AI in the Wild: A Meta-Analytic Evaluation of Depression Detection from Social Media Data

Xianglu Tang [1], Joyee W. Jin[1,2], Emily Ma[1], Xingyu Li [1],

[1]Institute for Human-centered AI, Stanford University
[2]Department of Computer Science, University of Toronto
xianglutang111@gmail.com, joyee.jin@mail.utoronto.ca, emilylihanma88@gmail.com, axyli@stanford.edu

Abstract

As AI moves into high-stakes, human-centered settings, we still lack clear evidence on *when* and *why* these systems succeed or fail. This meta-analysis synthesizes all empirical studies published between 2022 and 2025 that use social-media data to predict depression, quantifying pooled accuracy and testing study-level moderators. By showing how data sources and model architecture shape outcomes, we offer an empirical foundation for a more reliable, socially aware deployment of AI in mental health.

Across 67 studies, overall performance is strong (pooled $r \approx 0.80$) and climbs even higher in 2024, driven by deep, transformer-based and multimodal systems. The gains, however, are uneven: post-level binary detectors improve the most, user-level severity estimation still lags, and results hinge as much on label provenance and platform context as on model size—highlighting a persistent gap between leaderboard success and clinically meaningful reliability.

To address that gap, we propose a *Psych-Aligned Evaluation Framework* that maps predictions onto validated symptom dimensions and adds three deployment-critical tests—PHQ error, temporal stability, and clinician agreement. This framework converts single-number benchmarks into a multidimensional yardstick for real-world, psychologically meaningful depression detection.

Datasets and Appendix —
https://github.com/zjoyjin/Meta-Analysis-DataSet

Introduction

Mental disorders burden one in eight people worldwide (World Health Organization 2022), yet clinical screening remains labor-intensive and unevenly distributed. Social-media traces promise real-time, population-scale signals of depression, inspiring a decade of machine-learning research. Narrative reviews have mapped this terrain (Chancellor and De Choudhury 2020; Liu et al. 2022), and the first quantitative synthesis has now pooled studies from 2008–2023 (Phiri et al. 2025). However, prior work still centers on headline accuracy: it lacks system-level insight into reliability, generalizability, and ethical implications. Existing leaderboards seldom expose variability, social bias, or deployment readiness, leaving open the question of why AI systems work, or fail, when deployed "in the wild."

To close this gap, we meta-analyze 67 studies published between 2022 and 2025. Performance climbs steeply in the latest cohort, led by deep, multimodal models, yet gains remain uneven: post-level binaries improve most, whereas user-level severity still lags.

We then outline the persistent divide between benchmark success and clinical readiness, introducing our **Psych-Aligned Evaluation Framework**—a blueprint for translational AI that combines PHQ-error, temporal stability, and clinician agreement with standard metrics to enable safer deployment.

Contributions

1. A comprehensive, up-to-date synthesis that moves beyond benchmark datasets;
2. Quantitative evidence of context-sensitive moderators;
3. Guidance toward better benchmarks and evaluation standards that surface generalization gaps and social risk; and
4. A path to more reliable, equitable, and human-centered AI models, illuminating the socio-technical conditions under which they succeed or fail.

Related Work

Method Evolution in Depression Detection

Since 2015, social-media depression detection has progressed through three methodological waves. Early studies relied on hand-crafted text features. These included n-grams, TF-IDF weights, lexicon-based sentiment such as VADER, and psychologically grounded LIWC categories. Researchers paired these features with traditional classifiers, including Support Vector Machines, Random Forests, and Naïve Bayes (Salas-Zárate, Valencia-García, and García-Díaz 2022; Phiri et al. 2025). These models captured lexical patterns, which transferred poorly across platforms and user populations.

The second wave adopted neural networks, enabling automatic learning of longer-range syntactic and semantic dependencies (Zhang et al. 2023). The most recent wave leverages transformer-based language models, such as BERT and RoBERTa, achieving accuracy through large-scale

pre-training and task-specific fine-tuning (Padmaja et al. 2025). However, no previous research has thoroughly evaluated these new architectural improvements in terms of their computational expense, interpretability, or cross-domain effectiveness. Our meta-analysis systematically compares these trade-offs.

Feature Landscape

The field has moved beyond simply searching for specific words; it now leverages intricate, AI-generated understandings of language. Hand-crafted features remain valuable for theory-driven interpretation; for example, counts of negative affect words, first-person pronouns, and absolutist language align with cognitive-behavioral constructs such as self-focus and cognitive distortion (Chancellor and De Choudhury 2020). Neural approaches increasingly replace these cues with embeddings, such as word2vec, GloVe, or contextualized transformer vectors (Trifu et al. 2024). Beyond text, some studies have begun to explore multimodal pipelines that integrate images, posting rhythms, or social-interaction networks. However, these efforts remain limited and inconsistent (Trotzek, Koitka, and Friedrich 2018).

