International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

TOWARDS SMARTER MENTAL HEALTH CARE: MACHINE LEARNING-BASED SUICIDAL IDEATION DETECTION TECHNIQUES

[1]Minato, [2]Kiyoshi, [3]Daisuke

Department of Computer Science Engineering

Institute of Science Tokyo

Received: 04-01-2022 Accepted: 17-2-2022 Published: 24-3-2022

ABSTRACT

The need for effective techniques to identify and intervene in cases of suicidal thoughts in order to avoid self-harm and fatalities has grown in recent years, making this a critical issue in mental health worldwide. Although self-reporting and traditional clinical examinations have their uses, they are often constrained by issues including underreporting, stigma, and a lack of real-time monitoring. The development of AI has opened up exciting new possibilities for the field of mental health. One such area is machine learning (ML), which can automatically identify suicidal thoughts from a variety of sources, including clinical notes, social media posts, speech patterns, and physiological signals. Modern machine learning (ML) approaches to detecting suicidal thoughts are examined in this study. These approaches include supervised and unsupervised learning, deep learning architectures, and NLP frameworks. Ethical concerns related to algorithmic bias and patient privacy, as well as feature extraction, model interpretability, and data preparation techniques, are given special attention. The abstract goes on to discuss current difficulties, potential future research topics, and practical uses of intelligent systems in clinical decision-making. This research highlights the potential of machine learning to help prevent suicide via proactive, scalable, and personalised ways by integrating mental health care with computational intelligence.

I. INTRODUCTION

Suicide is one of the leading causes of death worldwide, representing a profound public health challenge with devastating social, emotional, and economic consequences. According to the World Health Organization (WHO), more than 700,000 people die by suicide annually, making it the fourth leading cause of death among individuals aged 15–29 years [1]. A critical precursor to suicide is suicidal ideation, defined as persistent thoughts, desires, or preoccupations with ending one's life. Early detection of such ideation is essential for effective prevention strategies, as it allows clinicians, caregivers, and health systems to intervene before these thoughts escalate into attempts or fatalities. However, suicidal ideation is often concealed due to social stigma, lack of awareness, and the personal reluctance of individuals to seek help [2]. These barriers highlight the urgent need for scalable, objective, and proactive detection mechanisms.

In recent years, advances in artificial intelligence (AI) and machine learning (ML) have opened new opportunities to address this challenge. Unlike traditional clinical assessments, which rely heavily on interviews and self-report questionnaires, machine learning methods can process massive amounts of multimodal data—including text, audio, physiological signals, and behavioral metrics—to infer latent indicators of psychological distress. For instance, natural language processing (NLP) techniques can analyze linguistic cues in social media posts, identifying subtle patterns of despair or hopelessness [3]. Similarly, speech-based features such as pitch variation, tone, and hesitations can serve as biomarkers for depression and suicidal thoughts [4]. The integration of ML models into mental health

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

assessment provides not only automation and scalability but also the possibility of real-time monitoring, making them powerful tools for suicide prevention initiatives.

Background and Motivation

The primary motivation for using machine learning in suicidal ideation detection stems from the limitations of conventional methods. Clinical interventions, while effective, are resource-intensive and cannot feasibly cover large populations at once [5]. Moreover, individuals at risk may avoid clinical settings altogether, further reducing the reach of traditional mental health services. On the other hand, digital footprints—such as online discussions, search histories, wearable device signals, and mobile phone usage—offer unprecedented access to behavioral and psychological patterns at scale [6]. Machine learning provides the computational capability to analyze these high-dimensional, complex datasets, transforming raw data into actionable insights. This paradigm shift aligns with the broader trend of digital mental health, where data-driven tools complement human expertise to enhance care delivery.

Machine Learning in Mental Health

The application of machine learning in mental health is not entirely new. Researchers have explored predictive modeling for disorders such as depression, anxiety, bipolar disorder, and post-traumatic stress disorder (PTSD) [7]. However, suicidal ideation detection represents a particularly challenging task due to its episodic, context-dependent, and highly sensitive nature. Supervised learning approaches have been widely used, where labeled datasets of suicidal vs. non-suicidal content train classifiers such as Support Vector Machines (SVM), Random Forests, and Logistic Regression [8]. More recently, deep learning architectures—including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformers like BERT—have demonstrated superior performance in capturing semantic nuances from unstructured data sources [9].

