk of self-harm



- among young adults in hospital: A retrospective study. *Journal of Affective Disorders*, 323, 1-9.
- [14] Manchini, L. (2024). Benchmarking action recognition models for self-harm detection in studio and real-world datasets. *Journal of Healthcare Informatics Research*, 8(2), 145-158.
- [15] Patel, P. P., et al. (2024). Regulating AI in mental health: Ethics of care perspective. *Frontiers in Psychiatry*, 15, 1334567.
- [16] Melia, R., et al. (2023). Ethical and regulatory considerations in the use of AI and machine learning in nursing: A systematic review. *Journal of Nursing Administration*, 53(1), 12-19.
- [17] Thornicroft, G., et al. (2024). Ethical and legal considerations in healthcare AI: Innovation and policy for safe and fair use. *Lancet Psychiatry*, 11(2), 134-142.
- [18] Langarizadeh, M., et al. (2024). Ethical decision-making for AI in mental health: The Integrated Ethical Approach for Computational Psychiatry (IEACP) framework. *Nature Neuroscience*, 27(3), 445-456.
- [19] Ferrando, S. J., Finklestein, D. M., & Kontos, N. (2023). Ethical decision-making guidelines for mental health clinicians in the artificial intelligence (AI) era. *Journal of Psychiatric Practice*, 29(4), 234-242.
- [20] Yates, S. (2025). The ethical implications of AI in nursing. *Nurse.com*. Retrieved from <https://www.nurse.com/blog/ethical-implications-ai-nursing-nsp/>
- [21] Langarizadeh, M., et al. (2024). Ethical decision-making for AI in mental health: The Integrated Ethical Approach for Computational Psychiatry (IEACP) framework. *Nature Neuroscience*, 27(3), 445-456.
- [22] Patel, P. P., et al. (2024). Regulating AI in mental health: Ethics of care perspective. *Frontiers in Psychiatry*, 15, 1334567.
- [23] Thornicroft, G., et al. (2024). Ethical and legal considerations in healthcare AI: Innovation and policy for safe and fair use. *Lancet Psychiatry*, 11(2), 134-142.
- [24] Melia, R., et al. (2023). Ethical and regulatory considerations in the use of AI and machine learning in nursing: A systematic review. *Journal of Nursing Administration*, 53(1), 12-19.
- [25] Ferrando, S. J., Finklestein, D. M., & Kontos, N. (2023). Ethical decision-making guidelines for mental health clinicians in the artificial intelligence (AI) era. *Journal of Psychiatric Practice*, 29(4), 234-242.
- [26] Yates, S. (2025). The ethical implications of AI in nursing. *Nurse.com*. Retrieved from <https://www.nurse.com/blog/ethical-implications-ai-nursing-nsp/>
- [27] Wright, N., et al. (2023). The psychiatric mental health nurse's ethical considerations regarding the use of coercive measures: A qualitative interview study. *International Journal of Mental Health Nursing*, 32(1), 156-167.

Conflict of Interest

The author declares that there is **no conflict of interest** related to this study.

Source of Funding

The study was **self-funded**, and no external financial assistance was received.