Label Quality and Ethical Compliance

Ground-truth labeling and ethical adherence remain persistent weaknesses in social-media depression detection. Approximately 75% of datasets rely on weak supervision, including subreddit membership, self-disclosure keywords, or crowd annotations, with only a minority leveraging standardized clinical interviews or rating scales (Yang et al. 2024; Jin, Ye, and Li 2024). Even "gold-standard" clinical scales introduce subjectivity and inter-rater variability (Zhang et al. 2025; Jain et al. 2018). Meanwhile, reporting of ethical practices, privacy, informed consent, and data governance remains inconsistent and often under-documented.

Deployment Challenges

Despite strong benchmark performance, real-world deployment of social-media depression detection systems remains rare. Models often fail to generalize across new platforms, languages, or cultural contexts, revealing brittle domain shift (El-Sappagh, Al Shorman, and Abdelrazek 2025). Temporal robustness is rarely evaluated, making systems vulnerable to performance degradation as slang and platform norms evolve (Ferrario, Bianchi, and Rossi 2024).

Deep architectures also limit clinical trust due to their opacity, and explainable-AI techniques have yet to see widespread adoption (Ibrahimov, Aliev, and Gasimov 2024). Our study introduces a Psych-Aligned Evaluation Framework that integrates methodological, ethical, and deployment considerations, offering a roadmap for clinically trustworthy social-media depression detection.

Method

Corpus Identification and Search Strategy

Following PRISMA-2020 (Page et al., 2021), we searched IEEE Xplore, PsycINFO, PubMed, Scopus, Web of Science, and the ACM Digital Library for peer-reviewed English articles published 1 Jan 2022–30 Mar 2025, using modular strings that fuse depression, social-media, machine-learning, and prediction keywords; the full database-specific syntax appears in Appendix Table A1. Searches were limited to peer-reviewed English articles.

After automated and manual deduplication, 4,361 records proceeded to screening (see PRISMA flow diagram, Appendix Fig. 1). A human reviewer and GPT-4 Turbo—using the prompt reproduced in Appendix A—independently screened all titles and abstracts, achieving high inter-rater reliability ($\kappa = 0.92$). Conflicts were resolved by the human reviewer after full-text inspection, leaving 67 eligible studies and 649 effect estimates (344 novel models, 305 baseline models).

Inclusion Rules

Eligible studies met five criteria: (i) they predicted depression, (ii) analyzed social-media content, (iii) used machine-learning methods, (iv) reported performance metrics convertible to Pearson r, F1, or accuracy, and (v) detailed their algorithms, feature sets, label definitions, and evaluation procedures. Sentiment-only analyses and non-peer-reviewed papers were excluded.

Coding Procedure

We developed a codebook spanning study context, modelling choices, and labeling practice. Two raters independently applied the coding schema, with disagreements resolved through discussion.

Effect-Size Computation

We extracted three performance indices—Pearson's r, accuracy, and F1-score—and standardized them as follows. If r was not reported, we (i) converted AUC to Cohen's d and then to r (Hanley and McNeil 1982; Borenstein et al. 2009) or (ii) derived odds ratios from sensitivity/specificity and transformed them to r (Lipsey and Wilson 2001). Each r was Fisher-z transformed ($z = \operatorname{atanh} r$), with sampling variance $1/(n-3)$. Because these metrics capture distinct facets of model performance and are not directly comparable, we ran separate random-effects meta-analyses for z-transformed correlations, accuracy, and F1.

Meta-Analytic Strategy

All analyses were conducted in **R 4.3**. We used **metafor** for random-effects modeling and meta-regression, **stats** for linear models and t-tests, and **dplyr/tidyr** for data wrangling.

- **Random-effects synthesis**
  Each study's headline metric (Accuracy, F1, or r) was mapped to a common z-scale via $z = \operatorname{atanh}(\cdot)$; for r this is the standard Fisher transformation with sampling variance $1/(N-3)$, and for bounded indices (Accuracy, F1) we used the same $1/(N-3)$ as a sample-size–based working variance because confusion-matrix details were rarely reported. We fitted a null REML model with `metafor::rma()` to obtain the pooled effect and residual heterogeneity.

Layer	Variables captured	Coding format / notes
Study metadata	Publication year, sample size, social-media platform(s), data modality	Year treated as a continuous moderator; other fields recorded verbatim
Algorithm paradigm	*Traditional ML* (SVM, RF, LR); *Deep Learning* (CNN, LSTM, Bi-GRU, transformer); *Hybrid* (graph + text ensembles)	One categorical code per model row; studies supply multiple rows if they evaluate several paradigms
Processing depth (dummy)	**Deep processing = 1** $\rightarrow$ Deep Learning, Hybrid **Shallow processing = 0** $\rightarrow$ Traditional ML	Derived from the paradigm code; used in moderator analyzes
Feature engineering	Lexicon-based, statistical (TF-IDF / BOW), static embeddings, contextual embeddings, graph embeddings, multimodal fusion	Multi-label: each feature family coded as an independent binary flag
Feature function	Sentiment, discrete emotion, psychological theme, behavioral rhythm, personality indicators	Multi-label; one binary flag per function category
Label source	Clinical diagnosis, self-report scale (PHQ-9, CES-D), self-declaration (user-posted diagnosis), human annotation (manual or semi-automated)	One categorical code per model row
Task granularity	**Level:** user, post, sentence; **Output type:** binary, multiclass, regression; **Target:** depression (0/1), severity, tendency, depression vs. other disorders	Three facets stored separately; concatenated for subgroup analyzes

Table 1: Coding framework for study-level variables.

