Vol 6, Issue 2, July 2026 eISSN: 2805-444X

JOURNAL OF COMMUNICATION, LANGUAGE AND CULTURE

From Conversations to Insights: Analysing Social Networks for Early Mental Health Detection - A Systematic Review of Causal Inference and Deep Learning

Umber Nazir[1*], Nur Shazwani Kamarudin[1], Mazlina Abdul Majid[2]

[1] Faculty of Computing, Universiti Malaysia Pahang, Al-Sultan Abdullah, Pekan, Pahang, Malaysia
[2] Centre for Artificial Intelligence & Data, Universiti Malaysia Pahang, Al-Sultan Abdullah, Pekan, Pahang, Malaysia

*Corresponding author: pez24016@adab.umpsa.edu.my; ORCID iD: 0000-0002-8351-989X

ABSTRACT

Early detection of mental health issues is crucial for timely intervention, reducing the severity of conditions, and improving overall well-being. Social networks have emerged as valuable platforms for identifying mental health issues, thanks to user-generated content and social interactions. Artificial intelligence (AI), especially causal inference and deep learning, has great potential for analysing large-scale social media data. It helps identify patterns and relationships that enable earlier and more accurate prediction of mental health issues. The paper aims to systematically review academic articles on the applications of AI, including deep learning and causal inference in social networks for early mental health detection. The systematic review initially considered 1,018 academic articles retrieved from major scholarly databases, including IEEE Xplore, Scopus, and ScienceDirect. The review was conducted in accordance with the PRISMA framework to ensure a transparent and rigorous screening and selection process. After careful review, the articles were filtered down to 90 for full analysis to present a classification framework based on four dimensions: Applications in mental health, methods/techniques, datasets used, and challenges. It was identified that AI continues to significantly outperform humans in terms of accuracy, efficiency, and early intervention for mental health detection. Methods and techniques map directly to relevant keywords such as AI applications, causal inference, deep learning, and social networks. Although AI shows great promise in mental health detection, challenges such as data bias, privacy risks, and a lack of model interpretability remain. Combining causal inference with deep learning can create personalised mental health interventions. Still, future research must prioritise explainable AI, privacy-preserving methods, and ethical data collection to ensure responsible, transparent AI applications in mental health care.

Keywords: Structural Causal Models, Machine learning, artificial intelligence, social network

Received: 29 November 2025, **Accepted:** 14 May 2026, **Published:** 30 July 2026

1.0 Introduction

Early mental health intervention and detection are important for reducing the progression of mental health conditions and enhancing recovery outcomes (Torous et al., 2025). According to King et al. (2024), depression, anxiety, and suicidal ideation are critical to address, but it can be timely to intervene by identifying early warning signs and symptoms. Studies emphasise that early detection considerably improves the efficacy of therapeutic interventions and decreases the long-term effects of mental health disorders (Patel et al., 2018). The frequency of mental health issues globally brings attention to the need for tools and techniques to detect these symptoms at an early stage. According to Patel et al. (2018), numerous studies highlight that recognising early warning signs, such as changes in mood, behaviour, sleep patterns, or social withdrawal, can enable more targeted and proactive interventions. Early detection facilitates more personalised treatment approaches, reduces healthcare costs, and mitigates long-term psychological, social, and economic consequences (Hassan & Omenogor, 2025).

Furthermore, with the rising global burden of mental health issues, there is an increasing demand for scalable, accessible, and efficient tools that can assist in identifying at-risk individuals before crises occur (Sah et al., 2025). This underscores the need for advanced methodologies, including the integration of technology and data-driven approaches, to support early screening and monitoring of mental health conditions across diverse populations. This paper is a systematic review of the academic literature exploring the application of artificial intelligence, deep learning, and causal inference techniques in social network analysis for the identification of mental health disorders. More specifically, this review intends to evaluate the application of AI techniques and algorithms for the identification of psychological distress, depression, anxiety, and other mental health disorders in the analysis of behavioural data and content use by users on social networking websites. By reviewing the techniques employed, the data analysed, and the challenges identified, this study provides a comprehensive review of the potential of the emerging field of AI-based mental health monitoring using social network data. Social networks have become an important source of data for mental health research by providing access to large-scale, real-time expressions of emotions, behaviours, and social interactions. In contrast to conventional clinical or survey-based data, Twitter, Reddit, and Quora provide an endless stream of unprompted information about users' psychological conditions, which can be analysed at the population level to identify patterns in mental health. According to Chancellor et al. (2019), linguistic and sentiment features extracted from social media, combined with machine learning methods, can be effectively used to support early diagnosis of conditions such as depression. Language patterns and sentiments from Twitter posts have been effectively applied to develop predictive models for identifying depression, and machine learning algorithms demonstrate superior performance compared to conventional baseline methods through the exploitation of contextual and interaction-based features(Yang & Li, 2025).

Extensive studies of Reddit conversations, especially those made in subreddit communities devoted to mental health, also demonstrate how users exchange subtle emotional moments and feelings related to suicidality and distress. Based on advanced natural language processing, Bauer et al. (2024) have found the existence of specific linguistic patterns, such as expressions of hopelessness, burden, and seeking support, which can be aligned with clinical theories of suicidal behaviour and with the perspectives of social discourse as a digital genetic observation of mental health risk. This kind of work is ideal because it illustrates how communication on social platforms goes beyond the presence of keywords and can indicate latent psychological patterns that can be used to track and provide early notifications (Bauer et al., 2024).

In addition to data availability, social media is important for communicative and educational functions in mental health awareness and intervention. Munir and Ahmed (2025) demonstrated that platforms are now being used not only to monitor mental health symptoms but also to share health information, provide peer support, and run campaigns to educate people. Health communication studies demonstrated that social media communication can be used to build health awareness and health education, and that users and health professionals can use the media to provide evidence-based information, personal experiences, and coping strategies; this involvement has been linked to enhanced health literacy and use of the media in care processes (Munir & Ahmed, 2025). Furthermore, meta-analytic data on social-media-based mental health initiatives show that even simple, systematic support, provided with the help of the social network, may yield observable progress in depression and anxiety symptoms. It demonstrates the potential of digital communication as a type of therapy when used within a clinical framework (Zhang et al., 2025).

Nonetheless, social network communication is complicated: social networks offer a wide audience and an immediate nature to this process; however, this result can increase social comparison, cyberbullying, and emotionally charged content, which may harm vulnerable populations. Recent systematic reviews by Cabezas-Klinger et al. (2025) demonstrated the connection between exposure to negative or harmful information and psychosocial processes of online interaction and anxiety, depressive symptoms, and disturbed sleep patterns among adolescents and young adults, and underscore that the quality of communication, rather than its quantity, is actually important to mental health outcomes. However, social media is also an important area of inquiry and a communication strategy in the mental health industry, where the ability to detect at an earlier stage, intervene, and provide support to the population is unmatched in terms of opportunity when properly utilised (Cabezas-Klinger et al., 2025).

Causal inference and deep learning techniques have evolved into radical approaches for analysing intricate, multifaceted mental health data. Causal inference helps establish cause-and-effect relationships, allowing researchers to distinguish factors contributing to mental health conditions and to design effective interventions. As described by Pearl & Mackenzie (2018), Methods such as propensity score matching, instrumental variables, and structural causal models are widely applied to unravel causation from correlation. On the contrary, Deep learning excels at handling large-scale, unstructured data, such as social media posts, by extracting features and patterns using neural network architectures like CNNs, RNNs, and transformers. According to the latest research, integrating these techniques can enhance the accuracy and reliability of mental health predictions (Althoff et al., 2016; Pearl & Mackenzie, 2018).

Jiao et al. (2024) conducted an in-depth survey of the combination of causal inference and deep learning, offering an optimistic approach to improve early mental health detection by addressing the shortcomings of traditional methods. Deep learning models are good at detecting patterns and predicting from large datasets, they are limited in their ability to establish causal relationships, which are important in understanding the root causes of mental health conditions (Zhao et al., 2022). On the other hand, causal inference further enhances this by identifying cause-and-effect relationships (Jiao et al., 2024). Schulam and Saria (2019) highlighted the potential of combining these techniques to improve early mental health detection and enhance the reliability of AI tools and solutions.

Recent strategies for early mental health detection face major challenges and limitations, as discussed by Jiao et al. (2024), including false correlations, data quality issues, and model interpretability. Deep learning models are prone to overestimating correlations that may not reflect causal relationships, leading to unreliable predictions and the risk of bias. Moreover, social media data is characterised by noise, missing information, and a lack of labelled datasets, which undermine the accuracy and robustness of models (Zhao et al., 2022). Challenges include ethical concerns like data privacy and the misuse of sensitive information. As a solution, Lipton (2018) and Rudin (2019) advised combining causal methods to infer causal relationships and deep learning techniques to handle noisy, unstructured data efficiently.

2.0 Literature Review

2.1 Artificial Intelligence (AI) for early mental health intervention

Artificial Intelligence plays a vital role in early mental health intervention by utilising machine learning (ML) techniques to predict and detect mental health conditions, thereby enabling timely intervention and improved well-being (Chikersal et al., 2020; Hazra-Ganju et al., 2023; Kuhail et al., 2024). AI-enabled models and techniques utilise Deep learning, natural language processing, predictive analysis, and causal inference methods for early mental health interventions (Zhao et al., 2022).

Oei et al. (2023) developed an ML model using decision trees and random forests to detect mental health impacts in stroke survivors, investigating key factors for early intervention. However, Myee et al. (2024) used sentiment analysis and emotion detection on social media posts (Twitter and Reddit) to predict signs of depression. There are several studies on depression detection using ML techniques and AI tools to improve early mental health interventions (Chikersal et al., 2020; Hazra-Ganju et al., 2023; Kuhail et al., 2024; Uddin et al., 2022). Sandulescu et al. (2024) integrated the Internet of Medical Things (IoMT) with AI by analysing physiological and behavioural data to develop predictive models for emotion detection in neurodegenerative diseases, enabling proactive mental health monitoring. Yoon (2024) introduced AI-based digital treatment methods customised for youth mental health care, powered by reinforcement learning, including chatbots and digital therapy.

Previous studies emphasise AI's radical possibilities for early mental health intervention. ML and DL models are integrated with NLP, IoT monitoring, emotion detection, and clinical decision support tools to develop customised, scalable, and preventive mental health solutions (Uddin et al., 2022).

2.2 Causal inference in mental health

Causal Inference determines the cause-and-effect relationship between variables rather than merely detecting associations (Ohlsson & Kendler, 2019). In mental health research, it is important to gain insights into the underlying factors of mental health conditions (Hogan, 2019). By distinguishing between correlation and causation, causal inference techniques can help identify risk factors, discover prospective approaches to mental disorders, and lead to effective and efficient interventions (Pearl & Mackenzie, 2018). According to Pearl (2009), causal inference provides an important tool for understanding the complex infrastructure of environmental, social, and biological factors that influence mental health.

There are several methods commonly applied in causal inference to analyse mental health outcomes. Propensity score matching (PSM) is a commonly used technique that seeks to estimate the impact of intervention (Ohlsson & Kendler, 2019). PSM is useful when random assignment is unrealistic, as it helps balance covariates between the treatment and control groups (Ohlsson & Kendler, 2019). Structural causal models (SCMs) are another method for describing the relationships between variables, allowing researchers to estimate intervention impacts (Imbens & Rubin, 2015; Pearl, 2009). SCM offers an intensive approach to analysing the impact of social networks on mental health (Abid et al., 2025). Beyond statistical or DL models, SCMs provide clarity, causal detection, and counterfactual analysis, making them a powerful tool for researchers and policymakers in mental health studies (Kawam et al., 2025).

Marchezini et al. (2022) applied counterfactual analysis and latent-variable modelling by integrating SCMs with DL and PSM to demonstrate that intervention groups were comparable when analysing mental health treatments. Several techniques and methods are used in previous studies, including Ghaffari et al. (2024)'s work on user activity analysis of a music application, which uses regression models to determine the relationship between social connection and diversity in listening choices. Ghaffari et al. (2024) used Difference-in-differences (DiD) techniques to do the behavioural comparison before and during COVID-19.

NLP techniques are commonly used in social media data to extract causal relationships from text. Doan et al. (2019) used NLP-based causal inference, leveraging linguistic structures, to detect causal statements in Twitter data. Rather than relying on traditional Causal Inference (CI) methods (SCMs/DiD), Doan et al. (2019) used text mining techniques to identify causal relationships in social media data. These techniques and methods enable researchers to create more accurate causal relationships from data, providing in-depth insights into factors influencing mental health conditions.

