# Clinician-annotated Social Media Corpus for Artificial Intelligence Based Mental Health Screening

Anusha M. D. [1] and Parameshwar R. Hegde [1,2,*]

[1] Department of Computer Science, Yenepoya Institute of Arts, Science, Commerce and Management, Yenepoya (Deemed to be University), Mangaluru, India
[2] Goran Apps, Yenepoya Technology Incubator, Mangaluru, India
Email: anugowda251@gmail.com (A.M.D.); param1000@yahoo.com (P.R.H.)
*Corresponding author

***Abstract*—The automated classification of mental health disorders from social media text holds significant promise for scalable early screening and clinical decision-making. However, progress in this domain remains constrained by the lack of clinically annotated corpora for training purposes. Existing datasets predominantly rely on crowdsourced, automated, or semi-supervised annotation strategies that lack the clinical credibility required for the reliable classification of disorders. We present International Classification of Diseases, 11th revision (MANAS-ICD11), a Reddit-based benchmark corpus of 3005 posts exclusively annotated by medical professionals, comprising one psychiatrist and two clinical psychologists, across five diagnostic categories aligned with the ICD-11. A novel three-task annotation design capturing disorder-specific keywords, clinical characteristics, and diagnostic labels achieved a strong Inter-annotator Agreement (IAA) with a Krippendorff's alpha value of 0.80. Six Artificial Intelligence (AI) models spanning Machine Learning (ML), Deep Learning (DL), and Transfer Learning (TL) architectures were benchmarked, with the Support Vector Machine (SVM) achieving the strongest performance on the held-out test set at 0.83 accuracy and a weighted F1-Score of 0.81, outperforming all DL and transformer-based models, including MentalBERT. Cross-validation additionally showed that Logistic Regression (LR) was a statistically more consistent performer across folds. The dominance of Term Frequency-Inverse Document Frequency (TF-IDF)-based classifiers highlights the discriminative richness of ICD-11 aligned clinical lexical patterns in the corpus. MANAS-ICD11 provides a clinically grounded and reproducible foundation for automated mental health screening with direct applicability in clinical and public health settings.**

***Keywords*—clinical annotation, disorder classification, International Classification of Diseases, ICD-11, mental health Natural Language Processing (NLP), MentalBERT, Reddit corpus**

## I. Introduction

Social media platforms have become key spaces where individuals openly express their emotions, personal troubles, and mental health issues without fear of consequences [1]. People tend to share mental health experiences online because of the stigma surrounding these conditions, low levels of mental health literacy, lack of healthcare access, and lack of willingness to seek professional assistance [2]. This is exacerbated by the global mental health crisis affecting the world [1]. According to Garg *et al.* [3], the World Health Organization (WHO) reports over 0.8 million suicides per year. The mental healthcare burden is equally alarming; by August 2021, England alone had 1.6 million people awaiting mental health services, while approximately 8 million more were denied specialist care on the grounds that their condition was not deemed severe enough [3]. Studies have further shown that 80% of people at risk avoid disclosing their distress in face-to-face clinical settings [3], making social media a critical and largely untapped resource for the early detection of mental health issues. Platforms such as Reddit support anonymous and long-form discussions, where users openly describe experiences of distress, intrusive thoughts, compulsive behaviours, and stress without fear of judgement [4]. These posts contain valuable linguistic cues for underlying mental health signs [4], and therefore offer an opportunity for scalable and cost-effective mental health screening that could support the early detection and automated analysis of mental health problems. Inspired by this potential, researchers have proposed social media-based datasets to systematically analyse mental health conditions using computational methods [3, 5]. Datasets such as Self-reported Mental Health Diagnoses (SMHD) [6], Dreaddit [7], Causal Analysis of Mental health issues in Social media (CAMS) [8], Low Self-Esteem (LoST) [9], and ReDSM5 [10] have facilitated large-scale analyses of mental health discourse using Machine Learning (ML) and Deep Learning (DL) techniques.

Although these resources have contributed meaningfully to progress in mental health classification, their diagnostic reliability and clinical credibility are limited because of labelling techniques, such as

Manuscript received June 17, 2026; revised July 7, 2026; accepted August 3, 2026; published November 12, 2026.

self-reported or crowdsourced annotations. Apart from annotation concerns, the majority of existing datasets are limited to binary classification tasks, most commonly distinguishing depression from control groups. This restricts the generalisability of the findings to the broader spectrum of clinical mental health conditions [3, 11].

These constraints emphasise the importance of expert-annotated datasets that span various disorders and include relevant clinical features [6]. To make accurate mental health predictions, a body of text must be provided with a clinical diagnosis rather than pattern matching or crowdsourced annotations. In addition, because the International Classification of Diseases, 11th revision (ICD-11) classification system [12] is a recognised global diagnostic tool developed by the WHO, it has clinical rigor and cross-national applicability. In contrast, current data are mostly platform-specific or based on the diagnostic categories of the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5), which is predominantly used in the United States, limiting their cross-cultural applications [13]. To address these gaps, this study proposes the MANAS-ICD11 corpus. The key findings of this study are as follows.

- We developed the MANAS-ICD11 corpus, a dataset based on Reddit that aligns with ICD-11 standards. It covers five diagnostic categories and was annotated by three medical professionals, achieving a Krippendorff's alpha value of 0.80, reflecting a strong Inter-annotator Agreement (IAA).
- We designed a three-task annotation framework that comprises keyword identification, clinical characteristic documentation, and diagnostic labelling to support detailed analysis.
- We evaluated six models, including traditional ML, DL, and transformer-based methods. This provides clear and reproducible baseline results for the classification of mental health disorders.

## II. Literature Review

Recently, many studies have been conducted on the extraction of mental health signals from social media platforms, partly because users tend to communicate their psychological distress more openly online than in clinical settings [14], The resulting datasets, computational methods used, and gaps in the current study are discussed in this section. One of the earliest and most widely cited resources is the SMHD, authored by Cohan *et al.* [6], which gathered Reddit posts from members who openly shared diagnoses of nine mental health conditions. The primary advantage of this corpus is its large size. However, because the labels were based solely on self-reported statements rather than clinical assessments, the classification performance was modest. The best-performing models were Logistic Regression (LR), Support Vector Machine (SVM), XGBoost, and Convolutional Neural Network (CNN), with the highest F1-Score of 0.55. Dreaddit [7], on the other hand, is a Reddit corpus for binary stress detection, where the annotation is obtained through crowdsourcing. Transformer models performed significantly better on this dataset, with an F1-Score of 0.81. However, the single-label use and lack of clinical validation affect the applicability of these results in real clinical diagnostic scenarios.

Subsequent researchers, acknowledging the limitations inherent in crowdsourced annotation, have increasingly relied on domain expertise to enhance the rigor and clinical credibility of their labelling methodologies [8, 9]. Garg *et al.* [9] introduced CAMS, an expert-informed corpus designed for causal analysis of mental health expressions. In this study, models such as LR, Support Vector Machine (SVM), CNN, and a hybrid CNN-Long Short-Term Memory (LSTM) achieved a moderate F1-Score of 0.63, constrained by the inherent complexity of modelling causal relationships. Garg *et al.* [9] also proposed LoST, a Reddit corpus for detecting low self-esteem, annotated by a clinical psychologist and a social Natural Language Processing (NLP) researcher. Models including Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional Encoder Representations from Transformers (BERT), RoBERTa, and XLNet achieved an accuracy of 0.88 on this corpus, affirming the benefits of expert annotation for sequential and contextual architectures.

Garg *et al.* [15] also proposed WellXplain, an expert-curated corpus for wellness dimensions, with 78% accuracy using transformer models, although it has narrow applicability for disorder classification, given its limited scope of wellness dimensions. Bao *et al.* [10] created ReDSM5, a Reddit corpus based on DSM-5 criteria for depression detection, on which transformers and Large Language Models (LLMs) scored 0.49 macro F1, highlighting the clinical ambiguity of labelling text-based information for a single disease [16]. To address the challenge of scalability, Raihan *et al.* [17] created MentalHelp, a semi-supervised dataset comprising 14 million Reddit instances labelled using an ensemble of language models. Domain-adapted transformer models achieved an average F1-Score of 0.83 on this dataset; however, the lack of clinical supervision in the labelling process leaves the diagnostic accuracy uncertain, a gap that scale alone cannot close [18].

The corpus construction methodology has also received increasing attention in recent studies. Rani *et al.* [19] built a large Reddit corpus spanning the pre-pandemic, pandemic, and post-pandemic periods, with a manually annotated subset of 800 posts labelled for underlying root causes such as trauma, personality factors, and early life environment. Their annotation process combined keyword-based filtering with a whole-post review, guided by a consultant psychiatrist, and reached a Cohen's Kappa value of 0.869 through a consensus-based resolution protocol. This bears some resemblance to the annotation methodology adopted for MANAS-ICD11, though Rani *et al.* [19] focused on root-cause attribution rather than ICD-11 aligned diagnostic classification, and their annotation team, unlike ours, did not consist exclusively of clinical experts.

Research extending beyond Reddit has further clarified the potential and limitations of social media-based

mental health classification. Early work by Coppersmith *et al.* [20] demonstrated that linguistic signals correlating with mental health conditions could be quantified from Twitter data, laying the groundwork for later shared-task benchmarks such as CLPsych 2015 [21]. Merayo *et al.* [22] created an Instagram collection of celebrity mental health disclosures and used machine learning methods to measure the emotional reactions they caused, finding that transformer-based emotion classifiers outperformed traditional classifiers.