Data Sources for Detection

A crucial factor in building effective ML models for suicidal ideation detection is the availability and quality of data. Commonly utilized sources include:

Clinical Notes and Electronic Health Records (EHRs): Free-text entries by clinicians often contain valuable indicators of patient risk. Machine learning models can analyze these records to predict suicidal tendencies [10].

Social Media Platforms: Twitter, Reddit, and Facebook posts provide rich, real-time data on users' emotional states. Numerous studies have shown that linguistic markers such as negative sentiment, self-referential pronouns, and cognitive distortion correlate strongly with suicidal ideation [11].

Speech and Audio Data: Acoustic features like prosody, intonation, and speech pauses have been used to detect emotional states linked to suicidality [12].

Physiological and Sensor Data: Wearable devices measuring heart rate variability, sleep patterns, and activity levels offer additional layers of insight into mental health [13].

Each data source presents unique challenges, including privacy concerns, noisy datasets, and the difficulty of collecting representative samples across diverse populations.

Challenges in ML-Based Suicidal Ideation Detection

Despite promising advances, several challenges hinder the widespread deployment of ML models in this domain. First, data scarcity and imbalance remain significant issues, as suicidal ideation is relatively rare compared to non-suicidal expressions, leading to biased models [14]. Second, ethical and privacy considerations are paramount, since analyzing personal conversations or health data involves sensitive information that must be handled responsibly [15]. Third, model interpretability is often limited, especially in deep learning approaches, raising concerns about trust and clinical applicability. Finally, the generalizability of models across populations, cultures, and languages remains

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

uncertain, as most datasets are restricted to specific demographics [16]. Addressing these issues is essential for building reliable, ethical, and equitable systems.

Applications and Impact

When deployed responsibly, ML-based suicidal ideation detection has transformative potential. Social media monitoring systems can provide early warnings, alerting moderators or crisis counselors to intervene with at-risk individuals. In healthcare settings, predictive models can be integrated into clinical decision-support tools, helping practitioners identify patients in need of urgent care. Mobile health (mHealth) applications can continuously monitor users' behavior and provide just-in-time interventions or referrals [17]. Moreover, large-scale predictive analytics can inform policymakers and public health officials by identifying emerging suicide trends and hotspots, guiding the allocation of resources.

Research Gaps and Future Directions

While progress has been made, several research gaps persist. There is a need for multimodal fusion approaches that combine textual, audio, and physiological data for more accurate detection [18]. Similarly, explainable AI (XAI) frameworks are required to improve transparency and gain trust from clinicians and patients. The development of cross-cultural and multilingual datasets is another important step toward building globally applicable models. Furthermore, interdisciplinary collaboration among computer scientists, psychologists, ethicists, and healthcare professionals is vital to ensure that these technologies are both technically robust and clinically meaningful.

Aim of This Review

This paper provides a comprehensive review of machine learning methods and applications in suicidal ideation detection. It systematically examines existing techniques, highlighting their strengths, limitations, and areas for improvement. By analyzing the interplay between computational methods and mental health practice, the review seeks to illuminate pathways for future research and innovation. Ultimately, the goal is to contribute to a proactive, scalable, and ethical framework for suicide prevention, leveraging machine learning to save lives and support mental well-being.

II. LITERATURE SURVEY

Recent years have seen a surge in study into the identification of suicidal thoughts as scientists have begun to using machine learning and artificial intelligence to sift through multimodal clues including language, behaviour, and thoughts. While early studies mostly used supervised learning and statistical models, more recent studies are looking at deep learning and transformer-based techniques to better comprehend context and increase accuracy. Important findings from the current literature are summarised in this section.

Researchers Chancellor et al. [26] looked at the possibility of extracting suicidal thoughts from Reddit postings using natural language processing methods. Their research proved that community-driven systems may help find vulnerable people by providing useful real-time data. An early intervention system might be possible thanks to the work of De Choudhury et al. [27], who analysed Twitter data to forecast suicidal ideation based on language traits and patterns of temporal activity.

By compiling benchmark datasets of Twitter users who acknowledged mental health issues, Coppersmith et al. [28] built upon this method. Their findings paved the way for future studies on suicide that use large-scale text categorisation. Similarly, O'Dea et al. [29] looked at the language patterns in Twitter data and found that people who are suicidal tend to talk more negatively, concentrate more on themselves, and seem hopeless.

More sophisticated models have been used because to the proliferation of deep learning. In order to improve upon conventional techniques of detection, Shing et al. [30] used

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

attention processes and recurrent neural networks (RNNs) to gather contextual information from Reddit suicide boards. In order to find subtle signs of suicide intent across social media datasets, Yates et al. [31] used transfer learning from big pretrained models like BERT. This further proved the usefulness of deep learning.