2.3 Deep learning (DL) techniques for mental health

Deep learning techniques are a subset of machine learning methods that use artificial neural networks with multiple layers to model complex structures in big data (Doan et al., 2019). Deep learning (DL) mostly achieves higher accuracy than traditional machine learning (ML), specifically for complex, multidimensional data such as images, text, and social media content (Chikersal et al., 2020; Hazra-Ganju et al., 2023; Kuhail et al., 2024). Unlike ML, which relies mainly on manual feature transformation, DL automatically extracts features from raw data, making it highly effective for large datasets (Ding et al., 2025). Traditional ML also performs well with smaller datasets and requires less computational power, making it more practical with limited resources. According to Razzaq and Shah (2025), DL excels at abstraction with large datasets but struggles with causal inference, unlike ML approaches that explicitly model causal relationships (Doan et al., 2019). For mental health prediction, a hybrid approach combining the accuracy of DL with the interpretability of ML could offer an effective solution (Razzaq & Shah, 2025).

There are several DL techniques for processing data, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). As explained by Alzubaidi et al. (2021) for Grid data processing, CNNs are commonly used in computer vision and text analysis to capture word patterns for sentiment classification. However, CNN in NLP, for example, in sentiment analysis or emotion detection in social media data (Nijhawan et al., 2022; Saranya & Kavitha, 2022). Another recent and innovative technique is transformers, widely used in NLP. It employs an attention mechanism to focus on relevant parts of the input, making it particularly effective for long-term dependencies such as machine translation and text summarisation (Jayasudha, 2025). Vaswani et al. (2017) applied these DL models to analyse social media data to predict mental health conditions based on user content.

Chancellor et al. (2019) stated that these methods are effective at predicting mental health risks and provide innovative approaches to detect these symptoms early through emotion and sentiment analysis. NLP and DL techniques and procedures have been successfully employed in social media posts like Twitter and Reddit (Dinu & Moldovan, 2021; Du et al., 2018; López-Ubeda et al., 2019), in psychological data (Obeid et al., 2019), and in voice data (Zainal et al., 2024) for sentiment analysis and emotion detection to predict mental health conditions. Some researchers used a multimodal approach, incorporating text, speech, images, and physiological signs to improve predictions (Malhotra & Jindal, 2020). Despite this advancement, key challenges remain, including data biases, privacy issues, model interpretability, and ethical concerns when dealing with sensitive mental health data (Nijhawan et al., 2022; Saranya & Kavitha, 2022; Yoon, 2024).

2.4 Causal Inference in Social Networks for Mental Health

Causal inference techniques are important for interpreting the social network's impact on mental health (Emmert-Streib & Dehmer, 2021). Ohlsson and Kendler (2019) have applied causal inference methods to analyse the patterns and trends of social interactions and mental health. Nowadays, the key sources of data for mental health research are Twitter, Reddit, and Instagram because of their large and diverse user bases, high levels of engagement, and rich, real-time content generation (Harrigian et al., 2021). These social media platforms provide researchers with discreet options to analyse social interactions, lexical patterns, and user behaviour, making them important for early mental health detection and sentiment analysis (De Choudhury et al., 2013).

Several causal models and methods are applied in mental health research using social media data; a few have been discussed here by other researchers. Temporal causal models have been used by Treur (2017) and Jabeen et al. (2020) to trace social interaction to study the impact of social influence on mental health. Meanwhile, the Bayesian network model integrates inferential models to explain mental trends. These models attempt to uncover underlying causal relationships between user behaviour, emotional states, and social interactions (Emmert-Streib & Dehmer, 2021). Unlike purely predictive models, which focus on trend analysis, causal Bayesian networks provide an explanatory framework for identifying risk factors and intervention strategies (Bernasconi, 2025). Moreover, recent advancements in decision-making trial and evaluation laboratory (DEMATEL) methods have been integrated with deep learning and text mining to analyse large-scale social media discussions on mental health (Zhao et al., 2021). These models quantify the influence of different social factors, such as stressors, peer support, and online therapy, on mental health, offering a more systematic approach to causal analysis (Zhao et al., 2021).

The challenges in the previous studies include data quality, model limitations, and ethical issues. Treur (2017) faced issues with model generalizability and validity, as simulated datasets do not reflect real-world scenarios. Emmert-Streib and Dehmer (2021) struggled with data bias, its limitations, and interoperability, making it hard to extract insights from social media data. Zhao et al. (2021) encountered complexity in feature selection, privacy issues, and linguistic variations that impact the accuracy and ethical considerations of mental health predictions based on social media data. The use of social media data for mental health research raises privacy concerns and ethical implications, as it involves the analysis of personal and sensitive information (Afrihyia et al., 2025). One of the main concerns is informed consent directly from the user, as social media allows researchers to collect publicly available data. Another critical issue is data privacy, as it involves personal information, which can lead to privacy violations if misused. Chancellor et al. (2019) explained the use of proper guidelines and ethical frameworks to handle the data, maintaining user confidentiality and prioritising individuals' mental health in research.

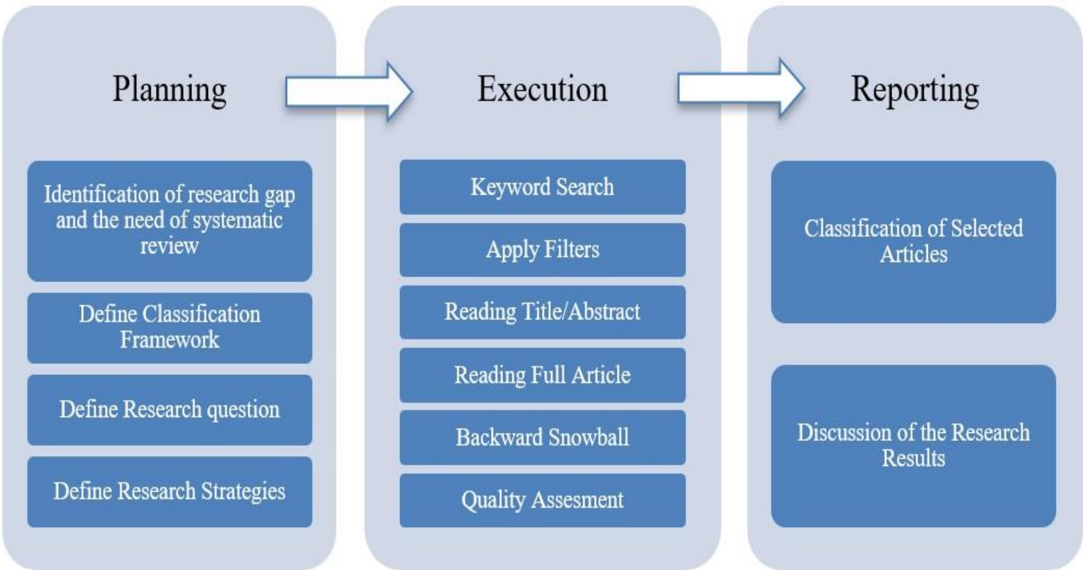
3.0 Methods

A systematic literature review (SLR) is a method for identifying, synthesising, and explaining all research relevant to the defined research scope, keyword, or related topics. It is also a methodology that includes the systematic collection, organisation, and assessment of previous studies and literature in the specific research field (Dabić et al., 2020; Paul et al., 2021). SLR has been widely used across several fields, including IT and mental health, as it provides a comprehensive overview of existing research and advances interventions and techniques in this complex field. Professionals and researchers in these industries rely on systematic reviews to stay informed about new developments, and, accordingly, Moher et al. (2009) noted that they serve as a foundation for technology guidelines. This study follows the structured approach outlined by Watson (2015), which comprises 3 phases for conducting an SLR: planning, execution, and reporting.

SLR are one of the best ways to ensure evidence-based practice in IT. According to Boell and Cecez-Kecmanovic (2015), the following structured literature search process makes SLR a highly effective research approach. However, Watson (2015) emphasised efficiency and noted that the real value of SLR lies in synthesising previous research, helping others develop deeper insights into key concepts and their connections.

Figure 1

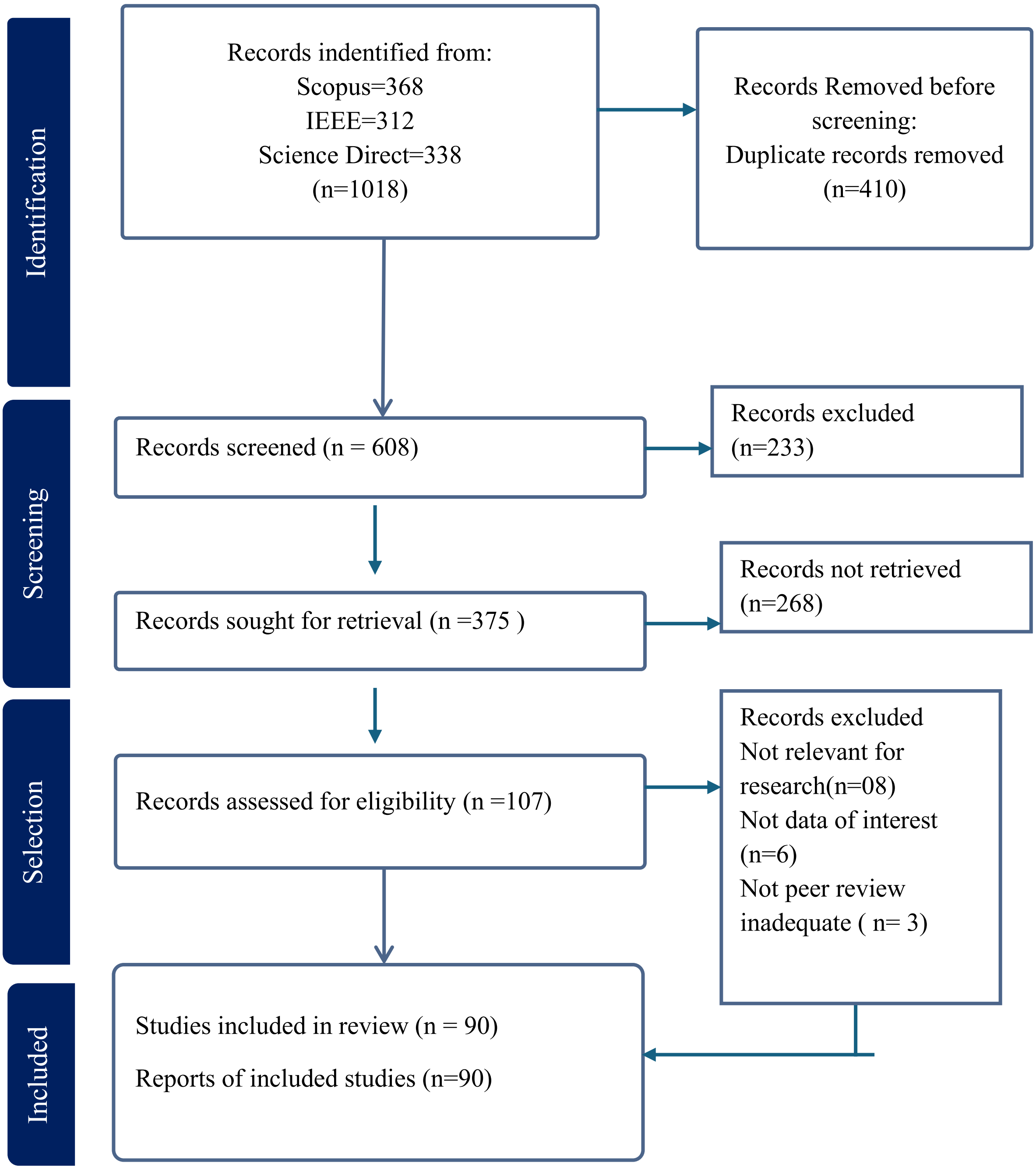
Systematic Literature Review Phases Adapted from Ali et al. (2023)



This study follows the systematic literature search phases defined by Ali et al. (2018, 2020, 2021), including three three-phase approaches outlined by Watson (2015). This SLR follows a three-phase approach, as shown in Figure 1. The planning phase includes identifying the requirements for the review, defining research objectives, and creating a classification framework. The execution phase involved a detailed literature search using keywords, filters, and full-text reviews. In the end, the reporting phase concludes by categorising, analysing, and discussing key points of the selected articles. The stages of the systematic literature review process, as outlined in the PRISMA framework, are illustrated in Figure 2.