The same authors [23] later investigated the polarity of Instagram mental health posts and patterns of stigma, achieving up to 96% accuracy with RoBERTuito and approximately 80% with hybrid CNN-LSTM models. Together, these findings suggest that clinically relevant linguistic patterns in mental health discourse are not platform-specific but generalise across social media contexts [22, 23]. Annotation quality is equally important in resource-constrained linguistic settings. Sarcasm detection in Kannada using ML on a manually annotated dataset [24] showed that rigorous expert annotation can yield reliable classification performance even when the corpus scale is limited, a principle that directly shaped the annotation methodology used in this study. This emphasis on architectural comparison has recently been extended to large-scale mental health benchmarks. Hasan *et al.* [25] evaluated five transformer architectures (BERT, RoBERTa, DistilBERT, ALBERT (A Lite BERT for Self-supervised Learning of Language Representations), and ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)) against LSTM variants on a seven-class Reddit corpus spanning six mental health conditions and a control group, finding that transformer models, particularly RoBERTa, consistently out-performed LSTM-based alternatives. Ovcharuk *et al.* [26] took a different approach, proposing a transformer-based ensemble of binary classifiers built on fine-tuned DistilBERT models for multi-label mental health detection. By decomposing the multi-class diagnostic tasks into independent binary problems, they found that inter-class confusion relative to the standard multi-class baselines was reduced. Both studies reinforce the value of systematic architectural comparison, although neither incorporates ICD-11-aligned labels nor clinician-annotated diagnoses comparable to the MANAS-ICD11 corpus.

Clinician-involved annotation has continued to expand more broadly in the field of mental health NLP. Holderness *et al.* [27] developed a gold-standard corpus of psychiatric electronic health records annotated by a team of clinical experts for trauma-related events and clinical information, achieving a high IAA across entity and relational tags. Lamparth *et al.* [28] introduced MENtal health Tasks AssessmenT (MENTAT), a clinician-annotated benchmark of real-world psychiatric tasks spanning diagnosis, treatment, monitoring, triage, and documentation, built to capture the ambiguity of everyday clinical decision-making rather than exam-style factual recall. Alhuzali *et al.* [29] proposed MentalQA, an Arabic corpus of 500 posts annotated by three expert annotators for question types and answering strategies in mental healthcare, achieving Fleiss' Kappa of 0.61 and 0.98, respectively. However, its focus on Q&A interaction typing rather than diagnostic classification, combined with its Arabic-language, non-ICD-11 scope, limits direct comparability to MANAS-ICD11. Most recently, Fouda *et al.* [30] introduced PsychiatryBench, a benchmark of 5188 question-answer items spanning 11 psychiatric reasoning tasks, curated and annotated exclusively by licenced psychiatrists from authoritative psychiatric textbooks and casebooks rather than social media or synthetic dialogue. These efforts reinforce the continued value of clinically grounded annotation frameworks, although none adopts an ICD-11 aligned multi-class diagnostic schema for social media text comparable to the present study.

TABLE I. COMPARISON OF MANAS-ICD11 WITH EXISTING MENTAL HEALTH DATASETS BY SOURCE, ANNOTATOR TYPE, ANNOTATION TASKS, AND INTER-ANNOTATOR AGREEMENT

| Corpus | Source | Size | Class | Diagnostic Framework | Annotator Type | Annotation Tasks | Agreement Metric | Agreement Score |
|---|---|---|---|---|---|---|---|---|
| SMHD [6] | Reddit | 900K+ | 9 | DSM-5 | Automated (pattern-matching) | Label only | Not reported | - |
| Dreaddit [7] | Reddit | 3553 | 2 | None | Crowdsourced | Label only | Krippendorff's α | 0.67 |
| CAMS [8] | Reddit | 5051 | 6 | DSM-5 | NLP experts | Label + Cause | Cohen's Kappa | 0.49 |
| LoST [9] | Reddit | 3251 | 2 | None | Psychologist + NLP researcher | Label only | Cohen's Kappa | 0.72 |
| ReDSM5 [10] | Reddit | - | 2 | DSM-5 | Automated | Label only | Not reported | - |
| WellXplain [15] | Reddit | 3092 | 6 | None | Psychologist + NLP researcher | Label + Wellness dim. | Cohen's Kappa | 0.71 |
| MentalHelp [17] | Reddit + Twitter | 14M | Multi-label | None | Semi-supervised (LLM ensemble) | Label only | Not reported | - |
| CLPsych 2015 [21] | Twitter | 1746 | 2 | None | Single human verifier | Self-reported + Label verification | Not reported | - |
| Rani *et al.* [19] | Reddit | 800 (annotated) | 4 | None | 2 annotators + psychiatrist consult | Label only | Cohen's Kappa | 0.87 |
| Hasan *et al.* [25] | Reddit | 105,000 | 7 | None | Automated + author validation | Label only | Cohen's κ / Krippendorff's α | 0.92 / 0.91 |
| MentalQA [29] | Arabic forums | 500 | N/A (Q&A) | None | 3 expert annotators | Q-type + strategy | Fleiss' Kappa | 0.61 / 0.98 |
| MANAS-ICD11 (Ours) | **Reddit** | **3005** | **5** | **ICD-11** | **Psychiatrist + 2 Psychologists** | **Keywords + Clinical Characteristics + Label** | **Krippendorff's α** | **0.80** |

Across all studies, a recurring pattern has been observed in which large-scale datasets derived from automated or crowdsourced annotations lack clinical credibility, whereas expert-annotated corpora remain restricted in both diagnostic scope and label granularity [6–10, 16]. While classification models applied to these corpora perform well on binary tasks, they often struggle with fine-grained multi-class disorder classification when corpus quality is limited. Compared with the existing mental health corpora reviewed in this study (Table I), MANAS-ICD11 combines clinician annotations by medical professionals, ICD-11-aligned multi-class labels, and a three-task annotation framework comprising disorder-specific keywords, clinical characteristics, and diagnostic labels. The corpus was designed to address the limitations of existing datasets.

## III. Materials and Methods

### A. The MANAS-ICD11 Corpus

The term "MANAS" originates from the Sanskrit word for "mind", indicating the dataset's focus on clinical aspects. The dataset is annotated in accordance with ICD-11, the diagnostic framework established by the WHO. This study is the principal contributor to the MANAS-ICD11 corpus. It was developed using a sequential pipeline that included selecting subreddits, scraping data, and manually filtering content. The dataset was labelled by domain experts focused on ICD-11, and the appropriate clinical keywords and characteristics were systematically identified. For each stage of the work, a careful description was provided to ensure the reproducibility of the work and to facilitate subsequent research on this corpus.

#### 1) Data collection and filtering

Reddit is one of the most popular online social platforms, with thousands of different communities of interest called "subreddits" [31]. Users on Reddit frequently share personal experiences, seek peer support, and discuss symptoms and coping strategies related to mental health. The platform provides a rich and ecologically valid mental health discourse, due to the conditions it offers [21]. Data collection was conducted using the Python Reddit Application Programming Interface (API) Wrapper (PRAW), an official Reddit API client that facilitates programmatic access to posts and metadata in accordance with Reddit's API Terms of Service. Posts were gathered from disorder-specific and general mental health subreddits covering the period from January 2019 to December 2024. The dataset was compiled from posts retrieved from subreddits such as depression, anxiety, bipolar disorder, Obsessive-Compulsive Disorder (OCD), Post-Traumatic Stress Disorder (PTSD), Attention-Deficit/Hyperactivity Disorder (ADHD), autism, Eating Disorders, Dissociative Identity Disorder (DID), Phobia, Panic Attack, social anxiety, and mental health.

For each post, the extracted metadata included the title, body text (selftext), subreddit name, author identifier, Coordinated Universal Time (UTC) timestamp of creation, post score, and the number of comments. Posts were filtered based on the following inclusion and exclusion criteria: posts had to share personal mental health stories in more than ten words to qualify. We removed those that were non-English, had removed body content, contained 10 or fewer words, were duplicates, or were unrelated to mental health, such as announcements, memes, or promotional content. After filtering, 3005 posts were retained for the expert analysis.

#### 2) Label schema

The source subreddits were mapped to five broader disorder categories consistent with the ICD-11 diagnostic groupings [12]. This consolidation was performed to reduce label sparsities and align the classification tasks with clinically meaningful categories. The five categories are defined as follows:

(1) Anxiety and Fear-Related Disorders (anxiety/fear): Conditions characterised by excessive fear or anxiety, often accompanied by avoidance behaviour.
(2) Obsessive-Compulsive and Related Disorders (OCRD): Disorders characterized by persistent obsessions or compulsions that cause significant distress or functional impairment.
(3) Disorders Specifically Associated with Stress (Stress-Related): Mental health conditions arising from exposure to stressful or traumatic experiences.
(4) Mood Disorders: Disorders characterized by a persistent disturbance in mood, such as prolonged sadness or loss of interest in activities.
(5) "Other": Posts that do not clearly fall into any of the above-mentioned diagnostic categories.