- **One-moderator meta-regressions**
  Candidate moderators—model depth, task level, baseline status, language, feature family, input modality, data source, annotation method—were entered individually as categorical predictors, again via REML. For each moderator we report the omnibus Q-test and the proportion of heterogeneity explained (R^2).
- **Robustness checks**
  Direction and independence were verified with OLS regressions that simultaneously controlled for all moderators above. Supplementary two-sample t-tests compared 2024 studies with earlier work; sensitivity analyses repeated all models after excluding baseline systems.

Quality, Ethics, and Bias Assessment

Beyond model variables, we designed a novel three-item *Data-Ethics Quality* (DEQ) rubric to assess procedural robustness: (1) transparent data provenance; (2) IRB or REC approval, or a justified exemption; and (3) explicit privacy safeguards. Two reviewers applied the rubric independently ($\kappa = 0.89$). DEQ scores were summarized descriptively and entered as moderators to test whether stronger ethical diligence predicts better model performance. These indicators were analyzed descriptively and included as moderators to assess whether ethical diligence correlates with technical performance.

Result

This section synthesizes findings from 67 social-media depression studies published between 2022 and 2025. We begin with an overview of the field's development and a discussion of key trends, followed by tracing the year-on-year rise in predictive accuracy, then quantify how model depth (deep vs. traditional) and input modality (unimodal-text vs. multimodal) drive performance variance. Finally, we examine deployment-critical gaps—scarce clinical labels, cross-platform fragility, temporal instability, privacy shortfalls, and limited explainability. Figure 1 provides a schematic of a typical social-media depression pipeline to orient these results.

Figure 1 condenses the full pipeline of social-media depression research. It tracks the flow from data sources (Twitter, Reddit, etc.) through task choices (user vs. post, longitudinal, multi-task), label quality (crowd tags $\rightarrow$ clinician diagnoses), and modalities (mostly text, little image/audio). Modeling splits into traditional ML, deep nets, and hybrids, all fed by embeddings, statistics, or lexicons. Outputs range from binary screens to PHQ-9 regression, and evaluation spans basic metrics to cross-platform and ethics checks.

Trends in Depression Detection Research

What 'Depression Detection' Usually Means in Practice
Most social-media depression models treat online activity as a sensor stream but differ in the signal unit they trust. Between 2022–2025, 79.1% of studies focus on post-level data—isolated tweets or forum entries—while only 22.4% aggregate user-level histories. This design choice aligns with a broader binary focus: across both levels, approximately 80% of models perform binary classification (depressed vs. not). Nuanced tasks remain rare: severity estimation appears in just 17.5% of post-level and 13.3% of user-level work, and disorder differentiation falls below 3.5%. Only one user-level study addresses depressive tendency, and just four studies take a prognostic approach—modeling future symptom trajectories rather than current status. In sum, current ma-

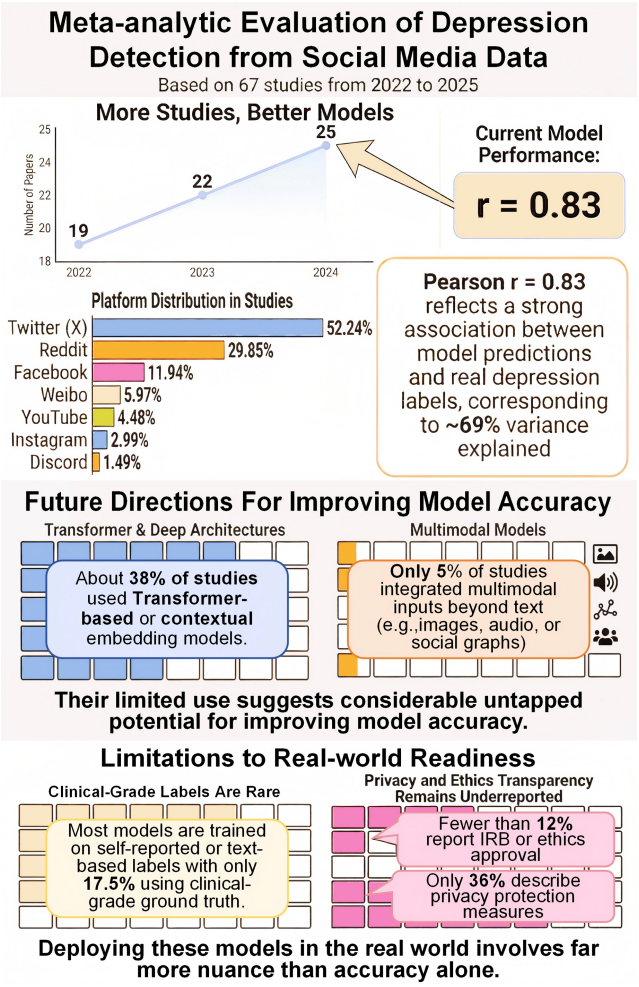

Figure 1: System overview: from social media data sources through modeling pipelines to clinical/ethical evaluation. Labels condensed for clarity.

chine learning approaches capture depression in a shallow, static manner rather than modeling it as a continuous and dynamic psychological state.