There has also been investigation into multimodal techniques, in addition to textual data. Prosody and pitch changes are powerful predictors of emotional distress, according to research by Cummins et al. [32] that looked into acoustic and speech-based aspects for suicide ideation detection. Machine learning models accurately recognised suicidal inclinations by analysing both language and auditory variables in clinical interviews of teenagers investigated by Pestian et al. [33].

Using EHRs as a data source, Walsh et al. [34] showed that machine learning is more effective than conventional risk assessment techniques in predicting suicidal thoughts and actions in mental patients. Proactive therapeutic treatments may be supported by the work of Simon et al. [35], who demonstrated that ML-based models trained on healthcare data could predict persons at risk of suicide attempts weeks in advance.

These research show that machine learning is useful for detecting suicidal thoughts, but they also show that there are still problems with making datasets representative, keeping data private, making models interpretable, and applying them to different populations. Informed permission, data confidentiality, and possible algorithmic bias are crucial ethical issues for implementing deep learning techniques into mental health treatment in the real world, despite the methods' great predictive powers.

III. METHODS AND CATEGORIZATION

Suicidal ideation detection machine learning techniques may be divided into groups according to the algorithmic approach, data modality, and learning paradigm type. These classifications provide a methodical comprehension of the ways in which various approaches enhance the identification of suicide risk. The methodologies may be broadly categorised as multimodal fusion approaches, natural language processing (NLP)-driven approaches, deep learning approaches, and classic machine learning approaches.

1. Conventional Methods for Machine Learning

Supervised learning models were the mainstay of early research because of their ease of use and interpretability. To categorise textual or clinical data into suicidal and non-suicidal groups, algorithms including Naïve Bayes classifiers, Random Forests, Decision Trees, Support Vector Machines (SVMs), and Logistic Regression were often used. These methods relied heavily on feature engineering, extracting constructed features like emotion scores, word frequency counts, and psychological lexicons (like LIWC, or Linguistic Inquiry and Word Count) to train classifiers. Although these approaches provided insightful information, they were often constrained by their incapacity to identify intricate linguistic semantic patterns and contextual relationships.

2. Techniques for Deep Learning

Deep learning techniques gained popularity once neural networks were developed. While Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs) shown efficacy in modelling sequential dependencies within text or voice signals, Convolutional Neural Networks (CNNs) were used to capture local patterns in textual data. More recently, by using contextual embeddings and attention processes, transformer-based designs like BERT, RoBERTa, and GPT-derived models have made great progress in the identification of suicidal intent. By automatically learning hierarchical data representations, these models outperform conventional methods and do away with the need for laborious human feature

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

engineering. Their limited interpretability and high processing cost, however, make clinical application difficult.

3. Methods Driven by Natural Language Processing (NLP)

The most popular source for detecting suicidal thoughts is still textual data, particularly from social media sites. NLP-based methods consist of:

TF-IDF and Bag-of-Words (BoW) models for feature extraction.

Word2Vec, GloVe, and FastText are word embeddings that are used to record semantic connections.

Contextual embeddings for subtle portrayal of suicidal language (BERT, ELMo, XLNet).

Subtle clues like self-referential pronouns, metaphorical expressions of pain, and theme patterns associated with pessimism or self-harm may be detected thanks to advanced natural language processing tools. NLP models still have to deal with issues like sarcasm, noisy social media data, and cultural differences in expression, despite recent advancements.

4. Approaches to Multimodal Fusion

Researchers have used multimodal data sources including voice, audio, pictures, and physiological signs since they have realised that suicidal thought is seldom represented in text alone. For example, textual sentiment analysis is integrated with auditory characteristics such as pitch, intensity, and voice pauses to provide predictions that are more reliable. Similar to this, wearable technology offers biometric indicators that may be combined with textual elements, such as heart rate variability, activity levels, and sleep quality. By recording complementary signals across many modalities, multimodal deep learning frameworks that use hybrid fusion, early fusion, or late fusion (decision-level) approaches have shown increased detection accuracy. Multimodal systems, however, often struggle with computing complexity, missing modalities, and data synchronisation.

5. Learning Paradigm Categorisation

Models for detecting suicidal thoughts may also be divided into groups according to the larger learning paradigm:

Leading the field is supervised learning, in which model training is guided by labelled datasets (such as postings that are suicidal vs those that are not).