#### 3) Annotation process and consolidation

The annotation process was conducted by a panel of three mental health experts comprising one psychiatrist and two clinical psychologists with extensive experience in the diagnosis and treatment of mental health disorders. Their participation followed the established annotation practices widely adopted in mental health NLP corpora [9, 32]. To facilitate systematic annotation, a corpus of 3005 Reddit posts was divided into batches of 500 posts. Each batch was assigned to the experts, and subsequent batches were distributed only after the completion of the previous batch to maintain annotation consistency throughout the process. Each post was independently reviewed by all three experts, ensuring that every sample received three independent annotations [33]. For every post, the annotators were instructed to:

- Identify clinically relevant keywords in text of post,
- Specify the clinical characteristics supporting their diagnostic reasoning,
- Assign one final diagnostic label from the five ICD-11-aligned categories described in Section III.A.2.

A representative example of this three-task annotation process is presented in Table II. To prevent annotation bias, each expert received an individual spreadsheet containing only their assigned batch and completed the annotation independently without access to the responses of other annotators, thereby ensuring blind annotation.

After the annotation phase, the outputs of all three experts were systematically consolidated. The identified keywords were normalised and deduplicated using Python set-based operations, whereas the clinical characteristics were merged and standardised using a custom preprocessing function that handled missing values, converted text to lowercase, removed extraneous whitespace, and segmented the text into distinct clinical phrases. The final diagnostic labels were determined using a majority voting strategy, whereby the label assigned by at least two of the three experts was selected as the gold standard annotation, thereby improving consistency and minimizing individual annotator bias [33]. In cases where all three experts assigned different diagnostic labels and no majority agreement was reached, the corresponding posts were jointly reviewed through discussion until a consensus was reached.

TABLE II. REPRESENTATIVE ANNOTATION EXAMPLE FROM THE MANAS-ICD11 CORPUS ILLUSTRATING THE THREE-TASK ANNOTATION PROCESS AND MAJORITY VOTING STRATEGY

| Annotation Component | Annotator 1 | Annotator 2 | Annotator 3 |
|---|---|---|---|
| Keywords | panic attack, derealization, depersonalization, hyperventilation, intrusive thoughts | out-of-body sensations, distorted reality, negative thoughts, shaking, fear of recurrence | panic attack, derealization, substance-triggered, breathlessness, anticipatory anxiety |
| Clinical Characteristics | Acute panic episode with dissociative features (derealization and depersonalization) triggered by substance use, accompanied by somatic symptoms and persistent intrusive cognitions | Severe anxiety episode with perceptual disturbances, loss of reality testing, prolonged recovery period, and emerging fear of recurrence consistent with panic disorder | Substance-induced panic attack with out-of-body sensations, anticipatory anxiety the following day, and somatic hyperarousal suggesting anxiety and fear-related presentation |
| Label | Anxiety & Fear-Related Disorders | Anxiety & Fear-Related Disorders | Anxiety & Fear-Related Disorders |

Note: Sample Post: "I had a severe panic attack after smoking and felt 'out of reality'... everything around me felt distorted and I couldn't process what was happening... I felt like I jumped out of my body and couldn't trust anyone... I had intrusive thoughts, imagined different realities, and wondered if I was still me... I experienced shaking, breathlessness, and fear of dying... it took 5–6 h to calm down... even the next day I felt scared it might happen again, with heart fluttering and exhaustion".

*4) Inter-Annotator Agreement (IAA)*

The reliability of the final diagnostic label was assessed using Krippendorff's alpha, computed with the Python krippendorff package. Standard inter-annotator agreement coefficients, such as Krippendorff's alpha, are computed over a fixed, discrete response space in which annotators select from a common set of categories, with agreement defined as the extent to which independent raters converge on the same category [32]. Among the three annotation tasks, only the diagnostic label satisfied this requirement, as the labels were selected from the five predefined ICD-11-aligned categories. By contrast, the keyword and clinical characteristic tasks were open-ended free-text responses with no fixed category set, allowing annotators to describe clinically equivalent content in different wording. Because there is no common categorical frame against which to compute exact agreement for free text, standard IAA metrics do not meaningfully apply to these two tasks. Instead, they were consolidated through the normalisation procedure described in Section III.A.3, rather than formally scored for agreement.

The final diagnostic labels from the three annotators were arranged into a reliability data matrix and encoded numerically, with the nominal level of measurement specified as appropriate for categorical classification data. This procedure yielded $\alpha = 0.80$, indicating substantial agreement among the expert annotators.

*5) Dataset statistics*

The MANAS-ICD11 corpus consists of 3005 annotated social media posts. The posts have, on average, 10.78 sentences and 20.65 words per sentence, which are rich and descriptive in nature. The corpus was partitioned into a training set of 2404 posts (80%) and a test set of 601 posts (20%) using stratified sampling to preserve the class proportions across both splits [3]. Table III presents the class-wise distribution of posts across the five diagnostic categories.

TABLE III. TRAIN/TEST CLASS DISTRIBUTION OF THE MANAS-ICD11 CORPUS ACROSS FIVE DIAGNOSTIC CATEGORIES

| Diagnostic Category | Train | Test | Total |
|---|---|---|---|
| Anxiety and Fear-Related | 790 | 198 | 988 |
| Mood Disorders | 532 | 133 | 665 |
| OCRD | 512 | 128 | 640 |
| Stress-Related | 436 | 109 | 545 |
| Other | 134 | 33 | 167 |
| **Total** | **2404** | **601** | **3005** |

*6) Ethical considerations*

Data were collected programmatically via the official PRAW API client between January 2019 and December 2024 under the Reddit API Terms of Service applicable during that period. We note that Reddit's Data API Terms have evolved since our collection period, with the Reddit for Researchers program now positioned as the primary avenue for academic access. Our data collection took place under the terms governing the standard API access at the time.

As this study used publicly available secondary data, formal ethical approval was not required for this study. Several steps were taken to protect participants' privacy throughout the study. Usernames, profile identifiers, and any other personally identifiable metadata were removed from each post before annotation began, and no effort was made to trace posts back to real people. The only metadata retained alongside the post text was the subreddit name, UTC timestamp, post score, and comment count. Given the sensitive nature of mental health disclosures, all data were handled with strict confidentiality throughout annotation and analysis.

We evaluated the proposed MANAS-ICD11 corpus using a traditional social media text classification pipeline with three stages of NLP techniques. These involve the use of traditional ML with handcrafted features, DL with word embeddings, and transformer-based fine-tuning (Fig. 1). The multi-tiered evaluation strategy is aligned with previous studies on social media mental health classification [3] and emotion detection from social media data [34]. Such studies have shown that the classical and neural approaches can be benchmarked for a comprehensive assessment of the utility of the corpus and model generalisation. The proposed corpus was benchmarked using six AI models including ML, DL, and TL.Classical ML models, such as SVM [22], LR [20], and Random Forest (RF) [23], have been evaluated using Term Frequency-Inverse Document Frequency (TF-IDF) feature representation [35]. Furthermore, word embeddings were used to train deep learning models such as CNN and LSTM networks. Moreover, a domain-specific transformer model called MentalBERT [36] was fine-tuned using the MANAS-ICD11 corpus. To ensure fairness and reproducibility in the comparative analysis, a standard preprocessing pipeline was used for each model.

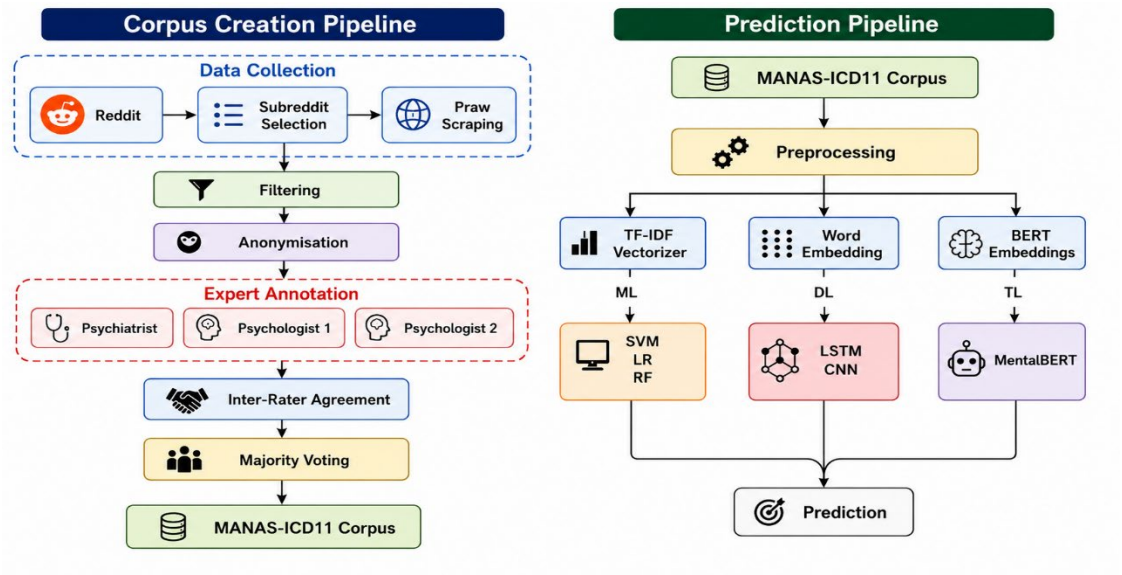



Fig. 1. MANAS-ICD11 corpus creation pipeline, from Reddit data collection through expert annotation and model benchmarking.

