

How Can We Build an AI To Diagnose Textually Ambiguous Mental Health Disorders from Subjective Patient Reports?

Dhairya Jain

Jayshree Periwal Global School, Jagatpura, Jaipur, Rajasthan 302033, India; jaindhairya21@gmail.com

ABSTRACT: Nearly 1 in 7 people in the world live with a mental disorder. Still, in many regions, only 7% of individuals get help due to a big treatment gap due to a lack of trained health professionals, cost issues, or stigma, especially in low- and middle-income countries. However, the advancements in the field of AI have been a big reason for the rise in interest in AI-based mental health screening tools that can analyze patient-generated texts, such as clinical notes, questionnaires, and social media posts. Previous studies in this domain have consistently demonstrated that AI systems often achieve strong results when diagnosing well-defined mental disorders such as major depressive disorder and generalized anxiety disorder, because they exhibit consistent linguistic markers and standard patient-reported outcomes. However, comparatively less progress has been made for mental health disorders characterized by overlapping and ambiguous symptoms, such as PTSD and subtypes of depression. These challenges highlight a critical limitation in current AI approaches, particularly in handling linguistic ambiguity and subjective patient-reported data. This paper presents a secondary analytical review of existing literature, focusing on model architectures, data sources, and evaluation frameworks, and proposes an uncertainty-aware AI architecture incorporating ambiguity detection, multi-label prediction, and human-centered evaluation.

KEYWORDS: Robotics and Intelligent Machines, Machine Learning, Mental Health, Diagnosis.

■ Introduction

The use of Artificial Intelligence (AI) in the mental health field has been growing for the last few years, especially in the areas of machine learning and natural language processing.¹⁻³ An AI system, through analyzing different forms of written language (clinical notes, self-report questionnaires, online text),⁴ has shown that it can detect mental health "signals" and come close to matching the performance of clinicians when diagnosing some well-defined mental health conditions (major depressive disorder, generalized anxiety disorder, etc.)⁵

These studies indicate that it may be possible to leverage AI tools to help expand access to mental health services on a larger scale through the use of Text-Based AI. Certain mental disorders can be easily classified or diagnosed using an AI tool. On the other hand, many disorders are difficult to classify or diagnose. The language used to describe depression subtypes, PTSD, etc, is very subjective, with a great deal of overlap between types of disorders, and will vary greatly between individuals. Because the majority of the information for diagnosing these disorders will come from patient-reported outcomes, and because these disorders do not have any objective diagnostic markers, they are often not diagnosed until the patient presents later in the course of a clinical presentation.⁶ Therefore, most of the current research on AI to date has treated the diagnosis of these disorders as a relatively straightforward classification task, placing little to no emphasis on factors like uncertainty, the need for explanation, and the fact that the majority of self-reported mental health experiences are very subjective. We still don't have a good understanding of the ways in which AI systems should be developed to consider the ambiguity of

language and how to more effectively incorporate Patient Reported Outcomes (PROs) into the development of AI systems.

Building AI systems that can diagnose and treat mental health disorders using inherently vague or ambiguous text will require examining three areas of research: Model Architecture, Model Training, and Model Evaluation. Recent advancements in the field of AI, such as chatbots, have shown the revolutionary potential of AI in the mental health field. While many chatbots are accurate in diagnosing disorders based on ideal datasets, their accuracy often decreases when used in the real world with subjective datasets. There is a lot of scope in filling this gap, as nowadays the demand for healthcare services is ever-increasing, and many countries are experiencing a shortage of healthcare practitioners, especially physicians.⁷ This shortage is very crucial to consider, as this often leads to delayed treatments, misclassifications, and poor outcomes. Addressing this issue can help us to improve early identification of disorders, and further in making a responsible development of an AI specific to mental health contexts. Currently, we have only one form of input, and that is text, but in the future, we can add speech, behavior, and facial expressions as part of the input. This study carries out a secondary analysis of existing AI methods in mental health research. Previous studies are organized by disorder type, data source, and evaluation metrics. The model will be evaluated based on both traditional machine learning measures and human-based evaluation measures.