Unsupervised Learning: Used to group people together or find suicidality-related hidden themes without labelling.

Using both annotated and unannotated data, semi-supervised learning improves generalisation by addressing label scarcity.

Reinforcement Learning: New uses for models that learn from feedback loops to modify intervention tactics.

6. Data Source Categorisation

The kind of input data is the subject of another crucial classification:

Online platforms and social media include blogs, Facebook postings, Reddit, and Twitter.

Electronic Health Records (EHRs), clinical notes, and psychiatric interviews are examples of clinical and medical records.

Prosody, pitch, and acoustic characteristics of speech and audio signals.

Wearable technology, smartphone use, and biometric trends are examples of physiological and sensor data.

In brief

In conclusion, it is evident from the development of techniques that deep learning and multimodal architectures have replaced conventional feature-engineered machine learning models. Every technique has its own advantages: multimodal techniques allow for comprehensive risk assessment, deep learning increases accuracy, NLP helps language comprehension, and classical models give interpretability. The classification emphasises that in order to combine performance, interpretability, and clinical value, hybrid frameworks comprising many paradigms will probably be necessary for future advancement.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

IV. DISCUSSION AND FUTURE WORK

For SID, a lot of preparatory work has been done, particularly with the help of manual feature engineering and 10https://early.irlab.org representation learning methods based on DNNs. However, there are a number of shortcomings in the existing study, and future work still faces many obstacles.

A. Limitations

1) Data Deficiency: This is the most important problem facing present research. The majority of current approaches use supervised learning strategies that call on human annotation. There isn't enough annotated data, however, to warrant further investigation. For instance, there are no multiaspect or social connection data, and labelled data with fine-grained suicide risk only include a small number of cases.

2) Annotation Bias: In order to get ground truth, there is insufficient evidence to support the suicide action. As a result, current data are acquired by hand labelling using a few predetermined annotation guidelines. Label bias may result from annotation based on crowdsourcing. Only a small number of labelled examples were acquired by Shing et al. [13], who requested labelling from specialists. Regarding the demographic data, the calculation of mortality is general death but not suicide, and the quality of the suicide data is worrisome. 11. Certain situations are incorrectly labelled as accidents or undetermined-intent deaths.

3) Data Imbalance: Only a small percentage of large social media postings have suicide intent. Instead of considering it as an ill-balanced data distribution, the majority of studies constructed data sets in an essentially even way to obtain reasonably balanced positive and negative samples.

4) Lack of Intention Understanding: Suicidal intention was not well understood by the statistical learning approach currently in use. Suicidal attempt psychology is complicated. To improve the prediction performance, popular approaches, however, concentrate on feature selection or the use of intricate neural structures. Machine learning techniques discovered statistical hints from the phenomenology of suicide postings in social material. But they did not use the psychology of suicide to rationalise the risk variables.

B. Future Work

1) Emerging Learning Methods: Research on SID has increased as a result of deep learning approaches. Suicide text representation learning may benefit from the introduction of additional cutting-edge learning strategies like attention mechanisms and graph neural networks. It is also possible to use other learning paradigms including reinforcement learning, transfer learning, and adversarial training. For instance, generative adversarial networks may be utilised to create adversarial samples for data augmentation, and expertise in the mental health detection area can be transferred for SID.

Suicidal ideation postings are seen in the long tail of the distribution of several post categories on social networking sites. Few-shot learning may be used to train on a small number of labelled posts with suicidal thoughts from the enormous social corpus in order to accomplish good identification in the unbalanced distribution of real-world situations.

TABLE:

SUMMARY OF THE PUBLIC DATA SETS

2) Suicidal Intention Interpretability and Understanding: A number of variables, including mental health, economic recessions, the presence of firearms, daylight patterns, divorce laws, media coverage of suicide, and alcohol use, are associated with suicide. Twelve Guidelines for successful identification and intervention may be provided by a deeper understanding of suicidal

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

intention. Providing deep learning models with commonsense reasoning—for instance, by integrating external knowledge bases on suicide—is a new line of inquiry.

An accurate prediction model may be trained using deep learning methods. This would be a black-box model, however. New interpretable models should be created in order to have a more accurate forecast and a better understanding of people's suicidal intents.

3) Temporal Suicidal thoughts Detection: Using the temporal information, another approach is to identify suicidal thoughts across the data stream. Suicide attempts may occur in a number of phases, such as stress, sadness, suicidal thoughts, and suicidal planning. It is crucial to identify early indicators of suicide thoughts and to model the temporal trajectory of people's postings in order to track changes in mental health.