### B. Text Preprocessing

Raw Reddit posts contain a lot of noise in the form of URLs, special characters, informal language, and platform-specific markup. All posts were preprocessed using a standard pipeline before features were extracted and models were trained. Each post underwent the following sequential cleaning steps with URLs and hyperlinks were removed using regular expression matching, followed by user mentions and hashtags removal. Newline characters, digits, and punctuation were then removed, and the text was converted to lowercase to reduce vocabulary sparsities. Stopwords were removed using the Natural Language Toolkit (NLTK) English stopword list, and tokens with two characters or fewer were discarded because of their minimal semantic value. Finally, the Snowball Stemmer was applied to normalise morphological variations across the corpus.

### C. Feature Extraction

- TF-IDF for traditional ML models: Text was represented using TF-IDF vectorisation [35]. The vectoriser was configured with a maximum vocabulary of 10,000 features and an n-gram range of (1, 2), enabling the capture of both unigram and bigram representations [35]. To control term distribution, the minimum and maximum document frequencies were set to 2 and 0.95, respectively, while sublinear TF scaling was employed for logarithmic normalisation of term frequencies. The vectoriser was fitted exclusively to the training set and applied to the test set to prevent any data leakage. Formally, the TF-IDF weight for term $t$ in document $d$ is computed as:

$$\text{TF-IDF}(t, d) = \text{tf}(t, d) \times \log(N / \text{df}(t) + 1) \quad (1)$$

where tf($t$, $d$) is the frequency of term $t$ in document $d$, $N$ is the total number of documents, and df ($t$) is the number of documents containing term t [35].

- Word Embeddings for DL Models: For CNN and LSTM, the text was represented using a trainable Keras Embedding layer [34]. Each post was tokenised using a KerasTokenizer with a vocabulary of 20,000 tokens and an <OOV> token for unseen words. The sequences were padded or truncated to a fixed length of 200. Each token was mapped to a dense vector of dimension 300, initialised with random weights, and updated during backpropagation, allowing the model to learn task-specific representations directly from the MANAS-ICD11 corpus. Both the CNN and LSTM models were evaluated under identical configuration settings to ensure comparability.
- BERT tokenisation for MentalBERT: Unlike TF-IDF and word embeddings, MentalBERT processes text through its built-in Word-Piece tokenizer [37], which

was pre-trained on mental health-related text. Each post was tokenised to a maximum length of 128 tokens, with shorter posts being padded and longer posts truncated [38]. The tokenizer produces two outputs for each post input IDs representing each subword token and an attention mask that tells the models which tokens are real and which are padding. This approach preserves the disorder-specific clinical vocabulary that general-purpose tokenizers may incorrectly split into subwords.

Prior to model training, the five diagnostic labels were encoded as integers using LabelEncoder. For CNN and LSTM, these were further converted to one-hot vectors using to_categorical(). MentalBERT received integer labels directly as torch.long tensors.

### D. Machine Learning

- LR is a probabilistic linear classifier that models the posterior probability of each class using the multinomial softmax function for multi-class problems [23]. The model was trained using the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) solver, which is efficient for high-dimensional sparse feature spaces such as TF-IDF representations. A regularisation parameter of $C = 1.0$ was employed with L2 regularisation, and the maximum number of iterations was set to 1000 to ensure convergence.
- The SVM identifies the optimal hyperplane that maximises the margin between classes in a high-dimensional feature space [19]. LinearSVC was employed because it is computationally efficient for large, sparse TF-IDF feature matrices. A regularisation parameter of $C = 1.0$ was used with a maximum of 2000 iterations to ensure convergence.
- RF is an ensemble method that constructs multiple decision trees during training and outputs the majority-voted class [23]. A total of 500 trees were constructed with the square root of the total features considered at each split. Trees are grown to full depth, with a minimum of five samples required to split a node and a minimum of two samples per leaf to prevent overfitting [20].

For all three ML models, class weights were applied proportionally to account for the class imbalance present in the MANAS-ICD11 corpus, ensuring that minority categories contributed adequately to the learning process during training.

### E. Deep Learning

- LSTM captures long-range sequential dependencies in text through gated memory cells, making it well-suited for mental health posts which often contain extended narratives [34]. The architecture consists of an embedding layer with an output dimension of 300, followed by a SpatialDropout1D layer with a rate of 0.30. Two stacked LSTM layers were employed, the first with 128 units and the second with 64 units, each with a dropout rate of 0.30 and a recurrent dropout of 0.20. The output was passed through two dense layers with 128 and 64 units, respectively, with ReLU activation and dropout rates of 0.40 and 0.30. The final softmax layer outputs the predictions over the five diagnostic categories. The model was compiled with the Adam optimiser at a learning rate of $1 \times 10^{-3}$ and categorical cross-entropy loss [39].
- CNN are widely used in text classification tasks because of their ability to capture local syntactic and semantic features from text [29]. The architecture was adapted from Rahib *et al.* [40] and consisted of an embedding layer with an output dimension of 300, followed by a Conv1D layer with 256 filters and a kernel size of 3 with ReLU activation to extract local n-gram features. A GlobalMaxPooling1D layer was applied to extract the most salient feature across the sequence, followed by a dense layer with 256 units and a ReLU activation. A Dropout layer with a rate of 0.20 was applied to reduce overfitting. The final softmax layer outputs the predictions over the five diagnostic categories. The model was compiled with the Adam optimiser at a learning rate of $1 \times 10^{-3}$ and categorical cross-entropy loss.

### F. Transfer Learning

Transfer Learning (TL) refers to the adaptation of a model pre-trained on a large-scale corpus for a specific downstream task through fine-tuning [36]. This strategy is particularly effective when labelled data are scarce because the pre-trained model transfers rich linguistic knowledge and reduces the need for extensive task-specific datasets. In the field of mental health NLP, domain-specific pre-trained models are more likely to yield better results than general-purpose models because they can acquire the vocabulary and linguistic structure of a disorder [36].

- MentalBERT [36] is a domain-specific pre-trained language model based on the BERT architecture, pre-trained exclusively on mental health-related text collected from Reddit and other online mental health communities. Unlike general-purpose BERT models, MentalBERT captures specialized clinical vocabulary, disorder-specific linguistic patterns, and the informal self-disclosure style characteristic of mental health discourse on social media. This domain alignment makes MentalBERT particularly well-suited for the MANAS-ICD11 classification task, where posts span nuanced and overlapping ICD-11 diagnostic categories. For fine-tuning, each post was tokenised with the MentalBERT tokenizer, with a maximum sequence length of 128 tokens, and padding and truncation were applied to ensure uniform input dimensions. A classification head consisting of a linear layer over the [CLS] token representation is added on top of the pre-trained encoder using BertForSequenceClassification. The model was fine-tuned end-to-end for a maximum of 50 epochs with early stopping based on validation loss (patience = 5), a batch size of 16 using the AdamW optimiser [38] at a learning rate of $2 \times 10^{-5}$ and weight decay of 0.01. A linear learning rate scheduler with 10% warm-up steps was applied to stabilise early training, and gradient clipping with a maximum norm

of 1.0 was employed to prevent exploding gradients [37]. Class weights were incorporated into the cross-entropy loss function to address class imbalance, consistent with the other models.

## IV. Experimental Setup

### A. Experimental Setting

All experiments were conducted on Google Colaboratory Pro, equipped with an NVIDIA A100 Graphics Processing Unit (GPU) (40GB HBM2), providing sufficient computational resources for both DL and transformer-based model training. The ML models were implemented using Scikit-learn 1.6.1. DL models were built using TensorFlow 2.19.0 and Keras, while MentalBERT was implemented using the HuggingFace Transformers 5.0.0 library with PyTorch 2.10.0 as the backend framework. All models were trained and evaluated using the same stratified 80:20 train-test split described in Section III-A (5). Stratified sampling was applied to preserve class proportions across both splits. For the DL models, an additional 15% of the training set was held out as a validation set to monitor the training progress and apply early stopping.

In addition to the held-out test evaluation, stratified 5-fold cross-validation was performed for all six models [27]. Given the relatively small corpus size of 3005 posts, Cross-Validation (CV) was employed to ensure robust and reliable performance estimation, minimising the risk of result variability due to a single train-test split. The CV Mean and CV Std are reported for both accuracy and weighted F1-Score across five folds. The weighted Area Under the Curve (AUC) was computed via a one-vs-rest Receiver Operating Characteristic (ROC) analysis to evaluate the probabilistic discriminability across the five diagnostic categories [23].

### B. Evaluation Metrics

The model performance was evaluated using four standard metrics: accuracy, weighted precision, weighted recall, and weighted F1-Score. The weighted F1-Score was adopted as the primary evaluation metric to address the class imbalance in the MANAS-ICD11 corpus, as it reflects the varying support of individual diagnostic categories. These metrics are formally defined as follows:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (2)$$

$$\text{Precision} = TP / (TP + FP) \quad (3)$$

$$\text{Recall} = TP / (TP + FN) \quad (4)$$

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

where TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives, respectively. For multi-class evaluation, weighted averaging was applied across all five diagnostic categories, proportional to their support in the test set.

## V. Result and Discussion

Across all six models, traditional ML classifiers consistently outperformed the DL and transformer-based approaches, with SVM and LR emerging as the top performers. The four primary diagnostic categories were classified reliably, while the residual "Other" category remained the weakest across every model. The class-wise performance of the machine learning and deep learning models is summarised in Tables IV and V, respectively. Statistical significance and error analysis are discussed in the following subsections.

