Computational Analysis of Expressive Behavior in Clinical Assessment

Jeffrey M. Girard,[1] Dasha A. Yermol,[1] Albert Ali Salah,[2] and Jeffrey F. Cohn[3,4]

[1]Department of Psychology, University of Kansas, Lawrence, Kansas, USA; email: jmgirard@ku.edu
[2]Department of Information and Computing Sciences, Utrecht University, Utrecht, the Netherlands
[3]Departments of Psychology and Intelligent Systems, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[4]Deliberate AI, New York, New York, USA

Annu. Rev. Clin. Psychol. 2025. AA:1–29

https://doi.org/10.1146/annurev-clinpsy-081423-024140

Keywords

clinical assessment, expressive behavior, computational methods, computer vision, speech signal processing, natural language processing

Abstract

Clinical psychological assessment often relies on self-report, interviews, and behavioral observation, methods that pose challenges for reliability, validity, and scalability. Computational approaches offer new opportunities to analyze expressive behavior (e.g., facial expressions, vocal prosody, and language use) with greater precision and efficiency. This paper provides an accessible conceptual framework for understanding how methods from computer vision, speech signal processing, and natural language processing can enhance clinical assessment. We outline the goals, frameworks, and methods of both clinical and computational approaches, and present an illustrative review of interdisciplinary research applying these techniques across a range of mental health conditions. We also examine key challenges related to data quality, measurement, interdisciplinarity, and ethics. Finally, we highlight future directions for building systems that are robust, interpretable, and clinically meaningful. This review is intended to support dialogue between clinical and computational communities and to guide ongoing research and development at their intersection.

Contents

1. Introduction

In the absence of reliable and accessible biomarkers for most psychiatric conditions (Aftab & Sharma 2021, Venkatasubramanian & Keshavan 2016), clinical assessment continues to rely heavily on client self-reports, clinical interviews, and informal behavioral observations (Meyer et al. 2001). However, there are numerous threats to the reliability, validity, and scalability of each of these assessment procedures.

Client self-report is a valuable source of phenomenological information and is relatively inexpensive to collect (Stone et al. 2000). However, it can be limited by clients' reading ability, idiosyncratic understanding of items and response options, reactivity to being assessed, and level of insight into their own functioning (Althubaiti 2016). Client report can be easily scaled across large numbers of clients with computerized administration and scoring. It can also be scaled over time (i.e., longitudinally) using ecological momentary assessment (Shiffman et al. 2008). However, care must be taken as too frequent or extensive administration can be burdensome for clients and may lead them to disengage from assessment.

When paired with standardized protocols, clinical interviews tend to be more reliable and less dependent on client understanding (Meyer et al. 2001). However, maintaining these benefits requires substantial and ongoing clinician training. Interviews are also more time-intensive and costly to administer, which limits their feasibility and scalability.

Finally, clinician observation of clients' behavior (e.g., facial expressions, speech, thought processes, and motor activity) has long been a staple of clinical assessment, offering valuable insights into affective, cognitive, and social functioning (Bunney & Hamburg 1963, Norris et al. 2016). As with interviews, however, these observations are difficult to quantify reliably without formal tools and training, and contextual factors may complicate interpretation (Kirmayer 2005, Barrett 2022). Manual approaches also share the same scalability challenges, given their reliance on clinician time and judgment.

To address these issues, computational approaches can be used to improve the reliability, validity, and scalability of observational assessments of clients' **expressive behavior**, i.e., their observable verbal and nonverbal behaviors that communicate internal states. Recent advances in computer science, affective computing, and computational linguistics have made it possible to estimate clients' behavioral patterns with greater accuracy and efficiency, including body and face movements, vocal prosody and dynamics, and linguistic content and style. Compared to manual methods, these approaches provide more consistent and fine-grained measurement, and their capacity to combine scale with methodological rigor enables broader, more adaptive, and contextually responsive forms of assessment.

In this paper, we aim to provide an accessible conceptual framework for understanding how computational tools can enhance the clinical assessment of expressive behavior. Rather than offering a technical deep dive, our goal is to clarify key ideas, highlight points of intersection between clinical and computational perspectives, and organize recent advances into an intuitive structure. We begin with an overview of the goals, frameworks, and methods that characterize assessment in clinical psychology. Next, we review how techniques from computer vision, speech signal processing, and natural language processing are being used to analyze expressive behavior. We then present an illustrative review of recent research that has applied these approaches in clinical settings, highlighting their utility and current limitations. Finally, we discuss the promises and challenges of integrating these methods into clinical research and practice. This paper is intended for an interdisciplinary audience, including clinical scientists seeking to understand emerging computational methods and computer scientists interested in mental health applications.

2. Clinical Assessment

Clinical assessment is the process of systematically gathering and interpreting information about an individual's symptoms and functioning, often to inform and guide treatment. Clients' affect, behavior, and cognition are of foremost interest in clinical psychology and psychiatry, as they constitute the core components of mental health.