Figure 2

Stages of Systematic Literature Review based on the PRISMA framework



3.1 Planning Phase

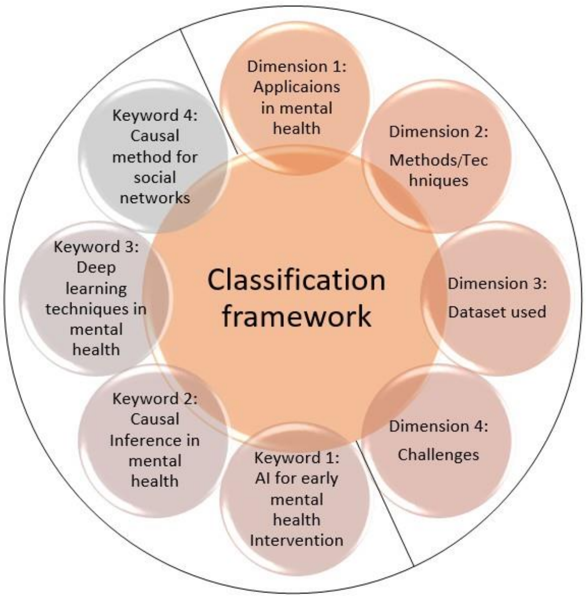
Initially, the planning phase determines whether a systematic review is needed. As mentioned earlier, key research has established how causal inference and deep learning contribute to early mental health intervention and detection, emphasising the methods/techniques, applications, datasets, and challenges. However, according to recent searches, no SLR has been conducted to support these findings and provide a detailed analysis of research and practical work in this field. The second step includes structuring a review protocol, which develops an infrastructure for understanding both the theoretical and practical aspects of this area (Ali et al., 2023).

3.1.1 Defining Classification Framework

This classification method is used in computing and healthcare research by Ali et al. (2018, 2020, 2023), explaining its applicability among diverse research areas. In this research, the proposed classification framework has been improved by adding additional dimensions, including the datasets used and the challenges they faced in the mental health sector. The framework is divided into four dimensions: application in mental health, methods/techniques, the dataset used, and challenges. Each of these dimensions was categorised from the keywords based on the results of the article selection for review, as shown in Figure 3.

Figure 3

Classification Framework



Source: Author

In each keyword, relevant dimensions were identified:

Application in mental health, where the studies have been applied within the mental health sector. These include applications in AI for early mental health intervention, particularly for detecting depression from social media posts (Myee et al., 2024), and for wearable-based mental health monitoring (Sandulescu et al., 2024). They also include applications in Causal inference in mental health, which consists of extracting health-related causal information from Twitter messages (Doan et al., 2019), counterfactual reasoning in mental health (Marchezini et al., 2022), and the moderating role of social interactions before and after COVID-19 (Ghaffari et al., 2024). The other keyword includes applications in deep learning techniques for mental health that consist of depression detection (Bokolo & Liu, 2022), detecting stress levels (Nijhawan et al., 2022), assessing the severity of depression (Muñoz & Iglesias, 2023), and identifying and classifying multiple mental illnesses (Dinu & Moldovan, 2021) from social media posts (Twitter, Reddit). The last keyword of this dimension is causal methods in social media for mental health, which includes social interactions to examine their influence on mental health (Treur, 2017), analyses of how data-driven social network science can predict mental health (Emmert-Streib & Dehmer, 2021), and social media-based mental health analysis and prediction using feature extraction techniques (Zhao et al., 2021).

Methods and techniques used in the studies according to every keyword. This includes methods in AI for early mental health intervention, like Decision Trees for mental health risk prediction (Oakey-Neate et al., 2020; Oei et al., 2023), sentiment analysis, emotion recognition to detect depression (Chikersal et al., 2020; Myee et al., 2024), and IoMT and emotion detection models (Sandulescu et al., 2024). Also, methods in Causal Inference for mental health include Difference-in-Differences (DiD) and two-way Fixed Effects to investigate the behaviour change before and after COVID-19 (Ghaffari et al., 2024) and NLP to extract health-related causal information (Doan et al., 2019). These include methods in deep learning techniques that consist of BERT, RoBERTa, and CNN for sentiment analysis to analyse mental health and depression (Bokolo & Liu, 2022; Dinu & Moldovan, 2021; Du et al., 2018; Malhotra & Jindal, 2020; Nijhawan et al., 2022; Zainal et al., 2024; Zanwar et al., 2023). Also, methods in causal methods for social media, Temporal Causal Network Models to study the social interaction impact on mental health (Treur, 2017), Causal Bayesian Networks to predict and explain mental health outcomes (Emmert-Streib & Dehmer, 2021), text Mining and topic modelling for feature extraction (Zhao et al., 2021).

Datasets refer to the data utilised in conducting the studies. These include datasets in AI for mental health intervention from social media posts (Twitter and Reddit) (Myee et al., 2024), textual data from psychologists' notes (Oakey-Neate et al., 2020; Uddin et al., 2022), Clinical data (Oei et al., 2023), Physiological and behavioural data (Sandulescu et al., 2024), and data from AI-driven systems (Chikersal et al., 2020; Hazra-Ganju et al., 2023; Kuhail et al., 2024; Yoon, 2024). They also include Causal inference for mental health using synthetic datasets (Treur, 2017) and data from social media platforms and online forums (Emmert-Streib & Dehmer, 2021; Zhao et al., 2021). It includes deep learning techniques dataset mainly from social media posts (Twitter and Reddit text data) (Almars, 2022; Alturayeif & Luqman, 2021; Bashar et al., 2022; Dinu & Moldovan, 2021; Du et al., 2018; Gongane et al., 2022; Hasan et al., 2023; López-Ubedá et al., 2019; Malhotra and Jindal, 2020; Martín & Camacho, 2022; Muñoz & Iglesias, 2023; Nijhawan et al., 2022), other types of healthcare data and clinical records (Kumari et al., 2024; Aleem et al., 2022; Martín & Camacho, 2022; Ozkanca et al., 2019; Obeid et al., 2019; Saranya & Kavitha, 2022). It also includes causal methods for social media, and the datasets used are synthetic (Treur, 2017) and social media posts (Emmert-Streib & Dehmer, 2021; Zhao et al., 2021).

Challenges are the issues the previous studies faced when using different methods, depending on the keywords mentioned for detecting and predicting early mental health. These include challenges in AI-based solutions for early mental health intervention, such as privacy and ethical considerations in data acquisition (Chikersal et al., 2020; Myee et al., 2024; Oei et al., 2023; Yoon, 2024). Furthermore, causal inference challenges in mental health include confounding by causality, data availability and reliability, and contextual understanding (Doan et al., 2019; Marchezini et al., 2022; Ghaffari et al., 2024). Moreover, challenges in deep learning for mental health consist of data bias, lack of standard, complexity in fusing multiple modalities, feature selection complexity and computational challenges (Al-Laith & Alenezi, 2021; Almars, 2022; Gongane et al., 2022; Kumari et al., 2024; Lin et al., 2020; Omarov & Zhumanov, 2023; Petrolini et al., 2022; Zhang et al., 2021; Zhao et al., 2021). Finally, the last keyword of this dimension refers to the challenges faced in causal methods in social media for mental health, which consists of model generalizability, data validation and interpretability, feature selection complexity, and ethical concerns for using personal data (Emmert-Streib & Dehmer, 2021; Treur, 2017; Zhao et al., 2021).

3.1.2 Real-World Case Studies

Based on Figure 2, there are several real-world applications of deep learning and causal inference for mental health prediction using social media. AI-driven mental health assessment primarily uses deep learning (DL) and causal inference techniques to analyse social media data, offering new insights into mental health trends and early intervention strategies (Jiao et al., 2024).

Several platforms now utilise social media data for mental health monitoring and early mental health detection (Chikersal et al., 2020; Hazra-Ganju et al., 2023; Kuhail et al., 2024). Facebook's AI-powered suicide prevention tool analyses user posts to identify distress signals and connects at-risk individuals with crisis support. Woebot and Wysa, AI-driven chatbots, employ sentiment analysis and conversational deep learning models to detect signs of anxiety and depression. Additionally, research-based models have explored social network dynamics, assessing how interaction patterns and network size correlate with mental health outcomes (Pearl & Mackenzie, 2018). Deep learning techniques such as transformers (BERT, GPT), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) are widely used for sentiment analysis, topic modelling, and emotion recognition in social media data (Kawam et al., 2025).

However, while DL excels in predictive accuracy, it struggles with interpretability and causal understanding (Pearl & Mackenzie, 2018). To address this, structural causal models (SCMs) and counterfactual reasoning are being integrated with deep learning to move beyond correlation and establish causal relationships between social network interactions and mental health states (Zhao et al., 2022). For example, previous research has applied causal methods to conclude whether increased social isolation directly contributes to depressive symptoms or if external spurious influencers play a role. Publicly available datasets are crucial for training deep learning models for mental health prediction (Jiao et al., 2024). Reddit Mental Health Datasets (e.g., r/depression, r/Anxiety, r/mental health), Twitter sentiment datasets, and DAIC-WOZ (a depression assessment dialogue dataset) are commonly used to analyse linguistic patterns and social engagement indicators (Myee et al., 2024). However, dataset biases, imbalanced representations of mental health conditions, and the dynamic nature of social media data present significant challenges (Kuhail et al., 2024). Integrating causal inference with deep learning provides a promising path forward for improving mental health prediction using social media data. While existing AI models demonstrate strong predictive capabilities, addressing challenges in comprehensibility, causal understanding, and ethical considerations is important (Hazra-Ganju et al., 2023).

The proposed ensemble deep learning model, incorporating social network size and causal inference, could offer a more transparent and actionable approach to mental health assessment, ultimately contributing to more effective intervention strategies.

3.1.3 Research Question

Defining the research question is the third step in the planning phase and is considered the most important in any systematic review (Paul et al., 2021). A systematic review achieves its goals by answering the research questions (Paul & Benito, 2018). The research question for this systematic review study is as follows:

What causal inference and deep learning methodologies are applied in social networks for early mental health detection?

Defining strategies for selecting articles: The fourth step in the planning phase involves defining these strategies. These strategies aim to identify primary studies that provide direct evidence related to the research question. To ensure a comprehensive and well-rounded selection, this study employed an integrated search strategy that combined automated database searches with a manual review process.

A broad automated search was conducted across multiple academic databases, including IEEE, ScienceDirect, and Scopus, as shown in the Figure. According to Rosado-Serrano et al. (2018), using an extensive search strategy helps incorporate the most relevant and credible sources. Filtering tools within each database were applied to refine the search results and minimise duplication.

Table 1

Review Selection Criteria

Criteria Inclusion	Inclusion	Exclusion
Type of Publication	Scholarly articles	Reports, and any other sources
Peer-reviews	Peer-reviewed	Non-peer reviewed
Publication year	Articles published from "2015 to 2024"	Articles that were published before "2015"
Language	English	Any language other than English

Table 1 outlines the inclusion and exclusion criteria for the systematic review. The criteria used to include the studies of interest were publication status, peer review, publication date, and language. Regarding publication, scholarly journal articles were included, while reports were excluded. Regarding peer review, only peer-reviewed publications were included, while non-peer-reviewed publications were excluded. With respect to date of publication, articles published between "2015" and "2024" were included, while those published before "2015" were excluded. Finally, with respect to language, English-language articles were included while all other articles were excluded.

In addition to the automated search, a manual screening process was conducted. This involved reviewing article titles and abstracts before conducting a full-text evaluation to filter out irrelevant studies. This manual approach ensured that only the most relevant and high-quality research articles were included in the review (Ali et al., 2023).

Table 1 outlines the eligibility criteria used for selecting articles. Establishing clear selection criteria ensures that the review is accurate, unbiased, and meaningful. When applied systematically, these criteria help reduce errors and enhance the reliability of the study's findings.

3.2 Execution Phase

During the execution phase, the strategies outlined in the planning phase were implemented to select the most relevant articles for this study. The process involved several key steps, as described further in this section.

As described by Paul et al. (2021), identifying search terms was an iterative process, starting with well-established keywords from recognised studies in the field. The process continued until no new relevant articles were found using the same search principles. The chosen academic databases offer advanced search functionalities, enabling the combination of multiple keywords to refine search results.

Keyword Search: For this study, the following keywords were identified based on the research question: "AI for early mental health intervention," "causal inference in mental health," "deep learning techniques for mental health," and "causal methods in social networks for mental health." The applied filters included research area (IT and mental health), publication year "2015-2024", and document type (journal articles and conference papers).

Reading Articles: Once search results were generated, each article was manually screened based on its title and abstract to assess relevance (Pucher et al., 2013). Articles that met the criteria underwent a thorough content analysis to extract relevant information.

Quality Assessment: Finally, a quality assessment was conducted to ensure that only high-value, credible studies were included (Hu & Bai, 2014). This ensured the accuracy, reliability, and relevance of the selected articles.

3.3 Summarising Phase

Table 2 presents the final selection of articles included in this systematic review. The initial keyword-based search identified 1,018 unique articles across multiple databases Figure 2. After applying filters, such as research domain, publication year, and document type, the number of articles was reduced to 608. A manual review was then conducted to eliminate studies that were irrelevant to the research topic.