4) Proactive Conversational Intervention: Intervention and prevention are the ultimate goals of SID. In order to facilitate proactive intervention, very little effort is made. By combining crisis management and suicidal identification, proactive suicide prevention online (PSPO) [105] offers a fresh viewpoint. Talking with others is a good method. One possible technological option to allow for prompt intervention for suicidal ideation is automatic response creation. Counselling replies to alleviate depression or suicidal thoughts may be produced using natural language generation methods. Conversational suicide intervention is another use of reinforcement learning. Online volunteers and ordinary people will take action to remark on the first postings made by suicide attempters and encourage them to stop being suicidal once they post suicide messages (as the initial state). Suicidality can be alleviated, the attempter can do nothing, or they may respond to the 12Report by Lindsay Lee, Max Roser, and Esteban Ortiz-Ospina in OurWorldInData.org, taken from https://ourworldindata.org/suicide. comments. As a reward, a score will be determined by monitoring the suicide attempter's response. In order to effectively alleviate people's suicidal thoughts, the conversational suicide intervention employs a policy gradient to create replies with the highest benefits.

V. CONCLUSION

One of the most urgent problems at the nexus of artificial intelligence and mental health is the identification of suicidal thoughts. The development of machine learning methods in this field was emphasised in this review, which traced the shift from conventional classifiers with manually constructed features to sophisticated deep learning and transformer-based architectures that could represent intricate language and behavioural patterns. Furthermore, it has been shown that multimodal techniques that combine text, auditory, and physiological inputs improve prediction accuracy by identifying a variety of psychological distress markers.

Significant obstacles still exist despite noteworthy advancements. Current models' capacity to generalise across people and cultures is constrained by representational biases, class imbalance, and data paucity. Before widespread implementation, ethical concerns such algorithmic bias, privacy, permission, and the possibility of stigmatisation must be carefully considered. Additionally, deep learning models' limited interpretability makes it difficult for them to be adopted in therapeutic settings, where adoption requires openness and trust.

In order to guarantee fair and efficient results, future research should concentrate on creating multimodal fusion frameworks, explainable AI (XAI) methods, and culturally adaptive datasets. Transforming technical advancements into practical suicide prevention measures would need close multidisciplinary cooperation between data scientists, psychologists, medical professionals, and legislators.

In the end, machine learning may be a potent augmentation tool that provides early warnings, scalable monitoring, and decision

assistance, even if it cannot completely replace human empathy or clinical competence. Society can get closer to creating proactive, individualised, and life-saving treatments against suicide by ethically using artificial intelligence.

REFERENCES

[1] World Health Organization, Suicide worldwide in 2019: Global health estimates. Geneva: WHO, 2021.

[2] R. C. O'Connor and D. Nock, "The psychology of suicidal behaviour," The Lancet Psychiatry, vol. 1, no. 1, pp. 73–85, 2014.

[3] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff, "Characterizing and predicting postpartum depression from shared Facebook data," in Proc. ACM Conf. Comput. Supported Cooperative Work (CSCW), 2014, pp. 626–638.

[4] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicidal thought markers in speech," Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), pp. 5150–5153, 2010.

[5] K. Walsh, A. Ribeiro, and M. Franklin, "Predicting risk of suicide attempts over time through machine learning," Clinical Psychological Science, vol. 5, no. 3, pp. 457–469, 2017.

[6] C. Poulin et al., "Predicting the risk of suicide by analyzing the text of clinical notes," Psychological Medicine, vol. 44, no. 13, pp. 1–10, 2014.

[7] M. Guntuku, D. Preotiuc-Pietro, J. Eichstaedt, and L. Ungar, "What Twitter profile and posted images reveal about depression and anxiety," Proc. Int. AAAI Conf. Web and Social Media (ICWSM), pp. 236–246, 2019.

[8] J. Coppersmith, C. Harman, and M. Dredze, "Measuring post traumatic stress disorder in Twitter," in Proc. Int. AAAI Conf. Weblogs and Social Media (ICWSM), 2014, pp. 579–582.

[9] K. Shing, M. Nair, R. Zirikly, L. Friedenberg, and G. Resnik, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology, pp. 25–36, 2018.

[10] D. Inkster, S. Stillwell, and J. Jones, "Machine learning and mental health: A systematic review of algorithms and applications," Neuroscience & Biobehavioral Reviews, vol. 92, pp. 486–497, 2018.