TABLE IV. Class-wise Precision (P), Recall (R), and F1-Score of Machine Learning Models on the Test Set

| Classes | SVM | | | LR | | | RF | | |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| Anxiety and Fear | 0.80 | 0.92 | **0.86** | 0.83 | 0.86 | 0.84 | 0.80 | 0.87 | 0.83 |
| Stress-Related | 0.85 | 0.86 | **0.86** | 0.86 | 0.86 | **0.86** | 0.84 | 0.84 | 0.84 |
| Mood Disorders | 0.85 | 0.83 | 0.84 | 0.84 | 0.86 | **0.85** | 0.81 | 0.86 | 0.84 |
| OCRD | 0.84 | 0.84 | 0.84 | 0.81 | 0.85 | 0.83 | 0.81 | 0.85 | 0.83 |
| Other | 0.40 | 0.06 | 0.10 | 0.30 | 0.18 | 0.22 | 0.00 | 0.00 | 0.00 |
| **Macro Avg** | 0.75 | 0.70 | 0.70 | 0.73 | 0.72 | 0.72 | 0.65 | 0.68 | 0.67 |
| **Weighted Avg** | 0.81 | 0.83 | **0.81** | 0.81 | 0.82 | 0.81 | 0.77 | 0.81 | 0.79 |

TABLE V. Class-wise Precision(P), Recall (R), and F1-Score(F1) of Deep Learning and Transformer Models on Test Set

| Classes | CNN | | | LSTM | | | MentalBERT | | |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| Anxiety and Fear | 0.79 | 0.88 | 0.83 | 0.79 | 0.84 | 0.82 | 0.83 | 0.85 | **0.84** |
| Stress-Related | 0.84 | 0.82 | **0.83** | 0.66 | 0.72 | 0.69 | 0.81 | 0.75 | 0.78 |
| Mood Disorders | 0.84 | 0.81 | **0.83** | 0.68 | 0.75 | 0.71 | 0.72 | 0.87 | 0.79 |
| OCRD | 0.84 | 0.87 | **0.85** | 0.82 | 0.56 | 0.67 | 0.82 | 0.84 | 0.83 |
| Other | 0.20 | 0.09 | 0.12 | 0.05 | 0.06 | 0.06 | 0.62 | 0.15 | **0.24** |
| **Macro Avg** | 0.70 | 0.69 | 0.69 | 0.60 | 0.59 | 0.59 | 0.76 | 0.69 | 0.70 |
| **Weighted Avg** | 0.79 | 0.81 | 0.79 | 0.71 | 0.70 | 0.70 | 0.79 | 0.80 | 0.78 |

### A. Performance of Machine Learning Models

Among the three ML models, SVM achieved the strongest overall performance, with an accuracy of 0.83 and a weighted F1-Score of 0.81, followed by LR at 0.82 and a weighted F1-Score of 0.81, and RF at 0.81 and a weighted F1-Score of 0.79. A closer look at class-wise performance reveals further distinctions among the three models. LR achieved the highest F1-Score in Mood Disorders (0.85) and shared the top score with SVM in Stress-Related Disorders (0.86). RF performed competitively in the four main diagnostic categories, with F1-Scores ranging from 0.83 to 0.84, but recorded an F1-Score of 0.00 in the "Other" category. Fig. 2 shows the ROC curves for the SVM model, with a weighted AUC of 0.93. Stress-related Disorders recorded the highest AUC of 0.97, while the "Other" categories recorded the lowest at 0.68.

To assess whether the observed performance differences were statistically meaningful, we computed the mean and standard deviation of the weighted F1-Score across the five cross-validation folds for each model (Table VI) and conducted paired t-tests comparing with SVM, the best-performing model on the held-out test set, with each of the other models (Table VII). Differences between the models were considered statistically significant at $p < 0.05$. LR significantly outperformed SVM ($t(4) = -3.11$, $p = 0.036$), and SVM significantly outperformed LSTM ($t(4) = 20.43$, $p < 0.001$). No statistically significant differences were found between the SVM and RF ($t(4) = 0.44$, $p = 0.681$), CNN ($t(4) = 0.78$, $p = 0.480$), or MentalBERT ($t(4) = 0.64$, $p = 0.559$) models.

### B. Performance of Deep Learning and Transfer Learning Models

Among the DL models, CNN attained an accuracy of 0.81 and a weighted F1-Score of 0.79, achieving the highest F1-Score of 0.85 in the OCRD category across all six evaluated models. LSTM recorded the lowest overall performance with an accuracy of 0.70 and a weighted F1-Score of 0.70, with particularly weak results for Stress-Related Disorders (F1 = 0.69) and OCRD (F1 = 0.67). MentalBERT achieved an accuracy of 0.80 and a weighted F1-Score of 0.78.

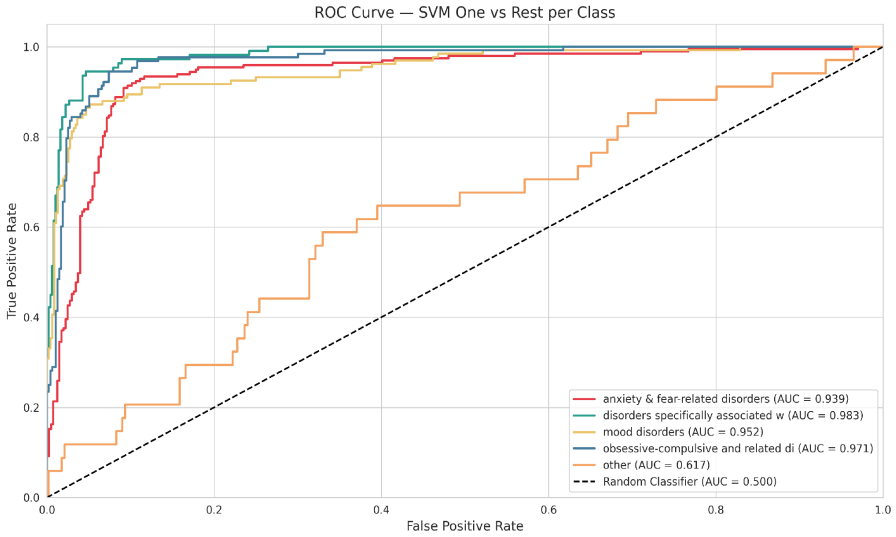



Fig. 2. ROC curves for the SVM model on the MANAS-ICD11 test set and Weighted AUC = 0.93.

TABLE VI. FOLD-WISE ACCURACY, F1-SCORE, AND WEIGHTED AUC FOR ALL SIX MODELS UNDER STRATIFIED 5-FOLD CROSS-VALIDATION

| Category | Model | Fold 1 | | Fold 2 | | Fold 3 | | Fold 4 | | Fold 5 | | Weighted AUC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | F1-Score | Acc | F1-Score | Acc | F1-Score | Acc | F1-Score | Acc | F1-Score | |
| Machine Learning (ML) Models | SVM | 0.82 | 0.80 | 0.80 | 0.78 | 0.81 | 0.79 | 0.79 | 0.78 | 0.81 | 0.79 | 0.93 |
| | LR | 0.81 | 0.80 | 0.81 | 0.81 | 0.81 | 0.80 | 0.79 | 0.78 | 0.82 | 0.81 | 0.94 |
| | RF | 0.81 | 0.79 | 0.78 | 0.76 | 0.81 | 0.79 | 0.78 | 0.77 | 0.82 | 0.81 | 0.93 |
| Deep Learning (DL) Models | CNN | 0.81 | 0.80 | 0.80 | 0.79 | 0.79 | 0.78 | 0.75 | 0.76 | 0.80 | 0.79 | 0.93 |
| | LSTM | 0.60 | 0.62 | 0.61 | 0.59 | 0.53 | 0.55 | 0.60 | 0.59 | 0.63 | 0.59 | 0.86 |
| Transfer Learning (TL) Models | MentalBERT | 0.80 | 0.80 | 0.76 | 0.76 | 0.79 | 0.79 | 0.79 | 0.79 | 0.80 | 0.80 | 0.93 |

TABLE VII. STATISTICAL COMPARISON OF CROSS-VALIDATION PERFORMANCE ACROSS ALL MODELS (PAIRED T-TEST VS. SVM)

| Model | Mean F1 ± SD | *t*-statistic | *p*-value | Significant |
|---|---|---|---|---|
| LR | 0.805 ± 0.010 | −3.11 | 0.036 | Yes |
| RF | 0.786 ± 0.020 | 0.44 | 0.681 | No |
| CNN | 0.785 ± 0.015 | 0.78 | 0.480 | No |
| LSTM | 0.590 ± 0.021 | 20.43 | <0.001 | Yes |
| MentalBERT | 0.785 ± 0.013 | 0.64 | 0.559 | No |

### C. Result Analysis

As shown in Tables IV and V, the four primary disease categories were consistently classified with high F1-Scores ranging from 0.67 to 0.86 across all models, demonstrating that the MANAS-ICD11 corpus provides adequate discriminative linguistic information for each category. Across all evaluated models, the "Other" category recorded the lowest F1-Scores, ranging from 0.00 to 0.24, consistent with its inherently residual and heterogeneous nature. Fig. 3 shows the confusion matrices for all six evaluated models. Across models, the predominant source of error was the mis-assignment of posts to the mood disorder category, with this trend being most prominent in the SVM, LR, and MentalBERT models. This trend is clinically relevant, with a number of keywords such as sadness, hopelessness, and emotional distress often appearing in several disorder categories in the same Reddit post, creating language overlap that is problematic for automated classifiers [16].