During this process, both empirical and conceptual articles that directly addressed the study's objectives were considered. As a result, 233 articles were removed, leaving 375 articles for further evaluation. Next, a full-text review was conducted, focusing on key selection criteria, including research objectives, research questions, data collection methods, applied methodologies, analytical techniques, and result presentation. Following this review, 268 articles were excluded due to lack of relevance, reducing the total to 107 studies. Finally, a quality assessment was conducted, resulting in the removal of 17 articles that did not meet the required standards. This led to a final selection of 90 high-quality articles for analysis, as shown in Figure 2. Figure 4 shows the graphical representation of Table 2, i.e., the review search results. Figure 5 shows the final selected articles after quality assessment**.**

Table 2

Review Search Results

	Automated Search Method		**Manual Search Method**		**Final Results**
Databases	**1st Strategy Keyword search**	**2nd Strategy Apply Filters**	**3rd Strategy Reading Title And Abstract**	**4th Strategy: Reading the Full Article**	**5th Strategy Quality Assessment**
Scopus	368	270	153	47	48
IEEE	312	175	117	45	26
Science Direct	338	163	105	15	16
Total	1018	608	375	107	90

Figure 4

Graphical Representation of Review Search Results

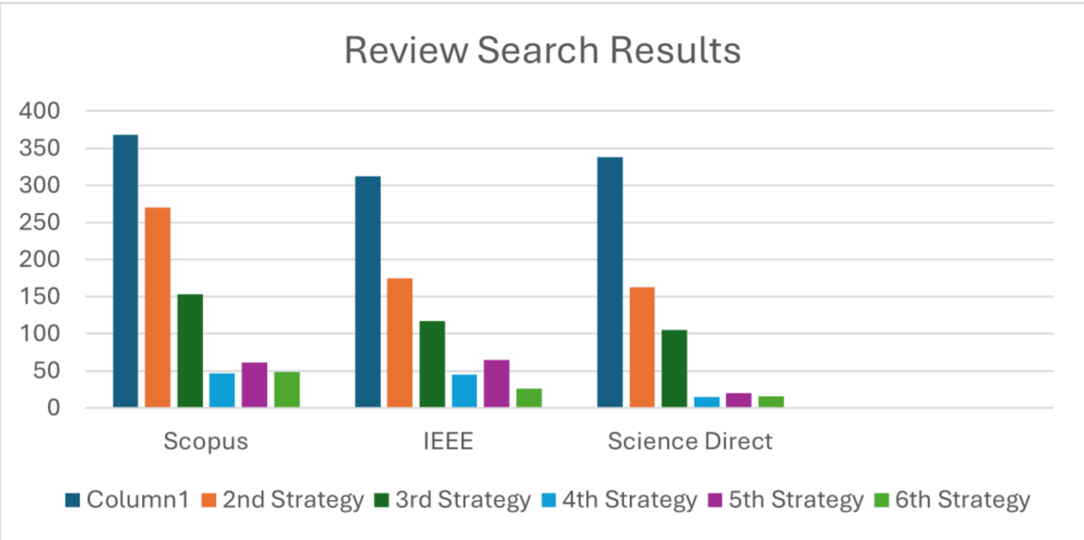
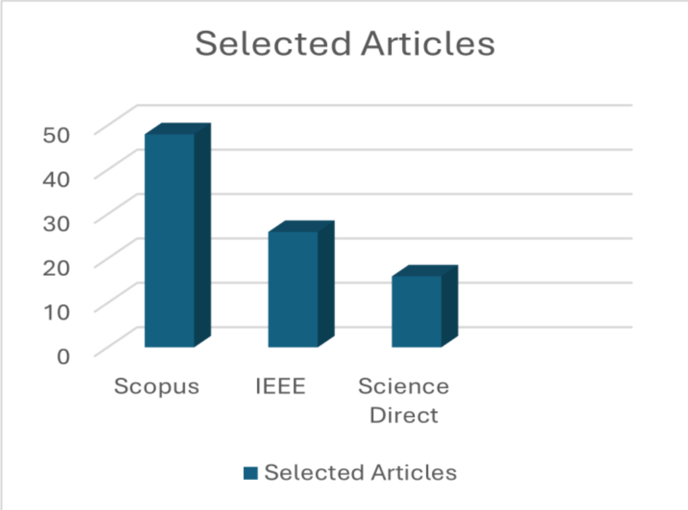


Figure 5

Selected Articles from Scopus, IEEE and Science



3.4 Reporting Phase

3.4.1 Research Classification Framework

The findings of a systematic review of causal Inference and deep learning techniques in mental health-related articles were presented and examined. The classification framework was applied by considering the following four dimensions: application in mental health, methods and techniques, dataset used, and challenges. Table 3 presents the findings for each dimension across all keywords.

Table 3

Research Classification Framework

Keyword	Dimension1: Applications in Mental Health	Dimension2: Methods/ Techniques	Dimension3: Datasets used	Dimension4: Challenges
Keyword 1: AI for early mental health Intervention	Depression detection from social media.	Transformer-based models.	Social media datasets.	Data quality and noise.
Keyword 2: Causal inference in mental health	Stress and anxiety detection suicide risk detection.	Causal inference techniques.	Clinical data.	Model interpretability.
Keyword 3: Deep learning techniques for mental health	Emotion and sentiment detection.	Multimodal deep learning models.	Speech and audio datasets multimodal data.	Privacy and ethical concerns.
Keyword 4: Causal methods in social networks for mental health	Effectiveness of mental health interventions.	Meta-heuristic optimization hybrid models.	IoT and wearable sensor data.	Dataset bias and generalizability feature fusion complexity.

For more details, see Tables 3, 4, 5, 6, 7 and 8. Table 3 illustrates the classification framework for AI in early mental health Intervention and comprises specific categories for this dimension, such as methods/techniques, application in mental health, datasets used, and challenges. Table 5 presents a classification framework for causal Inference in mental health, with specific categories for this dimension, including challenges, methods, applications in mental health, and the dataset used. Table 6 presents a classification framework for deep learning techniques in mental health, including specific groups such as multimedia processing, textual data processing, and others. Table 7 presents the classification framework for causal methods in social networks related to mental health, covering specific categories within this dimension.

4.0 Results and Discussions

4.1 Dimension 1: Application in Mental Health

4.1.1 Applications in AI for early mental health intervention

AI is increasingly applied to early mental health interventions, helping detect depression, stress, and other mental health conditions at an early stage through methods such as clinical decision support, wearable monitoring, sentiment analysis, and AI-assisted therapy (Ghaffari et al., 2024). AI-based approaches, such as decision trees and XGBoost models, have been used for post-stroke mental health risk prediction, enabling more targeted patient interventions (Oei et al., 2023). NLP models, including BERT and LSTMs, analyse social media posts and online therapy interactions to identify signs of depression and mental health distress (Myee et al., 2024; Uddin et al., 2022). AI-driven chatbots and digital therapeutics are also being developed for adolescent mental health interventions, using reinforcement learning to tailor responses based on user interactions (Yoon, 2024). According to Oakey-Neate et al. (2020), clinical decision support systems (CDSS) assist in personalised mental health treatments for antipsychotic users by using patient data for better intervention strategies. As discussed by Sandulescu et al. (2024) and Hazra-Ganju et al. (2023), wearable technology and digital platforms enable continuous mental health monitoring through emotion detection models and IoT tracking (Refer to Table 4).

4.1.2 Applications in causal inference in mental health

Causal inference techniques are increasingly used to understand the relationships between mental health disorders and various influencing factors, including social interactions, treatment efficacy, and behavioural responses (Ohlsson & Kendler, 2019). Difference-in-Differences (DiD) and Two-Way Fixed Effects (TWFE) methods have been implemented to analyse the effect of social interactions on mental health behaviours, such as changes in online music listening patterns during COVID-19 (Ghaffari et al., 2024). Marchezini et al. (2022) used counterfactual inference techniques in mental health care to assess intervention results for patients by latent variable modelling. Moreover, Doan et al. (2019) used NLP-based causal inference models to extract causal relationships from social media platforms. As explained by Pearl et al. (2016), natural experiments remain important for understanding causal models of mental health disorders through empirical evidence (Refer to Table 5).

4.1.3 Applications in deep learning techniques for mental health

Deep learning (DL) techniques are being used extensively in advanced mental health research to detect depression, analyse sentiment, measure stress, and detect cyberbullying (Razzaq & Shah, 2025). These applications span text-based social media analysis, speech-based depression detection, multimodal behavioural assessments, and clinical decision support systems (Yoon, 2024). Transformer-based models have been employed for analysing social media posts to detect early signs of depression, anxiety, and stress (Ding et al., 2025). In contrast, speech processing models have been used to assess emotional distress from voice recordings (Almars, 2022; Dinu & Moldovan, 2021; Zhang et al., 2021). Li et al. (2023) and Z. Zhang et al. (2024) have demonstrated that multimodal approaches analyse multiple data sources, including text, speech, video, and physiological signals, for risk prediction and prevention in patients with depression. Similarly, AI-based precision and prediction analysis in clinical mental health is also very important for prevention strategies (Martín & Camacho, 2022; Squires et al., 2023) (Refer to Tables 6 and 7).

4.1.4 Applications in causal methods in social networks for mental health

Social networks are the primary data source nowadays, mainly for causal methods to understand the relationship between social interactions and mental health (Yang & Li, 2025). These methods analyse how social dynamics, behaviour, and external influences affect mental health. Treur (2017) applied temporal causal network methods to analyse social interactions and their long-term impact on mental health. Emmert-Streib and Dehmer (2021) used causal Bayesian networks and predictive models to explain mental health outcomes through users' online behaviour on social media and online forums. Moreover, Zhao et al. (2021) performed mental health analysis through text mining and topic modelling to predict mental health conditions. These methods provide critical insights from social media data to predict mental health risks and identify coping strategies (Refer to Table 8).

4.2 Dimension 2: Techniques/Tools and Models

4.2.1 Techniques and methods in AI for early mental health intervention

Several AI techniques are used in mental health interventions, including machine learning, NLP, deep learning, reinforcement learning, and Internet of Medical Things (IoMT) technologies (Chikersal et al., 2020; Hazra-Ganju et al., 2023; Kuhail et al., 2024; Uddin et al., 2022). Decision trees, random forests, and XGBoost models play a critical role in clinical decision support, offering explainability in patient risk prediction and needs-based interventions (Oakey-Neate et al., 2020; Oei et al., 2023). Uddin et al. (2022) analysed large-scale text data using RNNs, LSTMs, and BERT to identify and predict depression symptoms. Chikersal et al. (2020) created an AI-assisted mental health platform using NLP conversation analysis to explain client interactions with the digital therapist.

Table 4

AI for Early Mental Health Intervention

Sources	Methods/ Techniques	Application in Mental Health	Dataset Used	Challenges
Oei et al. (2023)	Decision Trees, XGBoost, SHAP	Post-stroke mental health risk prediction	Clinical and demographic data from 1,780 post-stroke patients	Explainability is crucial for Decision Trees and XGBoost. Imbalanced data and ethical concerns.
Myee et al. (2024)	NLP, Sentiment Analysis, Emotion Recognition	Depression detection using social media data	Online Social media platforms like Twitter and Reddit	Cross-platform generalisation, and ethical concerns.
Uddin et al. (2022)	RNNs, LSTMs, BERT NLP models	Depressive symptoms prediction in text data	Large textual dataset of mental health-related content	Noise and biases. High computational resources. Generalisation across populations and languages.
Chikersal et al. (2020)	Random Forest, NLP for conversation analysis	Online mental health interventions	Digital therapy interaction data	Understanding client interactions. Accessing high-quality conversation datasets. Privacy and confidentiality.
Kuhail et al. (2024)	GPT-based chatbots, psychological scale assessment	AI vs. human therapy comparison	Data from empirical tests comparing AI-assisted therapy to human-led therapy	Human therapists over AI. while biases in psychological scales and small sample sizes in AI-human therapy Studies limit generalization
Oakey-Neate et al. (2020)	Clinical Decision Support Systems, Decision Trees	Needs-based interventions for antipsychotic users	Patient data on antipsychotic medications	Identifying patient needs. Require strict privacy controls. Clinician trust.

Sources	Methods/ Techniques	Application in Mental Health	Dataset Used	Challenges
Sandulescu et al. (2024)	IoMT, Emotion detection models	Wearable-based mental health monitoring	Physiological and behavioural data from wearable and Iot-based sensors	Sensor accuracy and battery life. Misinterpretation of signals and inconsistent device usage.
Hazra-Ganju et al. (2023)	Omni-channel AI (SMS, Chatbots), Supervised ML	Digital mental 20 health intervention in low-resource areas	Data from AI-driven digital health platforms	Stable internet access. Diverse regions, culturally relevant, and language barrier.
Yoon (2024)	Reinforcement Learning, Digital Therapeutics	AI for adolescent mental health & crisis management,	Data from AI-driven systems providing mental health	Large datasets, strict safeguards for adolescents. Immediate crisis responses.