Traditional measures of performance include accuracy, precision, recall, and the F1 -score, which assess the statistical accuracy of disorder classification as observed in other literature. Besides, human-centered indicators will be included through a Likert scale (an ordered rating system used in

surveys, Strongly agree - strongly disagree) to measure judgments of clarity, usefulness, appropriateness, and reliability, thus summarizing how patients and clinicians perceive the quality and credibility of the AI-produced screening results.⁸ This study adopts a secondary analytical review methodology, synthesizing existing literature in AI-based mental health diagnostics. Sources were selected based on relevance to model architectures, data inputs, and evaluation frameworks, with an emphasis on peer-reviewed academic research and established institutional reports.

■ Discussion

This paper has demonstrated how artificial intelligence works well in spotting mental health issues like major depression and anxiety, where speech patterns and symptoms are clear and consistent. Yet performance drops when dealing with conditions marked by unclear symptoms or overlapping traits such as PTSD or less typical forms of depression, where language may be subtle or indirect. Previous research papers tend to focus too much on numbers like accuracy rates instead of deeper concerns like how uncertain the model might be. Interpretation matters just as much, yet it often takes a back seat in current work. Even when models score high technically, they do not always earn user confidence if explanations feel opaque or irrelevant. Trust builds differently among individuals using chatbots or online platforms meant for emotional support. Clarity of response and whether people find them helpful shape actual usage far more than backend precision ever could. Outcomes depend heavily on personal experience, something algorithms rarely capture fully. Moving forward, intelligent systems should assist professionals rather than replace them. Including self-reports from patients alongside transparent results can guide better decisions. Oversight by trained clinicians ensures risks stay low while keeping care grounded in real needs.

■ Mental health diagnosis

How mental health disorders are currently diagnosed:

Mental health disorders mainly rely on clinical interviews and patient-reported outcomes, rather than visible symptoms like physical disorders. To diagnose a particular mental health disorder, clinicians use various diagnostic frameworks, such as DSM-5 and ICD-11,⁹ which are a few examples of these frameworks, as they assess symptoms based on how severe and how long a symptom has been present. Various questionnaires, such as PHQ-9 and GAD-7, also help clinicians in diagnosing mental health disorders. The questionnaires are primarily used in identifying the severity, too. Unlike physical illnesses, mental health diseases do not usually have objective biological markers, which makes the diagnosis much more dependent on the interpretation and judgment of the doctor.

Primary challenges in diagnosing hard-to-diagnose diseases:

Many mental health disorders are often considered complex/hard to diagnose disorders; this is basically because these kinds of disorders share common symptoms and they overlap

with each other. Some examples of overlapping symptoms are, for eg, fatigue, sleep disturbance, distress, etc. Mental health disorders are expressed through behavior, speech, and expressions, but each of them is subjective as it varies from culture to culture.¹⁰

Limitations of human diagnosis:

Clinical diagnosis faces several systemic challenges. First, assessments are time-intensive and require sustained interaction. Second, there is a global shortage of trained professionals, particularly in resource-constrained settings. Third, financial barriers and social stigma reduce accessibility. These limitations highlight the need for scalable systems that can assist clinicians in early-stage screening.

How can AI be incorporated in the diagnosis of mental health?:

Artificial intelligence offers amazing tools for analyzing mental health problems. These tools include the use of automated, continuous, and consistent processing of large volumes of text provided by the patient.¹¹ Artificial intelligence for mental health and mental illnesses: an overview." Recent advances in machine learning and natural language processing have enabled AI models to analyze written and spoken language. AI models can analyze textual data such as clinical notes, questionnaires, and patient-reported outcomes. AI-powered chatbots can offer non-stop support to a larger demographic area at a reduced cost.¹²

■ AI-integrated mental health diagnosis:

Existing works in diagnosing mental health with the help of AI:

Currently, clinical notes, self-report questionnaires, online forums, and chatbot conversations have been widely used to identify mental health signals using machine learning and NLP models. It can be used in screening for depression, detecting suicide risks, and monitoring anxiety. Such systems usually formulate mental health screening as a task of supervised classification with labeled datasets.¹³

Types of input data for AI screening:

Clinical notes: contain observations recorded by a doctor during the interview. NLP models can extract useful information (symptom indicators, diagnostic cues) and use it to diagnose the disorder.⁸

Self-report questionnaires: This type of input data includes standardized symptom measures such as PHQ-9, and these are widely used for the process of training an AI.