Data Sources & Modalities — Where the Field Actually Shops for Signals Depression-detection research continues to rely on a narrow band of social-media sources. Twitter accounts for 52.2% of all studies, followed by Reddit at 29.9%. All other platforms—Facebook, Weibo, YouTube, Instagram, and SMS—together contribute less than 18%. This hierarchy has remained stable from 2022 to 2024, with Twitter consistently dominating (40.0–57.9%) and Reddit trailing behind (21.0–30.0%). Such platform monoculture risks embedding Anglophone, text-centric biases in models intended for general use.

On the input side, homogeneity is even more pronounced. Across the 67 studies included in our coding, over 80% relied exclusively on textual data, most commonly raw posts or comment threads. Only 7.8% incorporated metadata, such as timestamps or follower counts, and just 5.2% employed multimodal pipelines that combined text with images or audio. Pure audio or video inputs were rare (2.6%), and LLM-enhanced approaches that used pseudo-labels or generated context appeared in only 1.3% of studies. Structured modalities, including sentiment graphs or longitudinal metadata, were almost entirely absent. This narrow input landscape reflects an emphasis on data that are convenient to collect, rather than data that most effectively capture mental health signals.

Diagnosis Label Sources Label quality underpins the credibility of any depression-prediction model, yet today's evidence base relies overwhelmingly on proxies rather than diagnoses. Human annotations—manual or semi-automated mood tags derived from text or subreddit flair—account for 76.5% of all labels, offering scalability at the cost of clinical grounding and demographic bias. Explicit self-disclosures in text contribute just 5.9% of studies, too sparse and platform-skewed to stabilize training sets. Psychometrically vetted self-report scales such as the PHQ-9 or CES-D appear in 10.3% of papers, improving rigor but still falling short of a formal diagnosis. The gold standard—clinician assessments or medical records—surfaces in only 7.4% of the corpus, a testament to the logistical barriers of healthcare partnerships. The field's dependence on non-clinical proxies may systematically overstate model performance, a concerning artifact for downstream applications.

Algorithm and Feature Choices—What the Models Are Actually Built From Granularity perspective—how machines stratify psychological signals. Across the corpus, most systems use *Traditional ML* such as SVM, Logistic Regression, and Random Forest (63.1%). Most models still operate at a coarse textual level: shallow/averaged embeddings or simple statistical features dominate (shallow embeddings 54.1%; n-gram/TF–IDF 43.8%). Concretely, these include frequency counts of words such as "sad" or "tired," bag-of-words/TF–IDF vectors, or averaged word2vec/GloVe embeddings—useful but largely capturing an overall emotional tint. Lexicon-based features (e.g., LIWC or VADER categories such as *negative affect*, *first-person pronouns*) appear less often (16.6%). Contextual embeddings (BERT-type), which track sentence-level meaning and irony, are used in 34.8% of model rows. Truly multidimensional inputs remain rare: multimodal pipelines that blend text with images or audio occur in only 4.3%, and social-graph features in just 2.2%.

Functional perspective—what psychological "components" are extracted. Feature choices echo this surface bias. Sentiment polarity registers in just 9.7% of models, serving as a crude emotional thermometer. Discrete emotions—anger, fear, joy—show up in only 3.1%, despite their clear links to action tendencies. Psychology-themed lexicons (loneliness, helplessness, rumination) appear in 23.3%, offering glimpses into maladaptive cognitions. behavioral rhythms (late-night posting, weekend withdrawal) feature in 3.1%, hinting at circadian disruption or social retreat. Personality cues such as first-person pronouns or syntactic complexity surface in fewer than 0.3%. Taken to-

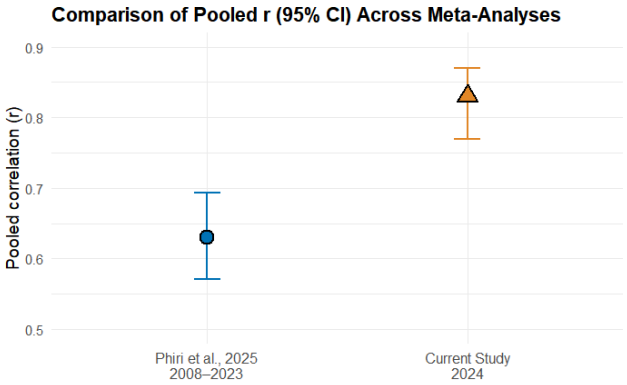

Figure 2: Comparison of pooled Pearson r (95% CI) across meta-analyses. Phiri et al. (2025) aggregated studies from 2008–2023, while the current study focuses on 2024. Error bars indicate 95% confidence intervals.

gether, today's algorithms mostly capture an emotional tint while overlooking deeper layers of cognition, behaviour, and personality—highlighting the current limits of machine understanding of the human mind.