Similarly, Yoon (2024) developed a personalised AI-driven digital therapeutic solution for youth mental health using reinforcement learning techniques. Sandulescu et al. (2024) integrated emotion-detection models with wearable sensors using IoMT technologies to monitor signals and behaviour indicators. Omni-channel AI systems, including chatbots and SMS-based mental health interventions, provide low-resource mental health support, leveraging supervised machine learning for adaptive responses (Hazra-Ganju et al., 2023) (Refer to Table 4).

Table 5

Causal Inference in Mental Health

Study	Methods/ Techniques	Application	Dataset Used	Challenges
Ghaffari et al. (2024)	Difference-in-Differences (DiD), Two-Way Fixed Effects (TWFE)	online music listening behaviours during COVID-19 and the role of social interactions	Online music listening data from 37,328 Last.fm users across 45 countries	Controlling confounding factors. Potential biases.
Marchezini et al. (2022)	Counterfactual Inference with Latent Variables	Modelled counterfactual reasoning in mental health care scenarios involving latent variables	Clinical data from mental health care settings	Complex confounders. Generalizability due to clinical data limitations. Ethical concerns.
Doan et al. (2019)	Natural Language Processing (NLP)	Extracted health-related causal information from Twitter messages	Twitter data containing health-related messages	noisy and short-text data. Unstructured text. Linguistic ambiguity
Pearl et al. (2016)	Randomised Clinical Trials (RCTs), Natural Experiments, Instrumental Variable Analysis	Understanding causal relationships in psychiatric disorders,	Large-scale psychiatric epidemiology datasets,	Confounding variables. Ethical concerns in experimental designs

Study	Methods/ Techniques	Application	Dataset Used	Challenges
Gao et al. (2019)	Causal Discovery, Structural Causal Models (SCMs)	Identifying causal factors in mental health using big data	Multi-dimensional big data sources (electronic health records, social media, wearable data)	High dimensionality, difficulty in establishing causality beyond correlations
Li et al. (2023)	NLP-based Causal Inference, Sentiment Analysis, Counterfactual Reasoning	Using NLP to infer causal relationships in social media interactions related to mental health	Social media datasets (Twitter, Reddit)	Data noise. Linguistic ambiguity. Need for explainable AI
Yuan et al. (2023)	Difference-in-Differences (DiD), Regression Discontinuity Design (RDD)	Analysing the impact of mental health coping 21 stories on social media users	Twitter dataset (tweets discussing mental health coping strategies)	Selection bias. Difficulty in establishing long-term effects

Table 6

Deep Learning techniques for mental health

Topics	Studies (Source)	Applications	Methods/ Techniques	Dataset Used	Challenges Faced
1. Text Based Depression & Sentiment Analysis	Y. Zhang et al. (2021), Almars (2022), Alturayeif & Luqman (2021), Kumari et al. (2024), Nijhawan et al. (2022), Muñoz & Iglesias (2023), López-Úbeda et al. (2019), Dinu & Moldovan (2021)	Monitoring and classification of depression trends using social media text	Transformer, Bi-LSTM, BERT, NLP sentiment analysis, RoBERTa, Feature-Based NLP, Random Forest, SVM	Twitter (2575 users), Arabic tweets, 20,000 labeled text samples, 92 research articles, Reddit dataset, Mental health-forum data, Spanish Twitter dataset	biases, Limited datasets, interpretability, Unstructured text processing, Computa-tional efficiency
2. SpeechBased Depression Detection	Zhao et al. (2021), Lin et al. (2020), Yin et al. (2023), Zainal et al. (2024)	Depression detection from speech signals	Multi LSTM, BiLSTM + 1D CNN fusion, Transformer and Parallel CNN, MFCCs, CNN for classification	DAIC-WOZ and MODMA speech datasets, DAIC-WOZ and AViD-Corpus, Unscripted	Capturing depression indicators, High computational complexity, Overfitting risks, Privacy concerns, Scarce datasets

Topics	Studies (Source)	Applications	Methods/ Techniques	Dataset Used	Challenges Faced
				speech datasets	
3. Emotion & Cyberbullying Detection in social media	Omarov & Zhumanov (2023), Hasan et al. (2023), Jadon & Kumar (2024), Geethanjali & Valarmathi (2024)	Emotion analysis & sentiment detection for mental health, Cyberbullying detection	Bi-LSTM, CNN, RNN, LSTM, Bi-LSTM Hybrid Models, Word embeddings (Word2Vec, GloVe, BERT, RoBERTa), CNNs	Kaggle dataset, Twitter, Instagram, YouTube, Social media posts (Twitter, Reddit)	Dependencies, Text distortion, Real-time challenges, Selecting optimal word embeddings, Model interpretability, Dataset biases
4. Multimodal Depression Detec-	Z. Zhang et al. (2024), Geethanjali & Valarmathi (2024), Junyi (2024)	Depression detection using multimodal cues (speech,	Multimodal Hierarchical Attention Model, Audio-Video,	Weibo dataset, Multimodal datasets, social media	Limited datasets, Privacy concerns, Feature
5. Machine Learning & Metaheuristic Approaches in Mental Health	Aleem et al. (2022), Saranya & Kavitha (2022), Jawad et al. (2023)	Depression diagnosis and prediction	SVM, Decision Trees, Cuckoo, CNN, ResNet, VGG, Chicken Swarm Intelligence, Self-Stabilizing Deep Neural Network (CSI-ISDNN)	Clinical records, social media datasets, Clinical datasets and mental health records	Data quality issues, Lack of standard, Convergence issues, Computational complexity, Instability in deep learning training
6. NLP & Sensitive Data Detection in Mental Health	Petrolini et al. (2022), Bokolo & Liu (2022), Du et al. (2018)	Automated detection of sensitive mental health related data, Suicide risk detection, Extracting psychiatric stressors	BERT, LSTM, BiTCN, CNN, RNN (NER), Transfer Learning, Logistic Regression, Random Forest, Na¨ıve Bayes, RoBERTa	Custom-built dataset of sensitive documents, Twitter dataset, Reddit dataset	Sensitive information, Privacy and ethical concerns, Model bias, High false-positive rates, Data noise, Imbalanced datasets
7. AI & Deep Learning Applications in Healthcare & Mental Health	Martín & Camacho (2022), Squires et al. (2023)	Overview of AI applications in health and mental health	CNNs, Neural Networks, Transfer Learning, Deep Learning, Precision Psychiatry	Clinical, Healthcare and social media data	Computational efficiency, High training costs, Lack of empirical validation, Explainability of AI models
8. Facial, Vocal, and Behavioural Mental Health Detection	Patil & Khedkar (2023), Rashed et al. (2024), Ezerceli & Dehkharghani (2024), Alshahrani et al.	Automated mental health detection using facial Images, voice, and behavioural analysis	CNN, VGG16, ResNet50, CNN-LSTM, Efficient Net, Bayesian Optimization, BERT, YOLO,	"FER-2013" dataset, Facial expression dataset, DB1 & DB2 image	Data variability, Ethical concerns, Privacy risks, Annotation consistency, Feature selection, Model ex-

Topics	Studies (Source)	Applications	Methods/ Techniques	Dataset Used	Challenges Faced
	(2024), Alshanketi (2024), W. Yang (2023), Malhotra & Jindal (2020)		MGADHF, VGG-16, Faster R-CNN, Word2Vec, Task & Feature	datasets, CEASEv2.0, SWMH datasets, 500 annotated children's drawings, Twitter	plain ability, Real-time deployment,

Table 8

Causal Methods in Social Networks for Mental Health

Study	Application	Methods/ Techniques	Dataset Used	Challenges
Treur (2017)	Simulated evolving social interactions to study their impact on mental health	Temporal Causal Network Models	Synthetic datasets and model simulations	No real-world social interactions Requires high computational power Lack of real-world validation for accuracy
Emmert-Streib & Dehmer (2021)	Explores how computational social network science can predict and explain mental health outcomes	Causal Bayesian Networks, Inferential and Predictive Models	Data from various social media platforms and online forums	with strong assumptions Unstructured and biased social media data Dynamic networks limit their effectiveness.
Zhao et al. (2021)	Social media-based mental health analysis and prediction using feature extraction techniques	Text Mining, Topic Modelling (LDA), Random Forest, DEMATEL & DANP	Data collected from Twitter and other social media sources	Complex NLP techniques for data extraction Privacy concerns Cultural variations

4.2.2 Techniques and methods in causal inference in mental health

A variety of causal inference methods are applied to mental health research, ranging from statistical modelling to deep learning-driven causal discovery (Emmert-Streib & Dehmer, 2021). DiD and RDD are often used to predict the effects of social interactions and mental health behaviours, as described by Yuan et al. (2023), to analyse coping strategies and online behaviours. Causal models such as PSMs and SCMs are applied to large-scale data from social media, health records, and physiological signals to identify causal relationships between stressors and health interventions (Gao et al., 2019). Meanwhile, NLP-based causal inference models integrate sentiment analysis and counterfactual reasoning to analyse mental health-related conversations on platforms like Twitter and Reddit, helping detect causal signals in user discourse (Li et al., 2023) (Refer to Table 5).

4.2.3 Techniques and methods in deep learning for mental health

Several deep learning architectures and machine learning models have been adopted in mental health research. NLP-based models, such as BERT, RoBERTa, and Bi-LSTM, have been applied to social media text analysis to detect linguistic markers of mental distress (Dinu & Moldovan, 2021; Zhang et al., 2021). Also, Speech analysis models such as LSTM, BiLSTMCNN fusion, and MFCC classifiers have been applied by Zhao et al. (2021), Yin et al. (2023), and Zainal et al. (2024) to identify depression measures in audio data. Multimodal approaches, including MHA, AVTF-TBN, CNN-LSTM, and Bayesian optimisation, have been utilised by Li et al. (2023) and Zhang et al. (2024) to improve the precision of stress and depression detection. Search algorithms such as Cuckoo Search Optimisation and Chicken Swarm Intelligence have also been incorporated into mental health diagnostics to improve predictive modelling (Aleem et al., 2022; Jawad et al., 2023). Finally, transformer-based classifiers (BERT, GPT) and self-attention ensemble models have been deployed for suicidal ideation detection and psychiatric stressor identification from online discussions (Du et al., 2018; Petrolini et al., 2022) (Refer to Tables 6 and 7).

4.2.4 Techniques and methods in causal methods in social networks for mental health

Various methods and modalities are used in causal relationship modelling with social media data for mental health. Treur (2017) applied Temporal Causal models to derive the social learning mechanisms underlying mental health outcomes. Causal Bayesian Networks are applied, combining a probabilistic model approach with inferential methods to model the impact of specific social behaviours on mental health (Emmert-Streib & Dehmer, 2021), thereby helping identify risk factors and the causal structure of online interactions. Text mining, topic modelling (LDA), and Random Forest methods are also frequently used to mine for indicators of mental states from social media text, speech, and images, yielding insights into mental states from language (Zhao et al., 2021). Zhao et al. (2021) also applied DEMATEL and DANP to understand causal relationships in mental health and social interaction (see Table 8).

These computational models are, however, closely aligned with a body of communication research that establishes that mental health in digital environments is strongly shaped by interpersonal communication processes such as self-disclosure, emotional expression, exchange of social support, peer feedback, dealing with stigma, negotiation of stigma, and networked social influence (Zhao et al., 2021). From a communication perspective, causal modelling can be viewed as an important approach to understanding the impact of communicative behaviours and interaction patterns/dynamics on mental health issues and to developing some degree of predictability of risk for mental health (or resilience) over time (Patil & Khedkar, 2023). Not only does causal modelling allow for this approach, but multimodal modelling extends this potential even further for communication researchers/scholars regarding the relevance of communication processes to these models and methods (Yang, 2023). Multimodal models allow researchers to understand predictors of mental health at the convergence of multiple modes of meaning-making and emotion capture, through text, speech, and images, whereas unimodal models limit this understanding to a single data type (Rashed et al., 2024). Yet these models come with complex architectures to be configured, trade-offs in the time and resources required to train/prepare them, the need for synchronised data across multiple modalities, and the inability to derive insights in real time, unlike unimodal models that focus on one modality/signal (Malhotra & Jindal, 2020). Furthermore, they have specific biases regarding data, modality, labeling and fairness that can have detrimental implications for predicting mental health issues across diverse populations (Ezerceli & Dehkharghani, 2024). Thus, causal and multimodal AI-based methods must be firmly rooted in communication theory to develop methodologies for detecting mental health issues that are socially grounded, interpretable, and ethically responsible (Alshahrani et al., 2024; Alshanketi, 2024).