Online text and social media: online posts from blogs and social media platforms capture a variety of expressions of distress and emotional states. These data are large-scale and reflect natural language use but are often context-dependent. Linguistic ambiguity and population bias are major challenges with this input type currently.

Model Architectures Used in Prior Studies:

Traditional machine learning models:

The most common traditional methods to analyze mental health data involve manually creating natural language processing based on simple pre-defined linguistic characteristics, such as term frequency (TF-IDF), word sentiment, or syntax (parts of speech). Traditional statistical methods that have been used in machine learning models include logistic regression, SVMs (Support Vector Machines), and Random Forests. However, their reliance on predefined features limits their ability to capture contextual meaning. In cases of textual ambiguity—where symptoms are expressed indirectly or overlap across multiple conditions—these models fail to distinguish between nuanced expressions. For example, phrases like “I feel off” or “something isn’t right” lack explicit markers and are often misclassified or ignored entirely. As a result, traditional models are effective only when symptoms are explicitly stated, but they struggle significantly with subjective and context-dependent language.¹⁴ This workflow is summarized in Figure 1.

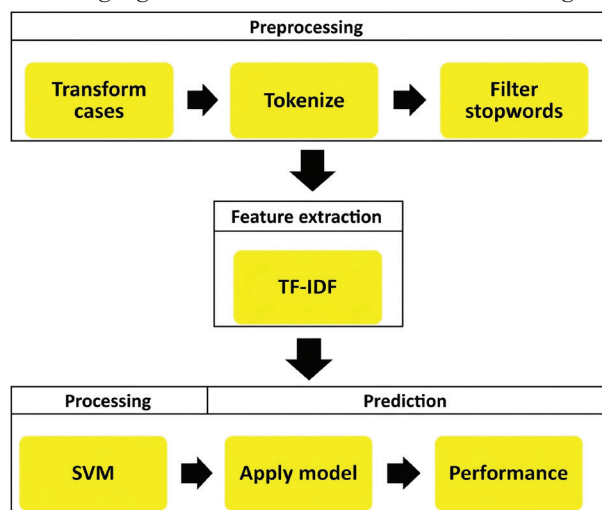


Figure 1: Figure 1 shows a traditional machine learning pipeline for mental health screening. It moves from preprocessing and TF-IDF feature extraction to an SVM model and performance evaluation.

Deep learning models:

Deep learning models, including convolutional neural networks, learn representations directly from raw text. They can capture local patterns such as key symptom phrases, improving performance over traditional methods. However, these models remain sensitive to variability in data and often fail to generalize when linguistic patterns deviate from training distributions, reducing their effectiveness in real-world scenarios.¹⁵ A simplified version of this process is presented in Figure 2.

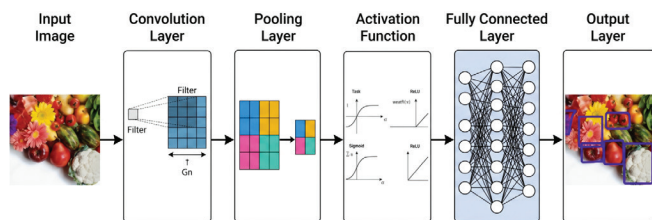


Figure 2: Figure 2 shows a deep learning classification architecture. It demonstrates how an input is processed through convolution, pooling, activation, and fully connected layers before producing an output prediction.

Transformer-Based Models (BERT, GPT, etc., for example):

Transformer-based models, such as BERT, represent the current state of the art in NLP. These models capture contextual relationships across entire text sequences using attention mechanisms.

Despite their advantages, transformers are not inherently equipped to handle ambiguity. Their performance depends heavily on training data quality. When inputs contain conflicting or incomplete symptom descriptions, model predictions can become unreliable.