Overall Performance

The first meta-analysis of social-media depression models and, after selecting one "best" system per study, reported a pooled Pearson $r = 0.63$ (95% CI 0.57–0.69) for work published between 2008 and 2023 Phiri et al. (2025) As shown in Figure 2, replicating their protocol but focusing on the most recent wave, we find a markedly stronger signal in 2024: eight new papers converge on $r = 0.83$ (0.77–0.87; see Fig. 2). When the window is widened to 2022–2024, the grand mean remains high at $r = 0.80$ (0.73–0.85). Such an effect explains roughly two-thirds of outcome variance—several times the median magnitude in social psychology. Performance, however, is uneven: the highest r reaches 0.95(Amanat et al. 2024), whereas the lowest bottoms out at r = 0.33 (Lyu et al. 2023).

Task granularity helps explain the spread: user-level binary classifiers average $r = 0.53$, user-level severity estimators drop to $r = 0.25$, and post-level binaries soar to $r = 0.75$.

Predictive Accuracy Rises Sharply Year over Year

A linear-trend analysis confirms a sharp, year-on-year performance climb. Across the full corpus, models published in 2024 outperformed their 2022–23 predecessors by an average $r = +0.16$ ($\beta = 0.158$, $p < .001$), Accuracy $= +0.09$ ($p < .001$), and F1 $= +0.08$ ($p < .001$). These gains hold after jointly controlling for task type (user vs. post), evaluation level, and baseline status (all $ps < .001$), indicating that the improvement is not a mere artifact of easier datasets or benchmark tuning.

Post-level binaries drive the steepest rise (2024 vs. pre-2024: $t = 5.05$, $p < .001$), yet even user-level classifiers show a significant uplift—Accuracy $t = 3.84$, $p < .001$; F1 $t = 2.03$, $p = .045$. Removing baseline models leaves the 2024 advantage essentially intact for r and Accuracy, though the F1 increment attenuates (n.s.).

A complete set of subgroup coefficients and p-values is provided in Appendix Table A2, confirming that the 2024 performance gain persists across *most* analytic slices; when pooling 2025 with 2024 (post-2024), the advantage attenuates and some post-level effects become non-significant.

Deep Processing Substantially Boosts Performance

Deep processing pays measurable dividends. A study–level meta-regression (with covariates) indicates that adopting deep processing—defined here as any deep-learning or hybrid architecture—explains a non-trivial share of between-study variation (F1: $R^2 = 7.9\%$; accuracy: $R^2 = 2.4\%$). At the aggregate level, deep models outperform shallow ones on both accuracy ($t = 4.11$, $p < .001$) and F1 ($t = 4.64$, $p < .001$).

Gains persist when results are stratified. For user-level classifiers: accuracy ($t = 3.64$, $p < .001$), F1 ($t = 3.77$, $p < .001$), and Pearson r ($t = 3.15$, $p = .010$) all favor deep/hybrid models. For post-level systems: accuracy ($t = 2.98$, $p = .003$), F1 ($t = 3.37$, $p < .001$), and Pearson r ($t = 3.65$, $p < .001$) also show significant improvements.

Crucially, these gains remain after adjusting for task, level, baseline status, platform, sample size, feature family, input modality, and annotation method: $\beta_{\text{Accuracy}} = 0.0498$ ($p < .001$), $\beta_{\text{F1}} = 0.0472$ ($p = .0218$), $\beta_r = 0.0609$ ($p = .0439$).

A complete set of subgroup coefficients and p-values is provided in Appendix Table A3.

Modality Matters—Beyond Disembodied Text

Adding non-text signals—images, audio, or social-graph features—delivers a measurable lift. A random-effects meta-regression using a text-only dummy explains a meaningful share of between-study variance ($R^2 \approx 3.6\%$ for F1; $\approx 2.5\%$ for accuracy). Compared with multimodal systems, text-only models show a deficit of -0.0738 in F1 ($t = -3.50$, $p < .001$) and -0.0453 in accuracy ($t = -2.80$, $p = .005$), while correlation (r) is unaffected ($p = .184$).

Pooled t-tests confirm the pattern: across all studies, multimodal pipelines improve accuracy ($t = 3.15$, $p = .002$) and F1 ($t = 4.87$, $p < .001$). Gains persist at both post-level ($t = 2.89$, $p = .005$ for accuracy; $t = 3.36$, $p = .0013$ for F1) and user-level (F1, $t = 3.81$, $p = .0003$), although the user-level accuracy boost falls just shy of conventional significance ($p = .061$).

In short, even sparse visual or behavioral cues sharpen decision boundaries beyond what "disembodied sentences" can achieve.