4.3 Dimension 3: Data Used

4.3.1 Data used in AI for early mental health intervention

The dataset used in AI for early mental health intervention varies from clinical records to social media and real-time sensor data. Different research required different data types; for clinical decision support models, medical datasets were utilised for mental risk assessment (Oakey-Neate et al., 2020; Oei et al., 2023). For NLP sentiment analysis and depression detection, social media data is more appropriate for analysing sentiment shifts and mental health indicators (Myee et al., 2024; Uddin et al., 2022). While AI-assisted therapy interventions use digital conversation data to analyse therapist-patient interactions and automate mental health responses (Chikersal et al., 2020). Wearable-based AI monitoring integrates physiological and behavioural data from IoT-based sensors, capturing real-time stress levels and emotional fluctuations (Sandulescu et al., 2024). Yoon (2024) used datasets on mental health support from an AI-driven digital platform for youth mental health. (Refer to Table 4)

4.3.2 Data used in causal inference in mental health

For casual inference, the most used datasets are social media and other online platforms, paired with real-time clinical and physiological data (Sandulescu et al., 2024). Studies on social interactions and mental health outcomes use large-scale behavioural datasets; for example, Ghaffari et al. (2024) used user records from an online music platform to analyse behaviour pre- and post-COVID-19. Mental health coping strategies on social media are assessed using Twitter datasets, in which researchers examine patterns of engagement with coping-related content (Yuan et al., 2023). Marchezini et al. (2022) used patients' clinical data to train counterfactual inference models that simulate intervention outcomes. Twitter and Reddit datasets are useful for identifying causal relationships in mental health using NLP-based causal inference (Doan et al., 2019; Li et al., 2023). At the clinical level, RCTs and causal discovery models utilise psychiatric datasets, integrating electronic health records, social media behaviours, and wearable device data to build comprehensive causal frameworks for mental health research (Gao et al., 2019; Pearl et al., 2016) (Refer to Table 5).

4.3.3 Data used in deep learning techniques for mental health

Research in deep learning for mental health is beginning to leverage heterogeneous, rich datasets of human communication and behaviour, including datasets spanning social media text and speech, clinical data, and even multimodal communication (Chikersal et al., 2020; Kuhail et al., 2024). Text-based studies often rely on large-scale social media datasets, from Twitter datasets (e.g., 2,575 users), Arabic tweets, Reddit threads, and even Spanish-speaking social media platforms, to identify linguistic patterns associated with depression and distressed individuals (Almars, 2022; López-Ubeda et al., 2019; Zhang et al., 2021).

Speech-based approaches further expand upon this scope by using speech datasets such as DAIC-WOZ, MODMA, and AViD-Corpus, which contain recordings of depressed individuals to identify acoustic and vocal patterns that may be indicative of depressive symptoms (Yin et al., 2023; Zhang et al., 2021). More recently, multi-modal datasets of text, speech, and physiological data, such as Weibo depression datasets, social media trackers, Empatica E4 sensor data, and CEASEv2.0 (a suicidal behaviour dataset), have been employed to generate more accurate representations of depressed individuals for their detection via these systems (Li et al., 2023; Moser et al., 2024; Zhang et al., 2024). Relatedly, clinical datasets from Electronic Health Records (EHR) and structured and unstructured self-reported mental health questionnaires are also being integrated to feed these AI-based diagnostic tools (Martín & Camacho, 2022; Obeid et al., 2019).

The focus on social media datasets is particularly interesting; this reflects the unique value of rich communication datasets collected in 'in-the-wild' communicative contexts, where communications related to mental health are embedded within everyday interactions, instances of self-disclosure, feedback from peers and networks, and even sources of networked social support (Pearl & Mackenzie, 2018). Communication on social media has an inherent timestamp, is inherently public in nature (and may also be in its treatment by the AI), and provides a resource for this type of unfiltered 'naturalistic' text-based communication at a scale that outpaces traditional models of (more structured) clinical data (Doan et al., 2019). In other words, social media communication presents a unique opportunity to train deep learning models for the creation of early-detection models (Razzaq & Shah, 2025). Beyond the implications for early detection models discussed here, however, this line of research may also have implications for communication practices related to those systems, in terms of how the systems communicate with users, how the platforms communicate to their users about new policies concerning content, and even the potential handling of ethical issues related to public communication, stigma, or fairness (Zainal et al., 2024). (Refer to Tables 6 and 7).

4.3.4 Data used in causal methods in social networks for mental health

Causal methods for social networks rely on big data and simulated social interaction on social media platforms. Treur (2017) used synthetic datasets to simulate social networks and evaluate their impact on mental health conditions. These simulations help model how peer influence and evolving social dynamics affect mental well-being (Treur, 2017). Social media-based research typically uses Twitter, online forums, and other social media datasets to analyse user-generated content related to mental health (Zhao et al., 2021). Moreover, for causal Bayesian Networks, social media and online platform data are utilised to predict mental health (Emmert-Streib & Dehmer, 2021). However, relying on publicly available social media data raises challenges related to privacy, consent and data quality. (Refer to Table 8).

4.4 Dimension 4: Challenges

4.4.1 Challenges in AI for early mental health intervention

Despite technological advances, AI for early mental health intervention poses various challenges, including explainability, biases, privacy concerns, and real-world limitations (Bernasconi, 2025). Oei et al. (2023) noted that decision trees and XGBoost models are effective for medical prediction but require clinical interpretability. NLP models analysing social media posts face challenges in detecting sarcasm, indirect references, and platform-specific variations (Myee et al., 2024). Deep learning models such as BERT and LSTMs require substantial computational resources and struggle to generalise across different populations (Uddin et al., 2022). AI-based digital therapy encounters privacy concerns and ethical risks, particularly in the confidentiality of therapist-client conversations (Chikersal et al., 2020). Mental health monitoring gadgets depend on sensor accuracy and user compliance, as patients may not consistently wear IoMT devices (Sandulescu et al., 2024). Hazra-Ganju et al. (2023) declared that low-resource settings can overcome internet connectivity issues and cultural adaptation barriers to ensure accessibility to everyone. Moreover, Yoon (2024) mentioned that AI-driven interventions for youth mental health require robust safeguards to control the use and ensure effective crisis response. (Refer to Table 4).

4.4.2 Challenges in causal inference in mental health

Despite its capabilities, causal inference in mental health research faces several challenges, including data quality, methodological limitations, and ethical concerns (Afrihyia et al., 2025). Causal research using DiD comes with the challenge of controlling confounders and maintaining parallel trend assumptions, thereby eliminating biases from external influences (Ghaffari et al., 2024; Yuan et al., 2023). It is challenging to model unobserved confounders, which limits generalisability across clinical settings (Marchezini et al., 2022). NLP-based causal models struggle with data noise, linguistic ambiguity, and accuracy issues, particularly when analysing short-text messages on Twitter and Reddit (Doan et al., 2019; Li et al., 2023). Pearl et al. (2016) mentioned that medical mental health research must address confounding factors and ethical concerns using RCTs and natural experiments. Gao et al. (2019) encountered high dimensionality and challenges with causal relationships when dealing with big data, leading to false causal inferences. (Refer to Table 5).

4.4.3 Challenges in deep learning techniques for mental health

Although DL plays an important role in mental health prediction and intervention, it faces several technical, ethical, and practical challenges. Almars (2022) identified challenges, including dataset limitations, language biases, and an imbalanced class distribution, and noted challenges with accurate model training and generalisation. For clinical settings, model interpretability and explainability become challenging using complex DL models (Moser et al., 2024; Zhang et al., 2024). AI-based mental health prediction model deployment can be complicated by risks such as ethical and privacy issues, real-time moderation, and potential biases (Hasan et al., 2023). However, in multimodal DL approaches, the complexity of feature fusion increases during behavioural analysis, limiting dataset scalability and real-world applicability (Li et al., 2023; Yin et al., 2023). In resource-limited healthcare settings, computational costs act as a barrier to widespread adoption (Martín & Camacho, 2022; Squires et al., 2023). (Refer to Tables 6 and 7).

4.4.4 Challenges in causal methods in social networks for mental health

In social networks, causal methods for mental health prediction pose several challenges, such as data complexity, ethical considerations, and model validation. High computational power is required for real-world social interaction, as Treur (2017) demonstrated using simulated datasets. Dynamic networks and unpredictable environments limit their structured nature, as captured by a causal Bayesian model (Emmert-Streib & Dehmer, 2021). Additionally, complex NLP techniques are required to extract information from social media text due to language barriers and cultural differences (Zhao et al., 2021). Ethical concerns due to user consent, privacy, and the consumption of public social media data can complicate causal inference in mental health research. Furthermore, the generalisability of findings across diverse populations and languages remains a key challenge (Refer to Table 8).

In brief, the advantages of using social media as a communication source for deriving communication and texting models for analyses of mental health conditions are offset by important considerations that extend beyond these technical issues. Mental health communication on social media is heavily contextualised by communication-based issues, such as self-presentation, as well as social norms around stigma, making it difficult for AI/ML models to avoid misinterpreting posts and misunderstanding their intent. The ethical considerations of privacy protection, informed consent, and the potential for misuse are especially important, as mental health communication often entails self-disclosure of sensitive information in public or semi-public digital spaces. Social media is also not a representative sample of all populations, resulting in demographic, behavioural, and selection biases in which voices are heard and whose distress is ignored. The spread of misinformation and harmful discourse also has the potential to shape mental health predictions and online support communities. Thus, deep learning and causal inference models for analysing mental health communication on social media should incorporate contextualised communication modelling and robust ethical considerations. Future progress in this field should aim for interpretable, fair AI systems, reduce biases in model training data, and use privacy-preserving techniques.

5.0 Research limitations

This research has some limitations. First, there was a lack of specific AI operations using causal inference and deep learning techniques that could not be accessed, as academic papers often lack details on how these methods work, given that these methods are largely proprietary. Second, despite deploying a thorough search strategy, some studies on AI in mental health were not included, such as grey literature and reports not published in the selected databases.

The nature of social media data presents challenges in establishing causality due to unobserved confounding and temporal ambiguity (Pearl, 2009). Deep learning models, while powerful, often operate as black boxes, lacking the interpretability crucial for mental health applications (Dinu & Moldovan, 2021; López-Úbeda et al., 2019). Data quality and its explainability are also concerns, as social media users do not reflect the general population, and mental health labels often lack clinical validation (Chancellor et al., 2019; De Choudhury et al., 2013). Moreover, Ethical concerns such as privacy, informed consent, and the risk of social exclusion further complicate research in this space (Petrolini et al., 2022). Addressing these limitations requires interdisciplinary efforts and the development of transparent, context-aware, and ethically grounded AI systems.

6.0 Conclusion

Integrating causal inference and deep learning in social networks offers a powerful approach for early mental health intervention. As mental health issues continue to rise globally, the growth in this research demonstrates how AI-powered systems can harness social media, speech, and behavioural data to detect early signs of depression, anxiety, and stress. Causal methods such as Difference-in-Differences (DiD), Propensity Score Matching (PSM), and Structural Causal Models (SCMs) identify causal relationships between social interactions and mental health outcomes. Meanwhile, deep learning models such as BERT, LSTMs, and multimodal fusion analyses of sentiment patterns, speech signals, and facial expressions enable scalable, accurate detection of depression, anxiety, and stress from social media and behavioural data.

Despite technological advancements, significant challenges remain, including model bias, privacy and ethical concerns, limited interpretability, and communication-related complexities. Causal inference models require careful control of confounding factors to estimate cause-and-effect relationships, particularly in observational social media settings where mental health expressions are shaped by context, social interaction, and self-disclosure practices. Similarly, deep learning models depend on large, labelled datasets and substantial computational resources, while also struggling to explain their predictions in ways that are meaningful to clinicians and sensitive to the communicative nuances of online discourse. Moving forward, integrating causal inference with deep learning holds strong potential for developing personalised, data-driven mental health interventions, but future research must prioritise explainable AI, privacy-preserving methods, and ethically grounded, communication-aware frameworks. By combining causal reasoning with advanced machine learning, researchers can move closer to building transparent, reliable, and equitable early mental health detection systems that respect the social and communicative contexts in which distress is expressed.

Acknowledgement

The authors would like to extend their heartfelt gratitude to the Ministry of Higher Education Malaysia for the approval of the Fundamental Research Grant Scheme 2024 (FRGS).

Conflict of Interest

The author (s) have declared that no competing interests exist.

Author Contribution Statement

UN: Conceptualization, Data Curation, Methodology, Validation, Writing – Original Draft Preparation. NSK: Project Administration, Writing – Review & Editing; MAJ: Project Administration, Supervision, Writing – Review & Editing.

Funding

This work is supported by the Fundamental Research Grant Scheme 2024 by the Ministry of Higher Education Malaysia, FRGS/1/2024/ICT02/UMP/02/3 (RDU240125).