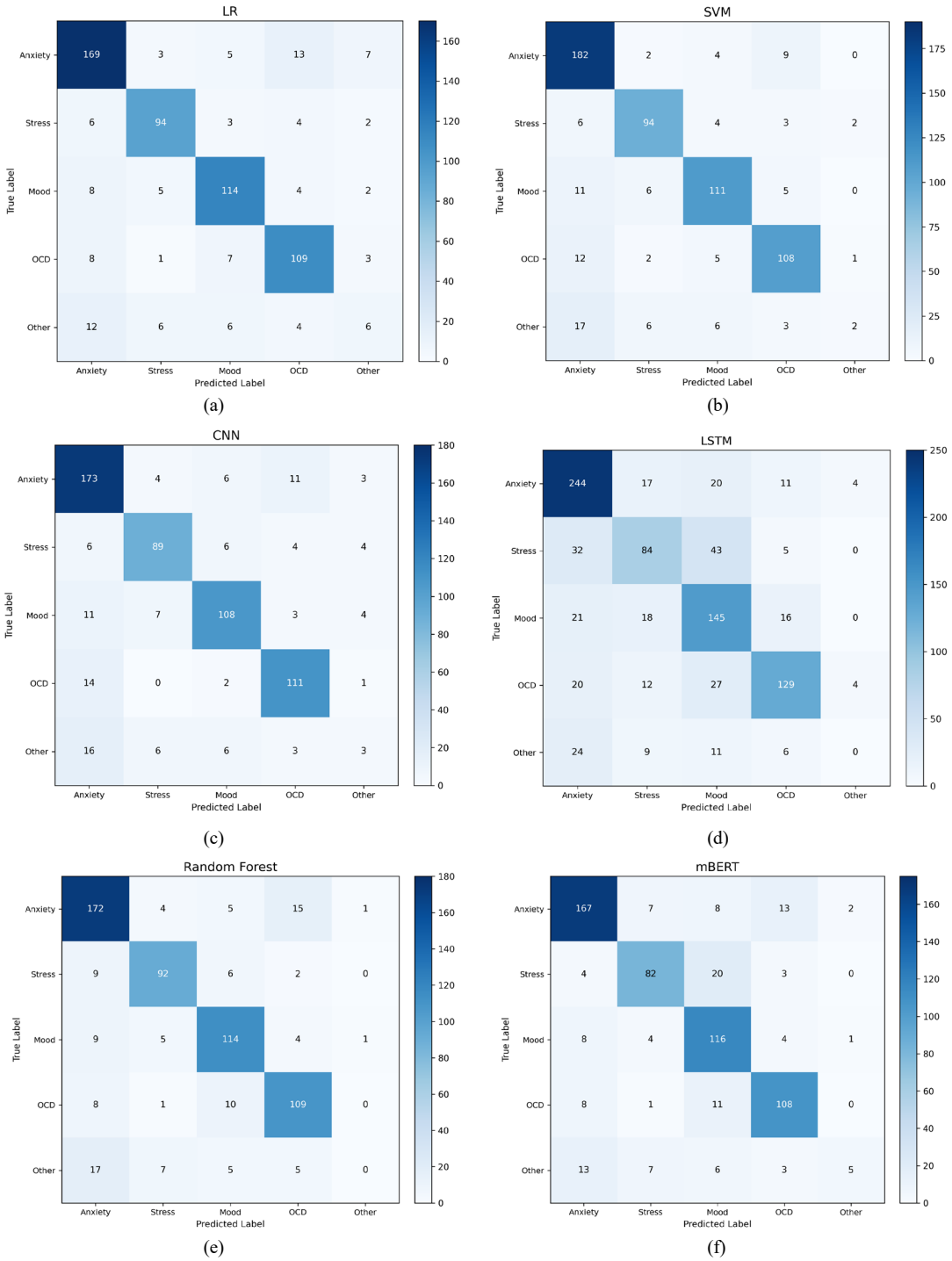



Fig. 3. Confusion matrices for all six evaluated models on the MANAS-ICD11 test set: (a) LR, (b) SVM, (c) CNN, (d) LSTM, (e) RF, (f) MentalBERT.

Table VIII shows that posts with unambiguous disorder-specific language were well classified under the SVM model. All six models were compared based on the test accuracy, Weighted F1-Score and Weighted AUC, as shown in Fig. 4. The main types of misclassifications were between categories with varying clinical overlap, especially between Anxiety/Fear and Stress-Related Disorders. The "Other" category continued to be the most difficult, mostly because of its residual and heterogeneous nature.

TABLE VIII. QUALITATIVE ERROR ANALYSIS OF ACTUAL VS. SVM-PREDICTED DIAGNOSTIC LABELS WITH CLINICAL INTERPRETATION OF CORRECT AND MISCLASSIFIED CASES

| Content Summary | Actual Label | Predicted Label | Model Interpretation |
|---|---|---|---|
| Post expresses persistent low mood, loss of interest, fatigue, and hopelessness | Mood Disorder | Mood Disorder | Clear depressive markers enable accurate classification |
| Recurrent panic attacks, intense fear, physical symptoms, anticipatory anxiety | Anxiety & Fear | Anxiety & Fear | Strong anxiety-related lexical patterns support correct prediction |
| Intrusive thoughts, repetitive compulsive behaviors, inability to control actions | OCRD | OCRD | Disorder-specific terms like intrusive and compulsion are highly discriminative |
| Trauma exposure, emotional flashbacks, heightened arousal, avoidance behavior | Stress-Related | Stress-Related | Trauma-linked expressions guide classification toward stress-related category |
| Identity confusion and emotional instability, no clear diagnostic indicators | Other | Anxiety & Fear | Ambiguous content leads model to default toward dominant anxiety class |
| Panic-like symptoms but also trauma triggers and past distressing experiences | Stress-Related | Anxiety & Fear | Linguistic overlap between anxiety and stress-related disorders causes misclassification |
| Therapy experiences in general manner without describing active symptoms | Other | Other | Absence of clinical symptoms correctly leads to non-disorder classification |
| Panic attack with physical symptoms resembling heart attack, acute distress | Anxiety & Fear | Stress-Related | Somatic symptom overlap leads to confusion between anxiety and stress-related classes |

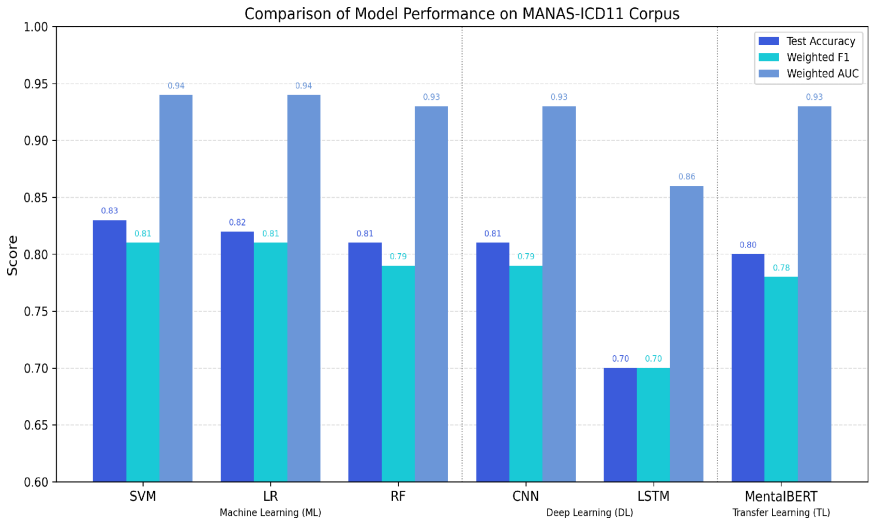



Fig. 4. Performance of all six models tested on the MANAS-ICD11 corpus.

*1) Classical ML vs. DL and transformer models*

One of the most notable findings of this study is that traditional ML models, such as SVM and LR, consistently outperformed DL and transformer-based approaches on the MANAS-ICD11 corpus. Although this result may appear counterintuitive in light of the recent dominance of transformer-based models in NLP benchmarks, it is consistent with the observations reported in comparable low-resource mental health classification settings [41]. The strong results obtained by the TF-IDF-based models indicate that mental health discourse in the MANAS-ICD11 corpus is characterised by rich disorder-specific lexical patterns. For example, stress-related posts often feature trauma and avoidance language, whereas obsessive-compulsive posts include compulsive and intrusive terminology, all of which are effectively captured by TF-IDF weighting. SVM's large-margin classification boundary is particularly well suited for exploiting such discriminative lexical signals in high-dimensional sparse feature spaces [22]. With only 2404 training instances, the corpus represents a low-data regime that is challenging for Deep Learning (DL) models. This is particularly relevant for LSTM architectures, whose ability to model long-range sequential dependencies from scratch typically depends on substantially larger training sets [34].