Real-World Deployment Gap and XAI

Despite steady gains in benchmark accuracy, social-media depression detectors remain ill-equipped for clinical or public-health use. Our review traces this shortfall to five persistent barriers: (i) scarce psychologically validated labels, (ii) fragile cross-platform transfer, (iii) minimal longitudinal testing, (iv) weak privacy and ethical compliance, and

	Binary framing in ML literature	Continuous framing in clinical practice
Classification form	*Depressed vs. Not-Depressed*	PHQ-9 total score (0–27) graded into mild / moderate / severe
Primary goal	Automated screening, real-time deployment, multimodal fusion	Gauge severity, plan treatment, track intervention efficacy
Data source	Passive social-media	Self-administered questionnaires (PHQ-9, CES-D)
Method / algorithm	ML models—SVM, LSTM, BERT, etc.	Self-reporting questionnaire scoring (no algorithm)
Sample size	Large	Small
Threshold decision	Single cut-off (e.g., PHQ-9 $\geq$ 10 = positive)	Multi-band system: 0–4 none, 5–9 mild, 10–14 moderate, etc.
Follow-up action	App-pushed CBT, auto-referral, continuous monitoring	Initiate therapy, issue sick leave, adjust medication
Real-world use	Fast, large-scale, but low diagnostic authority	Formal diagnosis, insurance claims, workplace adjustments

Table 2: Binary ML framing versus continuous clinical framing of depression.

(v) limited model explainability and reporting loss leading to performance attrition.

Together, these gaps reveal a field optimised for leaderboard scores within narrow data silos rather than for trustworthy, scalable mental-health intervention.

Clinical Label Scarcity & Granularity Gaps—Where Validation Still Falters Although the field has expanded rapidly, psychologically validated evidence remains scarce (see Table 2 for a head-to-head view of machine-learning vs. clinical labeling). User-level studies comprise only 22.4% of the corpus, leaving most work anchored to single posts—fleeting expressions rather than enduring disorders. Standard screening scales and clinician interviews provide just 17.6% of all labels, and models trained on these gold-standard sources perform significantly worse than those built on crowd tags ($t = -0.12$, $p < .001$), underscoring how far current systems fall from clinical reliability.

The gulf in Table 2 is clear: most ML studies still flip a single *depressed/not* switch, while clinicians position patients along a graded continuum and tailor care accordingly. Bridging this gap is not only academic: social-media streams offer unmatched scale, temporal depth, and behavioral nuance. If models could translate those signals into clinically aligned severity scores, they could enable early-warning, relapse monitoring, and personalized interventions. This requires importing clinical standards—multi-level thresholds, validated scales—into modeling pipelines and letting online data enrich rather than replace clinical decision-making.

Granularity also remains limited: about **80%** of models still perform binary classification, while severity grading appears in only **17.5%** of post-level and **13.3%** of user-level studies; cross-disorder differentiation falls to **3.5%** at the post and sentence level. Yet human affect is neither static nor binary. Affective-computing pioneers showed as early as 1997 that a 2D valence–arousal grid cannot distinguish fear from anger (Picard 1997), and psychology has long favored Russell's (Russell 1980) circumflex model treating emotion as *continuous*. Contemporary work likewise argues that dimensional assessments of mental health are "more reliable and valid than categorical diagnoses" (Lahey et al. 2017). Together, these patterns highlight scarce clinical labels and coarse, binary framing as central bottlenecks, motivating our **Psych-Aligned Evaluation Framework** Table 5.

Cross-Platform Fragility Only **5** studies in our corpus attempted cross-platform validation—training on one platform and testing on another—with an average performance drop of **42.5 percentage points in** r and **4.5% in F1** (Yang et al. 2020; Shen et al. 2018; Zhang et al. 2023). The slump is not a statistical fluke but a symptom of deep domain shift. Reddit users write extended, thread-based narratives; Twitter compresses thought into 280-character bursts laced with hashtags; Facebook mixes personal diaries with shared links. These contrasts reshape vocabulary, syntax, and even the emotional cadence of posts, so features that signal depression on Reddit—lengthy rumination, first-person pronouns—become faint or absent on Twitter. Because 83% of all datasets still come from Reddit and Twitter, the field has grown in a platform monoculture, optimizing for Anglophone, text-heavy habits while neglecting other languages, media types, and interaction styles. Without systematic multi-platform validation, models risk performing like dialect experts: fluent in one social vernacular, tongue-tied everywhere else.

Temporal Generalization and Longitudinal Data Only **4** studies in our corpus attempt true longitudinal prediction; the remainder rely on cross-sectional snapshots. Yet depression is intrinsically dynamic, waxing and waning over weeks or months. Models trained on single-time-point data can flag current distress but say little about future risk or relapse, whereas clinical practice routinely builds prognostic models to guide intervention. Longitudinal work is further hampered by sparse posting—many users go silent for days—and by the fact that online expression does not always mirror internal state (Chen 2020). These gaps hinder modeling of coherent mood trajectories and leave current models