For example, in cases where symptoms overlap between disorders (e.g., anxiety and PTSD), transformers may assign a single dominant label rather than reflecting uncertainty or

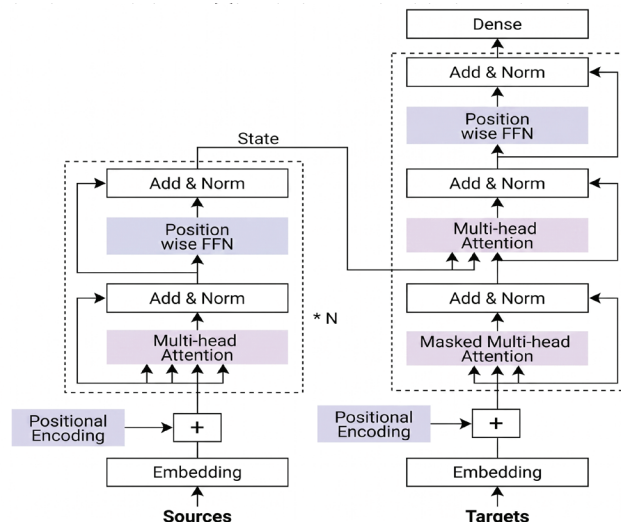


Figure 3: Figure 3 shows the basic transformer architecture, including embeddings, positional encoding, multi-head attention, feed-forward layers, normalization, and dense output. This helps explain why transformer models can capture context but may still struggle when patient language is incomplete or ambiguous.

Evaluation metrics used in previous studies

The majority of the studies test AI systems in terms of conventional performance measures like accuracy, precision, recall, F1 score, and PR-AUC. Although these measures are used to measure technical performance, they are not useful in measuring usability or trust. The recent studies focus on human-centered assessment with measures on a Likert scale for clarity, usefulness, appropriateness, and reliability. It is becoming evident in the literature that high accuracy is not the end goal for clinical deployment. The systems should also be able to express uncertainty and match the expectations of the user, especially in cases of ambiguous disorders.

Future workss

Improved model architectures for complex disorders:

Future studies need to develop models of different architectures to take into consideration the ambiguities in language used to describe the symptoms and the overlap in symptoms of disorders, and not treat the process of screening for mental health as a typical classification problem. The majority of today's systems are optimized for disorders that have clear and

standardized lists of symptoms and therefore fail miserably when attempting to classify disorders based on ambiguous, indirect, or heterogeneous language. Although transformer-based architectures have shown superior understanding capabilities, they also continue to have difficulties in interpreting culturally based symptom narratives. Therefore, future architectures that will be able to screen for complex disorders such as PTSD and depression subtypes must integrate representation learning with uncertainty modeling and structured symptom reasoning.

Multiple dataset integration:

While this review does focus mainly on text-based screening, future systems could benefit a lot from different data integration. When we combine patient language with physiological or behavioral signals, such as sleep patterns and heart rate, the variability may provide additional context that can help the AI system by reducing diagnostic ambiguity. Multimodal approaches have been suggested as a strong way to strengthen AI screening systems by compensating for the limitations of subjective language alone. Such integration should be carefully designed to respect the patient's privacy and ensure that non-textual signals are meaningful.

Human + AI collaboration in Mental health disorder diagnosis:

Future systems should adopt a human-AI collaborative model rather than fully automated diagnosis. In this framework, AI provides screening, risk scales, and symptom summaries, while clinicians will be responsible for final diagnostic decisions. Studies evaluating mental health chatbots show that user trust and perceived usefulness depend strongly on clarity, appropriateness, and safety rather than technical accuracy alone. A collaborative framework ensures that AI augments clinician judgment instead of replacing it and allows ambiguous cases to be escalated to professional evaluation.¹⁷ The collaborative workflow proposed for this issue is shown in Figure 4.