Ethics Statement

This research did not require IRB approval because it involved a retrospective analysis of fully anonymized data previously collected.

Data Availability:

The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding author.

Author Biography

Umber Nazir is a PhD researcher in Computing at UMPSA whose research focuses on artificial intelligence, deep learning, causal inference, social media analysis, and mental health-related studies. She has also contributed to research in disaster risk, flood vulnerability, and GIS-based analysis.

Ts. Dr. Shazwani Kamarudin is a Senior Lecturer in Software Engineering at UMPSA's Faculty of Computing. She has degrees in Computer Science from Universiti Sultan Zainal Abidin and a Ph.D. from Arizona State University. Her research focuses on social media analysis, especially mental health and human behaviour studies using social media data.

Professor Ts. Dr. Mazlina Binti Abdul Majid is an experienced academic at UMPSA's Faculty of Computing with 23 years of teaching experience. She earned her Ph.D. in Computer Science from the University of Nottingham, UK. She has held several leadership roles, including Deputy Dean, research group head, lab head, and Editor-in-Chief of a UMPSA journal. Her research focuses on simulation modelling, green technology, software usability testing, and software agents.

References

Abid, S. K., Roosli, R., Nazir, U., & Kamarudin, N. S. (2025). AI-enhanced crowdsourcing for disaster management: strengthening community resilience through social media. *International Journal of Emergency Medicine*, *18*(1), 201

Afrihyia, E., Chianumba, E. C., Mustapha, A. Y., Akomolafe, O. O., Omotayo, O., & Forkuo, A. Y. (2025). Protecting Mental Health Rights in the Digital Space: Legal and Ethical Considerations. *Journal of Frontiers in Multidisciplinary Research and Growth Evaluation*, *5*

Aleem, S., Huda, N. u., Amin, R., Khalid, S., Alshamrani, S. S., & Alshehri, A. (2022). Machine learning algorithms for depression: Diagnosis, insights, and research directions. *Electronics*, 11(7), 1111. https://doi.org/10.3390/electronics11071111

Ali, O., Abdelbaki, W., Shrestha, A., Elbasi, E., Alryalat, M. A. A., & Dwivedi, Y. K. (2023). A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *Journal of Innovation & Knowledge*, 8(1), Article 100333. https://doi.org/10.1016/j.jik.2023.100333

Ali, O., Ally, M., Clutterbuck, P., & Dwivedi, Y. (2020). The state of play of blockchain technology in the financial services sector: A systematic literature review. *International Journal of Information Management*, 54, 102199

Ali, O., Jaradat, A., Kulakli, A., & Abuhalimeh, A. (2021). A comparative study: Blockchain technology utilization benefits, challenges and functionalities. *IEEE Access, 9*, 12730–12749

Ali, O., Shrestha, A., Soar, J., & Wamba, S. F. (2018). Cloud computing enabled healthcare opportunities, issues, and applications: A systematic review. *International Journal of Information Management, 43*, 146–158. https://doi.org/10.1016/j.ijinfomgt.2018.07.009

Al-Laith, A., & Alenezi, M. (2021). Monitoring people's emotions and symptoms from Arabic tweets during the COVID-19 pandemic. *Information*, 12(2), 86. https://doi.org/10.3390/info12020086

Almars, A. M. (2022). Attention-based Bi-LSTM model for Arabic depression classification. Computers, Materials & Continua, 71(2), 3091–3106. https://doi.org/10.32604/cmc.2022.022609

Alshahrani, A., Almatrafi, M. M., Mustafa, J. I., Albaqami, L. S., & Aljabri, R. A. (2024). A children's psychological and mental health detection model by drawing analysis based on computer vision and deep learning. *Engineering, Technology & Applied Science Research, 14*(4), 15533–15540

Alshanketi, F. (2024). Revolutionising generalised anxiety disorder detection using a deep learning approach with MGADHF architecture on social media. *International Journal of Advanced Computer Science & Applications*, 15(1)

Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale analysis of counselling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 5, 463–476

Alturayeif, N., & Luqman, H. (2021). Fine-grained sentiment analysis of Arabic COVID-19 tweets using BERT-based transformers and dynamically weighted loss function. *Applied Sciences, 11*(22), 10694. https://doi.org/10.3390/app112210694

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8

Bashar, M. A., Nayak, R., & Balasubramaniam, T. (2022). Deep learning-based topic and sentiment analysis: COVID-19 information seeking on social media. *Social Network Analysis and Mining, 12*(1), 90. https://doi.org/10.1007/s13278-022-00916-0

Bauer, B., Norel, R., Leow, A., Rached, Z. A., Wen, B., & Cecchi, G. (2024). Using large language models to understand suicidality in a social media–Based Taxonomy of Mental Health Disorders: Linguistic Analysis of Reddit posts. *JMIR Mental Health*, *11*, e57234. https://doi.org/10.2196/57234

Bernasconi, A. (2025). From Real-World Data to Causal Models: A Bayesian Network to study cardiovascular diseases in Adolescents and Young Adults with Breast Cancer

Boell, S. K., & Cecez-Kecmanovic, D. (2015). On being "systematic" in literature reviews in IS. *Journal of Information Technology, 30*(2), 161–173. https://doi.org/10.1057/jit.2014.26

Bokolo, B. G., & Liu, Q. (2022). Deep learning-based depression detection from social media: Comparative evaluation of ML and transformer techniques. *Electronics, 12*(21), 4396. https://doi.org/10.3390/electronics12214396

Cabezas-Klinger, H., Fernandez-Daza, F. F., & Mina-Paz, Y. (2025). Associations between Social Media Use and Mental Disorders in Adolescents and Young Adults: A Systematic Review and Meta-Analysis of Recent Evidence. *Behavioral Sciences*, *15*(11), 1450. https://doi.org/10.3390/bs15111450

Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M. B., & De Choudhury, M. (2019). A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 79–88). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287587

Chikersal, P., Belgrave, D., Doherty, G., Enrique, A., Palacios, J. E., Richards, D., & Thieme, A. (2020). Understanding client support strategies to improve clinical outcomes in an online mental health intervention. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376728

Dabić, M., Vlačić, B., Paul, J., Dana, L.-P., Sahasranamam, S., & Glinka, B. (2020). Immigrant entrepreneurship: A review and research agenda. *Journal of Business Research, 113*, 25–38. https://doi.org/10.1016/j.jbusres.2020.03.013

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media, 7*(1), 128–137. https://doi.org/10.1609/icwsm.v7i1.14432

Ding, Z., Wang, Z., Zhang, Y., Cao, Y., Liu, Y., Shen, X., Tian, Y., & Dai, J. (2025). Trade-offs between machine learning and deep learning for mental illness detection on social media. *Scientific Reports*, *15*(1), 14497. https://doi.org/10.1038/s41598-025-99167-6

Dinu, A., & Moldovan, A. C. (2021). Automatic detection and classification of mental illnesses from general social media texts. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 358–366. https://aclanthology.org/2021.ranlp-1.41/

Doan, S., Yang, E. W., Tilak, S. S., Li, P. W., Zisook, D. S., & Torii, M. (2019). Extracting health-related causality from Twitter messages using natural language processing. *BMC Medical Informatics and Decision Making, 19*, 71. https://doi.org/10.1186/s12911-019-0785-0

Du, J., Zhang, Y., Luo, J., Jia, Y., Wei, Q., Tao, C., & Xu, H. (2018). Extracting psychiatric stressors for suicide from social media using deep learning. *BMC Medical Informatics and Decision Making, 18*(Suppl 2), 43. https://doi.org/10.1186/s12911-018-0628-2

Emmert-Streib, F., & Dehmer, M. (2021). Data-driven computational social network science: Predictive and inferential models for web-enabled scientific discoveries. *Frontiers in Big Data, 4*. https://doi.org/10.3389/fdata.2021.591749

Ezerceli, O., & Dehkharghani, R. (2024). Mental disorder and suicidal ideation detection from social media using deep neural networks. *Journal of Computational Social Science, 7*(3), 2277–2307. https://doi.org/10.1007/s42001-024-00293-7

Gao, G., Wang, T., Zheng, X., Chen, Y., & Xu, X. (2019). A systems dynamics simulation study of network public opinion evolution mechanism. *Journal of Global Information Management, 27*(4), 189–207. https://doi.org/10.4018/JGIM.2019100109

Geethanjali, R., & Valarmathi, A. (2024). A deep learning based hybrid model for maternal health risk detection and multifaceted emotion analysis in social networks. *International Journal of Applied Mathematics and Computer Science, 34*(4), 565–577. https://doi.org/10.34768/amcs-2024-0043

Gongane, V. U., Munot, M. V., & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: Current status and future directions. *Social Network Analysis and Mining, 12*(1), 129. https://doi.org/10.1007/s13278-022-01025-3

Harrigian, K., Aguirre, C., & Dredze, M. (2021). On the State of Social Media Data for Mental Health Research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access* (pp. 15–24). Johns Hopkins University. https://doi.org/10.18653/v1/2021.clpsych-1.2

Hasan, M. T., Hossain, M. A. E., Mukta, M. S. H., Akter, A., Ahmed, M., & Islam, S. (2023). A review on deep-learning-based cyberbullying detection. *Future Internet, 15*(5), 179. https://doi.org/10.3390/fi15050179

Hassan, E., & Omenogor, C. E. (2025). AI powered predictive healthcare: Deep learning for early diagnosis, personalized treatment, and disease prevention. *International Journal of Science and Research Archive.* 14(03), 806-823. https://doi.org/10.30574/ijsra.2025.14.3.0731

Hazra-Ganju, A., Dlima, S. D., Menezes, S. R., Ganju, A., & Mer, A. (2023). An omni-channel, outcomes-focused approach to scale digital health interventions in resource-limited populations: A case study. *Frontiers in Digital Health, 5*, 1007687. https://doi.org/10.3389/fdgth.2023.1007687

Hogan, J. W. (2019). Causal Inference in Statistics: A Primer Judea Pearl, Maria Glymour, and Nicholas Jewell, John Wiley & Sons, Ltd., Chichester, UK. *Biometrics*, *75*(2), 708–709. https://doi.org/10.1111/biom.13079

Hu, Y., & Bai, G. (2014). A systematic literature review of cloud computing in eHealth. *arXiv preprint* arXiv:1412.2494. https://arxiv.org/abs/1412.2494

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction.* Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Jayasudha, J. (2025). A review of Transformer-Based Models for Natural Language Processing (NLP). *International Journal for Research in Applied Science and Engineering Technology*, *13*(5), 7000–7004. https://doi.org/10.22214/ijraset.2025.71832

Jabeen, F., Gerritsen, C., & Treur, J. (2020). From victim to survivor: A multilayered adaptive mental network model of a bully victim. In L. Rutkowski et al. (Eds.), *Artificial Intelligence and Soft Computing* (pp. 679–689). Springer. https://doi.org/10.1007/978-3-030-61534-5_59

Jadon, A. K., & Kumar, S. (2024). Enhancing emotion detection with synergistic combination of word embeddings and convolutional neural networks. *Indonesian Journal of Electrical Engineering and Computer Science*. http://doi.org/10.11591/ijeecs.v35.i3.pp1933-1941

Jawad, K., Mahto, R., Das, A., Ahmed, S. U., Aziz, R. M., & Kumar, P. (2023). Novel cuckoo search-based metaheuristic approach for deep learning prediction of depression. *Applied Sciences, 13*(9), 5322. https://doi.org/10.3390/app13095322

Jiao, L., Wang, Y., Liu, X., Li, L., Liu, F., Ma, W., Guo, Y., Chen, P., Yang, S., & Hou, B. (2024). Causal Inference Meets Deep Learning: A comprehensive survey. *Research*, *7*, 0467. https://doi.org/10.34133/research.0467

Junyi, Z. (2024). The application value of behavior analysis based on deep learning in the evaluation of depression in art students. *Applied Mathematics and Nonlinear Sciences, 9*(1)

Kawam, B., Ostner, J., McElreath, R., Schülke, O., & Redhead, D. (2025). A causal framework for the drivers of animal social network structure. *PLoS Computational Biology*, *21*(9), e1013370. https://doi.org/10.1371/journal.pcbi.1013370

King, C. A., Gipson Allen, P. Y., Ahamed, S. I., Webb, M., Casper, T. C., Brent, D., Grupp-Phelan, J., Rogers, T. A., Arango, A., Al-Dajani, N., McGuire, T. C., & Bagge, C. L. (2024). 24-Hour warning signs for adolescent suicide attempts. *Psychological medicine*, *54*(7), 1272–1283. doi:10.1017/S0033291723003112

Kuhail, M. A., Alturki, N., Thomas, J., Alkhalifa, A. K., & Alshardan, A. (2024). Human–human vs. human–AI therapy: An empirical study. *International Journal of Human–Computer Interaction.* https://doi.org/10.1080/10447318.2023.2283419