Layer	Evaluation Focus	Core Metric (example)	Why It Matters
G1 Ethical Legitimacy	IRB approval; de-identified users; platform TOS followed; bias checks	*EthicsScore (0–5)*	Tools must use data lawfully and fairly to be deployable and trusted.
G2 Reporting Transparency	Sample size; full confusion matrix or AUC + N; shared code & data	*TranspScore (0–4)*	Clear reporting and open materials make findings checkable and reusable.
G3 Cross-Domain Robustness	Performance change across platform, language, or year	*ShiftRatio* (≤ 1)	A reliable detector should work beyond the one dataset it was trained on.
G4 Temporal Stability	Consistency over time; early risk flagging before crises	*RepeatStability* (scores stay stable over weeks), *EarlyWarn* (flags risk before symptom peak)	Depression changes over weeks; models should track the trend, not a single snapshot.
G5 Clinical Fidelity	Error vs. PHQ-9; macro-F1 on DSM-5 symptom labels	Δ*PHQ* / *SymptomF1*	Model outputs should line up with symptom scales clinicians actually use.
G6 Explainability	% of clinician-rated plausible explanations; overlap with DSM-based features	*ClinAgree* %	Human-understandable reasons are crucial for patient safety and regulator trust.

Table 3: Psych-Aligned Evaluation Framework: Six criteria for trustworthy depression detection.

ill-equipped for early prevention.

Data Privacy and Ethical Concerns Privacy is the steepest barrier to real-world deployment. Fewer than **12%** of studies report IRB approval, and only **36%** describe safeguards like anonymisation or compliance with platform policies. Yet we detect *no performance penalty for ethical compliance* ($p > .05$), suggesting that rigor need not dilute accuracy. Most users are unaware their posts may be mined for mental-health inference, raising unresolved questions about informed consent and opt-out rights (Chancellor and Choudhury 2020). Even anonymised corpora can be re-identified due to rich social-media metadata. Without explicit consent, robust privacy engineering, and regulatory alignment (e.g., GDPR, HIPAA), large-scale deployment risks legal and public backlash.

The "Black-Box" Problem Interpretability remains an outlier: only **three** studies in our corpus apply any form of explainable AI. Most models—particularly transformer-based—offer high accuracy but opaque reasoning, a fatal flaw in clinical contexts where practitioners must justify decisions to patients and regulators. Early work with SHAP values and attention maps shows that transparency improves clinician trust, yet such methods are rarely adopted(Tonekaboni et al. 2019). Until depression detectors routinely expose the features and logic driving each prediction—and align those explanations with clinical theory—their bedside utility will stay largely theoretical.

Reporting Loss and Performance Attrition Evidence synthesis is hamstrung by sparse reporting. Only **27.4%** of model entries supply both sample size and a complete confusion matrix (or AUC + N), the minimum needed to convert metrics into a common effect size r. Most others provide headline accuracy alone or omit denominators, making quantitative pooling and independent verification of robustness impossible. Until the field adopts PROBAST-style reporting standards for machine-learning studies(Moons et al. 2019), missing data will remain a deployment barrier alongside privacy and explainability. These gaps motivate our **Psych-Aligned Evaluation Framework**, which scores models on clinical validity, temporal stability, ethical compliance, and interpretability—not accuracy alone.

Psych-Aligned Evaluation Framework

Diagnostic accuracy is necessary but not sufficient: real-world deployment also demands reproducibility, context-robustness, clinical alignment, ethical soundness, and explainability. Guided by these requirements, we propose a six-layer *Psych-Aligned Evaluation Framework* (PAEF), a checklist for translating models from leaderboard success to bedside utility (Table 3). The PAEF reframes evaluation from "How high is the accuracy?" to "Why will this model succeed—or fail—in a sensitive, socially situated task?"

By foregrounding ethics, transparency, domain shift, temporal drift, clinical alignment, and interpretability, the framework echoes recent calls to move beyond static benchmarks toward AI systems tested in the wild that are reliable, equitable, and human-centered.

Conclusion

Overall, our synthesis shows that recent gains in social-media depression detection stem more from deeper architectures and richer modalities than from bigger datasets alone—but those gains concentrate on narrow, post-level tasks and fade for clinically nuanced targets. By mapping these findings onto our four framing pillars—real-world generalization, human-centered moderators, benchmark reform, and translational blueprints—we highlight the path from leaderboard success to trustworthy mental-health AI.

The *Psych-Aligned Evaluation Framework* distils that path into actionable checkpoints on ethics, stability, clinical fidelity, and explainability, offering researchers and practitioners a shared yardstick for genuinely human-centered deployment.

References

Amanat, A.; Rizwan, M.; Javed, A. R.; Abdelhaq, M.; Alsaqour, R.; Pandya, S.; and Uddin, M. 2024. Deep learning

for depression detection from textual data. *Journal of Network and Automation Online*, 15(1): 437–450.

Borenstein, M.; Hedges, L. V.; Higgins, J. P. T.; and Rothstein, H. R. 2009. *Introduction to Meta-Analysis*. Wiley.

Chancellor, S.; and Choudhury, M. D. 2020. Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review. *Current Opinion in Psychology*, 36: 79–84.

Chancellor, S.; and De Choudhury, M. 2020. Methods in Predictive Psychiatry with Social Media. *npj Digital Medicine*.

Chen, L. 2020. Overview of clinical prediction models. *Annals of Translational Medicine*, 8(4): 71.