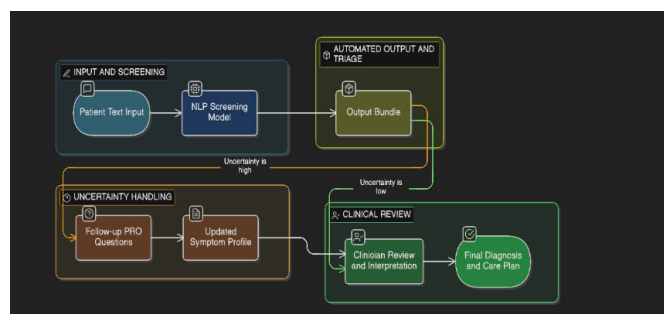


Figure 4: Figure 4 shows the proposed human-AI diagnostic workflow. Patient text is screened by an NLP model, uncertainty triggers follow-up PRO questions, and high-risk or unclear cases are routed to clinician review.

■ Proposed AI System Architecture

To address the limitations identified, this research proposes a conceptual AI architecture specifically engineered for the diagnosis of textually ambiguous mental health disorders.

The system consists of the following integrated components:

- Refined text input processing

- Contextual embedding utilizing transformer-based models
- Specialized ambiguity detection module
- Multi-label prediction layer
- Uncertainty estimation and probabilistic reasoning
- Human-AI interaction interface

Core Innovation:

The primary innovation of this architecture lies in its explicit modeling of linguistic ambiguity. Instead of dismissing vague patient inputs as noise, the system actively detects and quantifies uncertainty.

Multi-Label Prediction:

The framework generates multiple potential diagnoses rather than just a singular label. It mirrors real-world clinical reasoning in cases where symptoms significantly overlap across different mental health conditions.

Uncertainty Estimation:

Detailed confidence scores are produced for each diagnostic prediction, which allows the system to transparently communicate the degree of reliability associated with its outputs.

Human-AI Collaboration:

Diagnostic insights are presented to clinicians alongside confidence levels and ambiguity indicators, and they ensure that final interpretive decisions remain under professional human supervision. The final architecture is summarized visually in Figure 5.

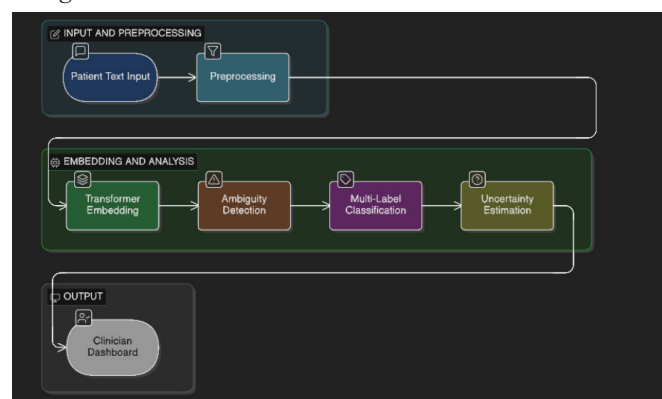


Figure 5: Figure 5 shows the proposed uncertainty-aware AI system architecture. It combines text preprocessing, transformer embeddings, ambiguity detection, multi-label classification, uncertainty estimation, and clinician-facing output.

■ Conclusion

This paper addresses the question of how AI systems can be strengthened to aid the screening of textually ambiguous mental health disorders using varying patient-reported outcomes and argues that this progress requires moving beyond accuracy-focused classification toward uncertainty-aware, practical, and much more human-centered design (one that can excel in understanding contexts). By using prior work across input data types, model architectures, and evaluation frameworks, this review highlights the gap between successful applications in

common disorders and the ongoing challenges associated with complex conditions such as PTSD and depression subtypes. The proposed focus on combining traditional performance metrics with a Likert-scale human evaluation aligns with the emerging healthcare AI frameworks that prioritize safety, transparency, and clinical collaboration. Ultimately, responsible deployment of AI in mental health care depends not on replacing clinicians, but on diagnostic reasoning and expanding access to early screening in a manner that acknowledges linguistic ambiguity and patient subjectivity. This research is limited by its methodological reliance on secondary literature synthesis and illustrative experimentation. Future work in this domain should prioritize validation utilizing diverse real-world datasets and rigorous clinical evaluation to ensure the practical efficiency of diagnostic frameworks.