Although MentalBERT was pre-trained on mental health-related Reddit text [36], its pre-training objective was primarily aligned with binary classification tasks rather than the nuanced five-class clinical diagnostic problem posed by the MANAS-ICD11 corpus. Distinguishing between closely related categories, such as Anxiety/Fear and Stress-Related, requires sensitivity to subtle clinical distinctions that a model pre-trained on coarser binary labels may not fully capture. These findings

suggest that domain-adaptive pre-training directly on ICD-11-aligned corpora may be needed to fully realize the potential of transformer models for fine-grained mental health classification [36, 41]. This is consistent with observations in other domain-specific classification settings, where LLM-based embeddings adapted to specialized terminology outperform general-purpose representations [42].

The statistical significance results (Table VII) further refine this interpretation. While SVM's aggregate metrics position it as the top-performing model on the held-out test set, LR significantly outperformed SVM across the cross-validation folds, suggesting that LR offers more consistent classification performance despite a comparable point-estimate weighted F1-Score in Table IV. In contrast, the performance gaps of SVM against RF, CNN, and MentalBERT were not statistically significant, indicating that these models are practically comparable on the MANAS-ICD11 corpus and that the differences observed in Tables IV and V may fall within the range of fold-to-fold variability rather than reflecting a genuine architectural advantage. Only the gap between SVM and LSTM was both large and statistically robust, confirming that LSTM's weaker performance is a reliable finding rather than an artefact of the train-test split. These results suggest that the classification advantage of TF-IDF-based lexical models over deep learning and transformer architectures, while directionally consistent with prior findings [41], should be interpreted with caution for architectures other than LSTM.

Despite these contributions, the present study is subject to several limitations. First, all six models were trained using standard default hyperparameter configurations to establish a consistent and reproducible baseline, rather than being individually tuned; this ensured a fair, uniform comparison but may not reflect the optimal achievable performance of each model. In particular, MentalBERT's default 128-token limit meant that 53.28% of posts in the corpus were truncated during fine-tuning (mean = 163.69 tokens, max = 2212), likely contributing to information loss in longer posts. Second, the relatively small training set of 2404 instances may have further constrained the performance of DL and transformer-based models compared to lexical approaches. Future work should address these constraints through hyperparameter optimisation, longer sequence lengths or hierarchical encoding for transformer models, corpus expansion, and multilingual data collection. Third, the corpus was restricted to English-language Reddit posts, which may limit generalisability to other languages and platforms.

*2) Comparison with existing mental health corpora*

The quality of the corpora on which they are trained is critical for developing reliable mental health NLP systems. To contextualise the MANAS-ICD11 corpus with respect to other existing mental health datasets, the different annotation strategies, diagnostic scopes, and source platforms of each are summarised in Table I. As shown in Table I, there is a trade-off between the scale of the corpus and its clinical credibility in the current corpora. Some large databases, such as the SMHD [7], provide coverage of more than 900,000 posts but rely on automated self-reported diagnosis patterns that are not clinically annotated. Likewise, the extent to which crowdsourced annotations are used to aid diagnosis is restricted in CLPsych 2015 [21] and Dreaddit [7].

MentalHelp [17] comprises 14 million instances but relies on semi-supervised labelling using an ensemble of language models. Introduces label noise and is not guaranteed to be clinically accurate. A major improvement would be datasets that include expert human annotations, such as LoST [9] and WellXplain [15]. Their annotation teams consisted of a clinical psychologist and a psychiatrist. However the social NLP researcher made decisions on diagnosis in addition to clinical reasoning, based on computational concerns. This approach may be effective for NLP-based activities but may not adequately reflect the clinical reasoning required for the accurate diagnosis of mental health disorders. More recent efforts have pushed clinical rigor further but still fall short of MANAS-ICD11's combination of diagnostic breadth and authentic patient discourse. PsychiatryBench [30] is annotated exclusively by licenced psychiatrists, but its question-answer items are drawn from psychiatric textbooks and casebooks rather than real patient-generated text, which limits its applicability to screening naturalistic social media disclosures. ReDepress [43] achieves clinician annotation of Reddit users with a reported Krippendorff's alpha of 0.76, yet its scope is restricted to a single disorder trajectory, depression relapse, rather than a multi-class diagnostic schema.

Among the corpora reviewed in this study, MANAS-ICD11 is, to our knowledge, the sole resource that integrates clinician-sourced annotations, naturalistic Reddit discourse, and a five-category ICD-11-aligned diagnostic structure. Beyond this combination, MANAS-ICD11 introduces a novel annotation process in which annotators first identify disorder-specific keywords from each post, use these keywords to derive the underlying clinical characteristics, and finally assign a diagnostic label grounded in this two-step clinical reasoning, ensuring that every label remains traceable to explicit clinical evidence. A more extensive search of unpublished or non-English corpora may uncover additional initiatives with similar objectives.

The MANAS-ICD11 corpus was developed to address existing gaps in the field. The annotation process was conducted solely by qualified medical professionals, including one psychiatrist and two clinical psychologists, with no involvement from NLP researchers during the labelling phase. Consequently, all diagnostic decisions within the corpus are based entirely on clinical judgment, thereby enhancing their clinical credibility among Reddit-based mental health data sets [32]. Furthermore, rather than assigning labels in isolation, each expert performed three tasks for every post: identifying disorder-specific keywords, documenting the clinical characteristics that supported the diagnosis, and assigning a final ICD-11-aligned label. Consequently, each label is grounded in explicit clinical evidence, providing a level of annotation granularity that appears to be absent from the

other corpora considered in this study. Third, the corpus adopts an ICD-11 aligned label schema defined by the WHO, consisting of five clinically meaningful diagnostic categories that reflect the inherent multi-disorder complexity of real-world mental health presentations, rather than simplifying the task to binary detection as most prior datasets do. "Collectively, these characteristics position the MANAS-ICD11 corpus as a highly interpretable and clinically grounded resource, offering a useful complement to existing mental health NLP benchmarks.

## VI. Conclusion

The MANAS-ICD11 corpus establishes a clinically grounded benchmark for multi-class mental health disorder classification, addressing a critical gap in the availability of medically annotated resources for computational psychiatry. The benchmark evaluation confirmed that ICD-11-aligned clinical annotation produces a learnable and discriminative corpus, with traditional ML models demonstrating competitive performance compared with domain-adapted transformer architectures. These findings suggest that clinical lexical richness, rather than model complexity alone, is a key determinant of classification performance in expert-annotated mental health corpora. The corpus has broad applicability in clinical decision support, early screening systems and public health surveillance. Future work will focus on improving class balance, exploring advanced transfer learning strategies, and developing an interactive application that identifies key linguistic indicators and delivers interpretable mental health data to support informed clinical decision making.

## Data Availability

The MANAS-ICD11 corpus is available on the HuggingFace Hub under a gated-access model. Researchers may request access for academic and non-commercial research purposes, subject to approval and agreement with ethical usage guidelines. The dataset is accessible at: https://huggingface.co/datasets/anushamd23/ManasICD11_dataset.

## Conflict of Interest

The authors declare no conflicts of interest.

## Author Contributions

Anusha M D was responsible for conceptualisation, data collection and curation, corpus development, annotation coordination, methodology design, software implementation, formal analysis, validation, and writing of the original draft. Parameshwar R Hegde contributed to the conceptualisation, research design, clinical expert coordination, methodology oversight, result interpretation, project administration, supervision of the overall research direction, and review and editing of the manuscript. All authors had approved the final version.

## Acknowledgment

The authors gratefully acknowledge the clinical expertise of Dr. Poonam Santhosh (MBBS, MD Psychiatry), Professor and Head of the Department of Psychiatry, K.V.G. Medical College, Sullia, and Consultant Psychiatrist, KMC Hospital, Mangaluru; Ms. Navyashree (M.Sc. Psychology in Clinical and Counselling Psychology), Counselling Psychologist, Indian Institute of Technology (IIT) Dharwad; and Ms. Namratha K.B. (M.Sc. in Psychology, MSW Medical and Psychiatry), Psychologist, K.V.G. Medical College and Hospital, Sullia, India, for their invaluable contribution to clinically annotating the Reddit dataset. The authors gratefully acknowledge the Research Fellowship and research facilities provided by Yenepoya (Deemed to be University), Mangaluru, India.

## References

[1] E. N. Mboweni, M. H. Mphasha, and L. Skaal, "Exploring Mental Health Awareness: A Study on Knowledge and Perceptions of Mental Health Disorders among Residents of Matsafeni Village, Mbombela, Mpumalanga Province," *Healthcare*, vol. 12, no. 1, 85, 2023. doi: 10.3390/healthcare12010085

[2] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Sci. Rep.*, vol. 10, no. 1, 11846, 2020. doi: 10.1038/s41598-020-68764-y

[3] M. Garg, "Mental Health Analysis in Social Media Posts: A Survey," *Arch. Comput. Methods Eng.*, vol. 30, no. 3, pp. 1819–1842, 2023. doi: 10.1007/s11831-022-09863-z

[4] M. M. T. Ayyalasomayajula, A. Agarwal, and S. Khan, "Reddit social media text analysis for depression prediction: using logistic regression with enhanced term frequency-inverse document frequency features," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 5, pp. 5998–6005, 2024. doi: 10.11591/ijece.v14i5.pp5998-6005

[5] N. P. Shetty, Y. Singh, V. Hegde, D. Cenitta, and D. K, "Exploring Emotional Patterns in Social Media Through NLP Models to Unravel Mental Health Insights," *Healthcare Technol. Lett.*, vol. 12, no. 1, 12096, 2025. doi: 10.1049/htl2.12096

[6] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and N. Goharian, "SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions," in *Proc. 27th Int. Conf. Comput. Linguistics (COLING)*, 2018, pp. 1485–1497.