El-Sappagh, S.; Al Shorman, O.; and Abdelrazek, S. 2025. Cross-Platform Challenges in Depression Detection. *Journal of Depression Research*.

Ferrario, A.; Bianchi, L.; and Rossi, F. 2024. The Role of Humanization and Robustness of Large Language Models in Conversational Artificial Intelligence for Individuals With Depression: A Critical Analysis. *JMIR Mental Health*.

Hanley, J. A.; and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1): 29–36.

Ibrahimov, K.; Aliev, F.; and Gasimov, R. 2024. Explainable AI for Mental Disorder Detection via Social Media: A survey and outlook. *arXiv preprint*.

Jain, S.; Kuppili, P. P.; Pattanayak, R. D.; and Sagar, R. 2018. Ethics in Psychiatric Research: Issues and Recommendations. *BMC Psychiatry*, 18: 206.

Jin, N.; Ye, R.; and Li, P. 2024. Diagnosis of Depression Based on Facial Multimodal Data. *Frontiers in Psychology*, 15: 10877283.

Lahey, B. B.; Krueger, R. F.; Rathouz, P. J.; Waldman, I. D.; and Zald, D. H. 2017. A hierarchical causal taxonomy of psychopathology across the life span. *Psychological Bulletin*, 143(2): 142–186.

Lipsey, M. W.; and Wilson, D. B. 2001. *Practical Meta-Analysis*. Sage.

Liu, Y.; Depp, C.; Jeste, D. V.; and Kim, H. C. 2022. Advancing computational methods for mental health diagnosis and treatment: Systematic review and future directions. *JMIR Mental Health*, 9(4): e36164.

Lyu, S.; Ren, X.; Du, Y.; and Zhao, N. 2023. Detecting depression of Chinese microblog users via text analysis: Combining Linguistic Inquiry Word Count (LIWC) with culture and suicide related lexicons. *Frontiers in Psychiatry*, 14: 1121583.

Moons, K. G. M.; Wolff, R. F.; Riley, R. D.; Whiting, P. F.; Westwood, M.; Collins, G. S.; Reitsma, J. B.; Kleijnen, J.; and Mallett, S. 2019. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine*, 170(1): 51–58.

Padmaja, S. M.; Godla, S. R.; Ramesh, J. V. N.; Muniyandy, E.; Sridevi, P.; El-Ebiary, Y. A. B.; and Devadhas, D. N. P. 2025. Depression Detection in Social Media Using NLP and Hybrid Deep Learning Models. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 16(2).

Phiri, D.; Makowa, F.; Amelia, V. L.; Phiri, Y. V. A.; Dlamini, L. P.; and Chung, M.-H. 2025. Detecting Depression on Social Media: A Review. *JMIR*. In press.

Picard, R. W. 1997. *Affective Computing*. MIT Press.

Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6): 1161–1178.

Salas-Zárate, M. d. P.; Valencia-García, R.; and García-Díaz, J. A. 2022. Depression Detection in Social Media. *Healthcare*, 10(2): 291.

Shen, T.; Jia, J.; Shen, G.; Feng, F.; He, X.; Luan, H.; Tang, J.; Tiropanis, T.; Chua, T.-S.; and Hall, W. 2018. Cross-Domain Depression Detection via Harvesting Social Media. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 1600–1606. AAAI Press.

Tonekaboni, S.; Joshi, S.; McCradden, M. D.; and Goldenberg, A. 2019. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In *Proceedings of the 2019 Machine Learning for Healthcare Conference (MLHC 2019)*, volume 106 of *Proceedings of Machine Learning Research*, 359–380.

Trifu, R. N.; Nemeș, B.; Herta, D. C.; Bodea-Hategan, C.; Talaș, D. A.; and Coman, H. 2024. Linguistic Markers for Major Depressive Disorder: A Cross-Sectional Study Using an Automated Procedure. *Frontiers in Psychology*, 15: 1355734.

Trotzek, M.; Koitka, S.; and Friedrich, C. M. 2018. Multimodal Approaches for Social Media Depression Detection. *arXiv preprint*.

World Health Organization. 2022. *World Mental Health Report: Transforming Mental Health for All*. Geneva: World Health Organization.

Yang, S.; Cui, L.; Wang, L.; Wang, T.; and You, J. 2024. Enhancing Multimodal Depression Diagnosis Through Representation Learning and Knowledge Transfer. *Frontiers in Psychiatry*.

Yang, X.; Gao, S.; Wang, T.; Yang, B.; Dang, N.; and Ye, K. 2020. gCAnno: a graph-based single cell type annotation method. *BMC Genomics*, 21(1): 823.

Zhang, W.; Xie, J.; Liu, X.; and Zhang, Z. 2023. Depression Detection Using Digital Traces on Social Media: A Knowledge-aware Deep Learning Approach. *arXiv preprint arXiv:2303.05389*.

Zhang, X.; Li, C.; Chen, W.; Zheng, J.; and Li, F. 2025. Optimizing Depression Detection in Clinical Doctor-Patient Interviews Using a Multi-Instance Learning Framework. *Scientific Reports*, 15: 90117.