Kumari, M., Singh, G., & Pande, S. D. (2024). Depressonify: BERT a deep learning approach for detection of depression. *EAI Endorsed Transactions on Pervasive Health & Technology, 10*(1). https://doi.org/10.4108/eetph.v10i1.2678

Li, Z., An, Z., Cheng, W., Zhou, J., Zheng, F., & Hu, B. (2023). MHA: A multimodal hierarchical attention model for depression detection in social media. *Health Information Science and Systems, 11*(1), 6. https://doi.org/10.1007/s13755-023-00237-3

Lin, L., Chen, X., Shen, Y., & Zhang, L. (2020). Towards automatic depression detection: A BiLSTM/1D CNN-based model. *Applied Sciences, 10*(23), 8701. https://doi.org/10.3390/app10238701

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue, 16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

López-Ubeda, P., Plaza-del Arco, F. M., Díaz-Galiano, M. C., Ureña-López, L. A., & Martín-Valdivia, M. T. (2019). Detecting anorexia in Spanish tweets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (pp. 655–663). https://doi.org/10.26615/978-954-452-056-4_077

Malhotra, A., & Jindal, R. (2020). *Multimodal deep learning-based framework for detecting depression and suicidal behaviour by affective analysis of social media posts*. EAI Endorsed Transactions on Pervasive Health and Technology, 6(21), e1. https://doi.org/10.4108/eai.13-7-2018.15962

Marchezini, G. F., Lacerda, A. M., Pappa, G. L., Miranda, D., Romano-Silva, M. A., Costa, D. S., & Diniz, L. M. (2022). Counterfactual inference with latent variables and its application in mental health care. *Data Mining and Knowledge Discovery*. https://doi.org/10.1007/s10618-021-00818-9

Martín, A., & Camacho, D. (2022). Recent advances on effective and efficient deep learning-based solutions. *Neural Computing and Applications, 34*(13), 10205–10210. https://doi.org/10.1007/s00521-022-07272-z

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ, 339*, b2535. https://doi.org/10.1136/bmj.b2535

Ghaffari, M., Khan, G. F., Singh, S. P., & Ferwerda, B. (2024). The impact of COVID-19 on online music listening behaviors in light of listeners' social interactions. *Multimedia Tools and Applications, 83*, 13197–13239. https://doi.org/10.1007/s11042-023-16079-1

Moser, M. K., Ehrhart, M., & Resch, B. (2024). An explainable deep learning approach for stress detection in wearable sensor measurements. *Sensors, 24*(16), 5085. https://doi.org/10.3390/s24165085

Munir, M. M., & Ahmed, N. (2025). Using social media platforms to raise health awareness and increase health education in Pakistan: Structural equation modeling analysis and questionnaire study. *JMIR Human Factors*, *12*, e65745. https://doi.org/10.2196/65745

Muñoz, S., & Iglesias, C. Á. (2023). Detection of the severity level of depression signs in text. *Applied Sciences, 13*(21), 11695. https://doi.org/10.3390/app132111695

Myee, M. K., Rebekah, R. D. C., Deepa, T., Zion, G. D., & Lokesh, K. (2024). Detection of depression in social media posts using emotional intensity analysis. *Engineering, Technology & Applied Science Research, 14*(5), 16207–16211

Nijhawan, T., Attigeri, G., & Ananthakrishna, T. (2022). Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data, 9*(1), 33. https://doi.org/10.1186/s40537-022-00575-6

Oakey-Neate, L., Schrader, G., Strobel, J., Bastiampillai, T., van Kasteren, Y., & Bidargaddi, N. (2020). Using algorithms to initiate needs-based interventions for people on antipsychotic medication: Implementation protocol. *BMJ Health & Care Informatics, 27*(1), e100155. https://doi.org/10.1136/bmjhci-2019-100084

Obeid, J. S., Weeda, E. R., Matuskowitz, A. J., Gagnon, K., Crawford, T., Carr, C. M., & Frey, L. J. (2019). Automated detection of altered mental status in emergency department clinical notes: A deep learning approach. *BMC Medical Informatics and Decision Making, 19*, 164. https://doi.org/10.1186/s12911-019-0894-9

Oei, C. W., Ng, E. Y. K., Ng, M. H. S., Tan, R. S., Chan, Y. M., Chan, L. G., & Acharya, U. R. (2023). Explainable risk prediction of poststroke adverse mental outcomes using machine learning techniques in a population of 1780 patients. *Sensors, 23*(18), 7946. https://doi.org/10.3390/s23187946

Ohlsson, H., & Kendler, K. S. (2019). Applying causal inference methods in psychiatric epidemiology. *JAMA Psychiatry*, *77*(6), 637. https://doi.org/10.1001/jamapsychiatry.2019.3758

Omarov, B., & Zhumanov, Z. (2023). Bidirectional long-short-term memory with attention mechanism for emotion analysis in textual content. *International Journal of Advanced Computer Science and Applications, 14*(6)

Ozkanca, Y., Göksu Öztürk, M., Ekmekci, M. N., Atkins, D. C., Demiroglu, C., & Hosseini Ghomi, R. (2019). Depression screening from voice samples of patients affected by Parkinson's disease. *Digital Biomarkers, 3*(2), 72–82. https://doi.org/10.1159/000502136

Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chisholm, D., Collins, P. Y., Cooper, J. L., Eaton, J., Herrman, H., Herzallah, M. M., Huang, Y., Jordans, M. J. D., Kleinman, A., Medina-Mora, M. E., Morgan, E., Niaz, U., Omigbodun, O., & Unützer, J. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet, 392*(10157), 1553–1598. https://doi.org/10.1016/S0140-6736(18)31612-X

Patil, M., & Khedkar, V. (2023). An automated system for depression detection based on facial and vocal features. *International Journal on Recent and Innovation Trends in Computing and Communication, 11*(7), 286–291. https://doi.org/10.17762/ijritcc.v11i7s.7001

Paul, J., & Benito, G. R. G. (2018). A review of research on outward foreign direct investment from emerging countries, including China. *Asia Pacific Business Review, 24*(1), 90–115. https://doi.org/10.1080/13602381.2017.1357316

Paul, J., Lim, W. M., O'Cass, A., Hao, A. W., & Bresciani, S. (2021). Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). *International Journal of Consumer Studies, 45*(4), O1–O16. https://doi.org/10.1111/ijcs.12695

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. Wiley

Petrolini, M., Cagnoni, S., & Mordonini, M. (2022). Automatic detection of sensitive data using transformer-based classifiers. *Future Internet, 14*(8), 228. https://doi.org/10.3390/fi14080228

Pucher, K., Boot, N., & De Vries, N. (2013). Systematic review: School health promotion interventions... *Health Education, 113*(5), 372–391. https://doi.org/10.1108/HE-02-2012-0008

Rashed, A. E. E., Atwa, A. E. M., Ahmed, A., Badawy, M., Elhosseini, M. A., & Bahgat, W. M. (2024). Facial image analysis for automated suicide risk detection with deep neural networks. *Artificial Intelligence Review, 57*(10), 274. https://doi.org/10.1007/s10462-023-10586-2

Razzaq, K., & Shah, M. (2025). Machine learning and deep learning paradigms: from techniques to practical applications and research frontiers. *Computers*, *14*(3), 93. https://doi.org/10.3390/computers14030093

Rosado-Serrano, A., Paul, J., & Dikova, D. (2018). International franchising: A literature review and research agenda. *Journal of Business Research, 85*, 238–257. https://doi.org/10.1016/j.jbusres.2017.12.049

Rudin, C. (2019). Stop explaining black box machine learning models... *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sah, A. K., Elshaikh, R. H., Shalabi, M. G., Abbas, A. M., Prabhakar, P. K., Babker, A. M. A., Choudhary, R. K., Gaur, V., Choudhary, A. S., & Agarwal, S. (2025). Role of artificial intelligence and personalized medicine in enhancing HIV management and treatment outcomes. *Life, 15*(5), Article 745. https://doi.org/10.3390/life15050745

Sandulescu, V., Ianculescu, M., Valeanu, L., & Alexandru, A. (2024). Integrating IoMT and AI for proactive healthcare. *Algorithms, 17*(9). https://doi.org/10.3390/a17090399

Saranya, S., & Kavitha, N. (2022). Robust feature selection with chicken swarm intelligence... *Journal of Theoretical and Applied Information Technology, 100*(10), 3337–3345

Schulam, P., & Saria, S. (2019). Can you trust this prediction? *AISTATS Proceedings, 89*, 1022–1031

Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Acharya, U. R., & Li, Y. (2023). Deep learning and machine learning in psychiatry. *Brain Informatics, 10*(1), 10. https://doi.org/10.1186/s40708-023-00212-5

Torous, J., Linardon, J., Goldberg, S. B., Sun, S., Bell, I., Nicholas, J., Hassan, L., Hua, Y., Milton, A., & Firth, J. (2025). The evolving field of digital mental health: Current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *World Psychiatry, 24*(2), 156–174. https://doi.org/10.1002/wps.21299

Treur, J. (2017). Modelling the dynamics of adaptive temporal-causal networks. *Computational Social Networks, 4*(1), 1–20. https://doi.org/10.1186/s40649-017-0044-z

Uddin, M. Z., Dysthe, K. K., Følstad, A., & Brandtzaeg, P. B. (2022). Deep learning for prediction of depressive symptoms in text. *Neural Computing and Applications, 34*(1), 721–744. https://doi.org/10.1007/s00521-021-06430-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6000–6010). Curran Associates, Inc. doi:10.5555/3295222.3295349

Watson, R. T. (2015). Beyond being systematic in literature reviews. *Journal of Information Technology, 30*(2), 185–187. https://doi.org/10.1057/jit.2014.30

Yang, W. (2023). Extraction and analysis of factors influencing college students' mental health based on deep learning. *Applied Mathematics and Nonlinear Sciences*

Yang, X., & Li, G. (2025). Psychological and Behavioral Insights from Social Media users: Natural Language Processing–Based Quantitative Study on Mental Well-Being. JMIR Formative Research, 9, e60286. https://doi.org/10.2196/60286

Yin, F., Du, J., Xu, X., & Zhao, L. (2023). Depression detection in speech using transformer... *Electronics, 12*(2), 328. https://doi.org/10.3390/electronics12020328

Yoon, S. (2024). AI-based digital therapeutics for adolescent mental health. *Information, 15*(10). https://doi.org/10.3390/info15100511

Yuan, Y., Saha, K., Keller, B., Isometsä, E. T., & Aledavood, T. (2023). Mental health coping stories on social media: A causal inference study. *Proceedings of the ACM Web Conference 2023*, 2677–2685. https://doi.org/10.1145/3543507.3583451

Zainal, N. A., Asnawi, A. L., Jusoh, A. Z., Ibrahim, S. N., & Mohd Ramli, H. A. (2024). Integration of MFCCs and CNN for multi-class stress speech classification. *IIUM Engineering Journal, 25*(2)

Zanwar, S., Wiechmann, D., Li, X., Qiao, Y., & Kerz, E. (2023). What to fuse and how to fuse: Emotion + personality fusion strategies. *Findings of ACL 2023*, 8926–8940. https://doi.org/10.18653/v1/2023.findings-acl.559

Zhang, Q., Huang, Z., Sui, Y., Lin, F., Guan, H., Li, L., Wang, K., & Neitzel, A. (2025). Social-Media-Based Mental Health Interventions: Meta-Analysis of Randomized Controlled Trials. Journal of Medical Internet Research, 27, e67953. https://doi.org/10.2196/67953

Zhang, Y., Lyu, H., Liu, Y., Zhang, X., Wang, Y., & Luo, J. (2021). Monitoring depression trends on Twitter during COVID-19. *JMIR Infodemiology, 1*(1), e26769. https://doi.org/10.2196/26769

Zhang, Z., Zhang, S., Ni, D., Wei, Z., Yang, K., Jin, S., Huang, G., Liang, Z., Zhang, L., Li, L., Ding, H., Zhang, Z., & Wang, J. (2024). Multimodal sensing for depression risk detection. *Sensors, 24*(12), 3714. https://doi.org/10.3390/s24123714

Zhao, Y., Liang, Z., Du, J., Zhang, L., Liu, C., & Zhao, L. (2021). Multihead attention-based LSTM for depression detection from speech. *Frontiers in Neurorobotics, 15*, 684037. https://doi.org/10.3389/fnbot.2021.684037

Zhao, Y., Yu, Y., Wang, H., Li, Y., Deng, Y., Jiang, G., & Luo, Y. (2022). Machine learning in causal inference: application in Pharmacovigilance. *Drug Safety*, *45*(5), 459–476. https://doi.org/10.1007/s40264-022-01155-6