INPLASY

INPLASY202690118

doi: 10.37766/inplasy2026.9.0118

Received: 28 September 2026

Published: 28 September 2026

Machine Learning and Natural Language Processing for Mental Health Detection in Clinical Text: A Scoping Review

Ding, XC; Lam, CLM; Yim, SH; Cheung, AK; Lee, TMC.

Corresponding author:
Xiaochen Ding

u3014640@connect.hku.hk

Author Affiliation:
InnoCentre of ClinicalNeuropsychology, The University ofHong Kong, Hong Kong SAR, China.The University of Hong Kong.

ADMINISTRATIVE INFORMATION

Support - No support.

Review Stage at time of this submission - Completed but not published.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY202690118

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 28 September 2026 and was last updated on 28 September 2026.

INTRODUCTION

Review question / Objective The primary objective of this systematic review is to evaluate the classification performance of machine learning and natural language processing models in detecting formally diagnosed mental health disorders using unstructured real-life clinical text.

Background The widespread use of clinical text data in Electronic Health Records (EHRs) has accumulated a vast amount of unstructured information regarding patients' psychological states. Natural Language Processing (NLP) and Machine Learning (ML) can efficiently extract this textual information to aid in the early screening and identification of mental health issues, such as depression and anxiety. This improves the accuracy of interventions and advances precision mental healthcare.

Rationale Existing reviews on digital mental health frequently overemphasize highly prevalent conditions while overlooking a broader spectrum of mental health disorders. Furthermore, current studies heavily rely on single-institution datasets lacking multicenter validation, and they frequently prioritize technical performance metrics over actual clinical applicability. There is a need to systematically map the current status of ML-based clinical text analysis to analyze algorithmic architectures, diagnostic performance, data representation, and methodological limitations to provide a scientific basis for future applications.

METHODS

Strategy of data synthesis Data will be systematically extracted using a standardized charting form. We will synthesize study-level metadata, clinical context (targeted mental health conditions), and data representations (clinical text source, textual format, and language). Methodological details regarding NLP feature

extraction techniques, ML classification models, class imbalance handling strategies, and performance metrics (e.g., F1-score, AUC-ROC, accuracy) will be collated. The evidence will be mapped narratively and visually using tables and charts to summarize temporal trends, geographic distributions, and algorithmic paradigms.

Eligibility criteria Clinical text data must originate from authentic formal healthcare or clinical settings, such as EHR narratives or clinical interview transcripts. Studies relying on non-clinical texts, like social media or public forums, are excluded. The ground truth for mental health disorders must be based on formal clinical diagnoses or documented psychiatric criteria. Studies relying solely on self-reported diagnoses are excluded. Studies must employ a formal measure or methodology utilizing machine learning and natural language processing to actively detect, predict, or classify at least one specific mental health disorder from textual data. Peer-reviewed journal articles and full-length conference papers featuring completed experimental research published in English. Preprints, non-English papers, literature reviews, and meta-analyses are excluded.

Source of evidence screening and selection The screening process will adhere to PRISMA-ScR guidelines and occur in two stages. Stage 1 involves title and abstract screening conducted independently by two reviewers. Stage 2 involves retrieving full texts of potentially relevant articles, which will be independently assessed by the same two reviewers. Discrepancies at all stages will be reconciled through consensus discussion or adjudicated by a third reviewer if consensus cannot be reached.

Data management EndNote 2025 will be utilized to manage citations and remove duplicates. Independent blinded review libraries will be established so reviewers cannot access each other's inclusion decisions prior to consensus meetings. A standardized data charting form will be pilot-tested on five randomly selected studies and refined prior to independent data extraction by two researchers.

Reporting results / Analysis of the evidence Results will be reported systematically using the PRISMA-ScR framework, including a detailed flow diagram of the study selection process. The charted data will be summarized to reflect methodological and clinical characteristics. Additionally, the risk of bias for the included machine learning models will be assessed using the Prediction Model Risk of Bias Assessment Tool across four relevant domains.

Language restriction English.

Country(ies) involved Hong Kong (SAR).

Keywords Mental Health; Clinical Text; Natural Language Processing; Health Records; Scoping Review.

Dissemination plans The synthesized findings will be disseminated through publication in a peer-reviewed journal to offer practical guidance for the global digital transformation of mental health services and to expand the frontiers of computational psychiatry research.

Contributions of each author
Author 1 - Xiaochen Ding.
Author 2 - Charlene L.M. Lam.
Author 3 - See Heng Yim.
Author 4 - Amanda K. Cheung.
Author 5 - Tatia M.C. Lee.