■ Acknowledgments

I want to acknowledge my mentor, Dr Siddharth Krishnan, and my TA, Samuel Lefcourt.

■ References

1. Mental disorders. <https://www.who.int/news-room/fact-sheets/detail/mental-disorders> (accessed 2026-02-20).
2. Mental Health ATLAS 2020. <https://www.who.int/publications/i/item/9789240036703> (accessed 2026-02-20).
3. D'Alfonso, S. AI in Mental Health. *Curr. Opin. Psychol.* 2020, 36, 112–117. <https://doi.org/10.1016/j.copsyc.2020.04.005>.
4. Is AI the Future of Mental Healthcare? | Topoi | Springer Nature Link. <https://link.springer.com/article/10.1007/s11245-023-09932-3> (accessed 2026-02-20).
5. Artificial Intelligence for Chatbots in Mental Health: Opportunities and Challenges | Springer Nature Link. https://link.springer.com/chapter/10.1007/978-3-030-67303-1_10 (accessed 2026-02-20).
6. Schnittker, J. *The Diagnostic System: Why the Classification of Psychiatric Disorders Is Necessary, Difficult, and Never Settled*; Columbia University Press, 2017.
7. The Rise of Artificial Intelligence in Healthcare Applications. In *Artificial Intelligence in Healthcare*, Academic Press, 2020; pp 25–60. <https://doi.org/10.1016/B978-0-12-818438-7.00002-2>.
8. Natural language processing applied to mental illness detection: a narrative review | npj Digital Medicine. <https://www.nature.com/articles/s41746-022-00589-7> (accessed 2026-02-21).
9. Regier, D. A.; Kuhl, E. A.; Kupfer, D. J. The DSM-5: Classification and Criteria Changes. *World Psychiatry* 2013, 12 (2), 92–98. <https://doi.org/10.1002/wps.20050>.
10. Liu, J.; Wang, H.; Li, L. Rethinking the Studies of Diagnostic Biomarkers for Mental Disorders. *Meta-Radiol.* 2025, 3 (1), 100135. <https://doi.org/10.1016/j.metrad.2025.100135>.
11. Artificial Intelligence for Mental Health and Mental Illnesses: an Overview | Current Psychiatry Reports | Springer Nature Link. <https://link.springer.com/article/10.1007/S11920-019-1094-0> (accessed 2026-02-21).
12. Arji, G.; Erfannia, L.; Alirezai, S.; Hemmat, M. A Systematic Literature Review and Analysis of Deep Learning Algorithms in Mental Disorders. *Inform. Med. Unlocked* 2023, 40, 101284. <https://doi.org/10.1016/j.imu.2023.101284>.
13. Analysis of depression in social media texts through the Patient Health Questionnaire-9 and natural language processing - Nam Hyeok Kim, Ji Min Kim, Da Mi Park, Su Ryeon Ji, Jong Woo Kim, 2022. <https://journals.sagepub.com/doi/full/10.1177/20552076221114204> (accessed 2026-02-21).
14. A novel text mining approach based on TF-IDF and Support Vector Machine for news classification | IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/abstract/document/7569223> (accessed 2026-02-21).
15. 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. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *J. Big Data* 2021, 8 (1), 53. <https://doi.org/10.1186/s40537-021-00444-8>.
16. End-to-End Transformer-Based Models in Textual-Based NLP. <https://www.mdpi.com/2673-2688/4/1/4> (accessed 2026-02-21).
17. JMIR mHealth and uHealth - Validity of Chatbot Use for Mental Health Assessment: Experimental Study. <https://mhealth.jmir.org/2022/10/e28082> (accessed 2026-02-21).

■ Author

Dhairya Jain is a Grade 11 student at Jayshree Periwal Global School, Jaipur. He plans to pursue Computer Science, with interests in how artificial intelligence and technology can be used to solve real-world problems. He actively engages in research, Model United Nations, and drama, combining technical curiosity with leadership and communication.