[7] E.Turcan and K. McKeown, "Dreaddit: A Reddit Dataset for Stress Analysis in Social Media," in *Proc. 10th Int. Workshop Health Text Mining Inf. Anal. (LOUHI)*, 2019, pp. 97–107. doi: 10.18653/v1/D19-6213

[8] M. Garg, C. Saxena, S. Saha, V. Krishnan, R. Joshi, and V. Mago, "CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts," in *Proc. 13th Language Resources Eval. Conf. (LREC)*, 2022, pp. 6387–6396.

[9] M. Garg, M. Gaur, R. Goswami, and S. Sohn, "LoST: A Mental Health Dataset of Low Self-Esteem in Reddit Posts," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, 2023, pp. 3854–3859.

[10] E. Bao, A. Pérez, and J. Parapar, "ReDSM5: A Reddit Dataset for DSM-5 Depression Detection," in *Proc. 34th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2025, pp. 6323–6327.

[11] Y. Chen *et al.*, "From symptom tracking to prevention—A transformer-based dynamic model for predicting mild cognitive impairment risk in older adults with depression: A longitudinal study based on CHARLS and CLHLS," *Medicine*, vol. 105, no. 19, e48653, 2026. doi: 10.1097/MD.0000000000048653

[12] World Health Organization, *Clinical Descriptions and Diagnostic Requirements for ICD-11 Mental, Behavioural and Neurodevelopmental Disorders*, Geneva, Switzerland: World Health Organization, 2024.

[13] D. A. Regier, E. A. Kuhl, and D. J. Kupfer, "The DSM-5: Classification and Criteria Changes," *World Psychiatry*, vol. 12, no. 2, pp. 92–98, 2013. doi: 10.1002/wps.20050

[14] M. Malgaroli, T. D. Hull, J. M. Zech, and T. Althoff, "Natural Language Processing for Mental Health Interventions: A Systematic Review and Research Framework," *Transl. Psychiatry*, vol. 13, no. 1, 309, 2023. doi: 10.1038/s41398-023-02592-2

[15] M. Garg, "WellXplain: Wellness Concept Extraction and Classification in Reddit Posts for Mental Health Analysis," *Knowl.-Based Syst.*, vol. 284, 111228, 2024. doi: 10.1016/j.knosys.2023.111228

[16] J. Pavez and H. Allende, "A Hybrid System Based on Bayesian Networks and Deep Learning for Explainable Mental Health Diagnosis," *Appl. Sci.*, vol. 14, no. 18, 8283, 2024.

[17] N. Raihan, S. S. C. Puspo, S. Farabi, A. M. Bucur, T. Ranasinghe, and M. Zampieri, "MentalHelp: A Multi-Task Dataset for Mental Health in Social Media," in *Proc. 2024 Joint Int. Conf. Comput. Linguistics, Language Resources Eval. (LREC-COLING)*, Torino, Italy, 2024, pp. 11196–11203.

[18] B. Bauer, R. Norel, A. Leow, Z. AbiRached, B. Wen, and G. A. Cecchi, "Using Large Language Models to Understand Suicidality in a Social Media-Based Taxonomy of Mental Health Disorders: Linguistic Analysis of Reddit Posts," *JMIR Ment. Health*, vol. 11, e57234, 2024. doi: 10.2196/57234

[19] S. Rani, K. Ahmed, and S. Subramani, "From Posts to Knowledge: Annotating a Pandemic-Era Reddit Dataset to Navigate Mental Health Narratives," *Appl. Sci.*, vol. 14, no. 4, 1547, 2024.

[20] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying Mental Health Signals in Twitter," in *Proc. Workshop Comput. Linguistics Clin. Psychol. (CLPsych)*, Baltimore, MD, USA, 2014, pp. 51–60. doi: 10.3115/v1/W14-3207

[21] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "CLPsych 2015 Shared Task: Depression and PTSD on Twitter," in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol. (CLPsych)*, Denver, CO, USA, 2015, pp. 31–39. doi: 10.3115/v1/W15-1204

[22] N. Merayo, A. Ayuso-Lanchares, and C. González-Sanguino, "Machine Learning and Natural Language Processing to Assess the Emotional Impact of Influencers' Mental Health Content on Instagram," *PeerJ Comput. Sci.*, vol. 10, e2251, 2024. doi: 10.7717/peerj-cs.2251

[23] N. Merayo, A. Ayuso-Lanchares, and C. González-Sanguino, "Machine Learning Algorithms to Address the Polarity and Stigma of Mental Health Disclosures on Instagram," *Expert Syst.*, vol. 42, e13832, 2025. doi: 10.1111/exsy.13832

[24] S. Chinchali and P. B. Patil, "Sarcasm Detection and Classification in Kannada Language Using Machine Learning Techniques on a Manually Annotated Dataset," *SN Comput. Sci.*, vol. 7, 36, 2026. doi: 10.1007/s42979-025-04557-2

[25] K. Hasan, J. Saquer, and Y. Zhang, "Mental Multi-class Classification on Social Media: Benchmarking Transformer Architectures against LSTM Models," in *Proc. 2025 IEEE Int. Conf. Mach. Learning Appl. (ICMLA)*, 2025, pp. 160–167.

[26] O. Ovcharuk, O. Mazurets, M. Molchanova *et al.*, "Detecting Mental Disorders in Social Media Using a Transformer-Based Ensemble of Binary Classifiers," *medRxiv*, vol. 12, 16.25342390, 2025.

[27] E. Holderness, B. Atwood, M. Verhagen, A. K. Shinn, P. Cawkwell, H. Cerruti, J. Pustejovsky, and M.-H. Hall, "Machine learning in psychiatric health records: A gold standard approach to trauma annotation," *Translational Psychiatry*, vol. 15, 260, 2025.

[28] M. Lamparth, D. Grabb, A. Franks, S. Gershan, K. N. Kunstman, A. Lulla, M. Drummond Roots, M. Sharma, A. Shrivastava, N. Vasan, and C. Waickman, "Moving Beyond Medical Exam Questions: A Clinician-Annotated Dataset of Real-World Tasks and Ambiguity in Mental Healthcare," arXiv preprint, arXiv:2502.16051, 2025.

[29] H. Alhuzali, N. Alasmari, and H. Alsaleh, "MentalQA: An Annotated Arabic Corpus for Questions and Answers of Mental Healthcare," *IEEE Access*, vol. 12, pp. 101155–101165, 2024.

[30] A. E. Fouda, A. A. Hassan, R. J. Hanafy, and M. E. Fouda, "PsychiatryBench: A Multi-Task Benchmark for LLMs in Psychiatry," arXiv preprint, arXiv:2509.09711, 2025. https://arxiv.org/abs/2509.09711

[31] K. E. Anderson, "Ask Me Anything: What Is Reddit?" *Libr. Hi Tech. News*, vol. 32, no. 5, pp. 8–11, 2015.

[32] R. Artstein and M. Poesio, "Survey Article: Inter-Coder Agreement for Computational Linguistics," *Comput. Linguistics*, vol. 34, no. 4, pp. 555–596, 2008. doi: 10.1162/coli.07-034-R2

[33] G. Marzi, M. Balzano, and D. Marchiori, "K-Alpha Calculator – Krippendorff's Alpha Calculator: A User-Friendly Tool for Computing Krippendorff's Alpha Inter-Rater Reliability Coefficient," *MethodsX*, vol. 12, 102545, 2024.

[34] A. Amanat *et al.*, "Deep Learning for Depression Detection from Textual Data," *Electronics*, vol. 11, no. 5, 676, 2022. doi: 10.3390/electronics11050676

[35] K. Sparck Jones, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," *J. Documentation*, vol. 28, no. 1, pp. 11–21, 1972. doi: 10.1108/eb026526

[36] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare," in *Proc. 13th Language Resources Eval. Conf. (LREC)*, Marseille, France, Jun. 2022, pp. 7184–7190.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technol. (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423

[38] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," arXiv preprint, arXiv:1711.05101, 2017. https://arxiv.org/abs/1711.05101

[39] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU," arXiv preprint, arXiv:2305.17473, 2023. https://arxiv.org/abs/2305.17473

[40] M. R. Rahib, A. H. Tamim, M. Z. Tahmeed, and M. J. Hossain, "Emotion Detection Based on Bangladeshi People's Social Media Response on COVID-19," *SN Comput. Sci.*, vol. 3, no. 2, 180, 2022. doi: 10.1007/s42979-022-01077-1

[41] M. Garg, "Towards Mental Health Analysis in Social Media for Low-Resourced Languages," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 23, no. 3, pp. 1–22, 2024. doi: 10.1145/3638761

[42] Z. E. Ergun and E. Sefer, "Financial Statement Fraud Detection via Large Language Models," *Intelligent Systems in Accounting, Finance and Management*, vol. 32, no. 4, e70021, 2025.

[43] A. K. Agarwal, S. Bhattacharjee, M. Rastogi, J. S. Jacob, B. Banerjee, R. Gupta, and P. Bhattacharyya, "ReDepress: A Cognitive Framework for Detecting Depression Relapse from Social Media," in *Proc. 2025 Conf. Empirical Methods Natural Language Process. (EMNLP)*, Nov. 2025, pp. 34652–34670.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).