2.1. Assessment Goals

Although terminology may differ across disciplines, the primary goals of assessment in clinical practice can be usefully organized into four broad objectives: screening, diagnosis, prognosis, and monitoring. We briefly review each of these core objectives and then describe several additional goals that are often emphasized in more comprehensive assessments.

First, assessment may aim to quickly identify individuals who are at elevated risk for psychopathology and may benefit from further evaluation, supportive resources, or referral to treatment. This goal, known as **screening**, is common in primary care, schools, and workplace wellness programs (e.g., Joseph & Hermann 1998, Kim et al. 2022, Søvold et al. 2021). Screening tools are typically designed for brevity over precision and are often calibrated to be overly inclusive, reducing the risk of missing someone in need.

Screening: An assessment goal focused on quickly identifying people at risk who may need further evaluation.

Second, assessment may aim to confirm the presence or severity of psychopathology by

Diagnosis: An assessment goal focused on identifying specific conditions by confirming their presence and excluding other possible explanations.

integrating more comprehensive information and ruling out alternative explanations (e.g., Spitzer 1983, Fernandes et al. 2017). These **diagnostic** assessments are widely used in clinical practice (e.g., to guide case conceptualization and treatment planning), by insurers (e.g., to determine reimbursement eligibility), and in legal contexts (e.g., for disability claims or litigation). Although they require more time and training than screening tools, diagnostic assessments are essential in high-stakes contexts.

Prognosis: An assessment goal focused on predicting future outcomes and the course of a person's condition over time.

Third, assessment may aim to predict a person's future outcomes and how their condition is likely to change over time (e.g., Croft et al. 2015, Fusar-Poli et al. 2018). These **prognostic** assessments are often used in clinical settings to guide decisions about treatment type, intensity, and how to involve clients and families in care planning. They also inform follow-up care, such as how often a client should be re-assessed. Prognostic models typically draw on risk and protective factors, which help clarify the conditions that shape recovery and long-term outcomes.

Monitoring: An assessment goal focused on tracking changes in a person's condition over time, such as during or after treatment.

Finally, assessment may aim to track how a person's condition is actually changing over time (e.g., Boswell et al. 2015, Malasinghe et al. 2019). This goal is often referred to as **monitoring** and is central to modern, data-informed approaches to care (Lewis et al. 2019). Monitoring can help evaluate whether treatment is producing the desired effects, detect early warning signs of symptom relapse, and identify fluctuations in mental health that correspond to life events, stressors, or environmental changes. When implemented effectively, monitoring can enhance shared decision-making, improve treatment precision, and support more responsive care. Moreover, it provides the longitudinal data needed to test and refine prognostic models by comparing predicted outcomes with actual trajectories.

During an in-depth assessment, clinicians may also strive to: evaluate the individual's risk of harm to self and others (e.g., Jobes 2023, Singh et al. 2011), identify the individual's strengths and resources for coping with condition-related stress (Tedeschi & Kilmer 2005), understand cultural and contextual influences on the individual's behaviors and functioning (Ryder et al. 2011, Kraemer et al. 2003), and build a collaborative relationship with the individual characterized by mutual trust, empathy, and respect (e.g., Hilsenroth et al. 2004, Rakel 2016). There are even strategies for adding clinical interventions to the procedure, thus creating "therapeutic assessments" (e.g., Finn & Tonsager 1997, Finn 2007).

2.2. Assessment Frameworks

Given that clinical assessment seeks to describe clients' symptoms and functioning, it is inherently linked to *nosology*—the systems used to define and organize these phenomena. Yet, how best to construct a nosological framework that accurately captures such patterns of psychological disturbance remains a topic of ongoing debate (First 2015, Rief et al. 2023). Thus, we briefly review the primary nosological approaches currently in use.

The first, and currently dominant, approach organizes psychopathology into discrete "disorders" or "syndromes" that are defined by specific criteria (e.g., obsessive-compulsive disorder and schizophrenia). Disorders are viewed as separate disease entities that are different in kind from normal functioning. Examples include the Diagnostic and Statistical Manual of Mental Disorders (**DSM**; American Psychiatric Association 2022) and the International Classification of Diseases (**ICD**; World Health Organization 2019). Assessment under this approach involves determining the presence or absence of diagnostic criteria and then assigning categorical diagnostic labels.

A second approach is that of the Research Domain Criteria (**RDoC**; Insel et al.

Table 1 Overview of Assessment Frameworks

Framework	Development	Objective	Conceptual Units
DSM & ICD	Top-Down	Descriptive	Diagnostic Categories
RDoC	Top-Down	Mechanistic	Functional Dimensions
HiTOP	Bottom-Up	Descriptive	Hierarchical Dimensions
Network Theory	Bottom-Up	Mechanistic	Causal Interactions

2010, Cuthbert & Insel 2013, Cuthbert 2022), which characterizes psychopathology in terms of transdiagnostic domains of behavior and functioning (e.g., cognitive, social, and arousal/modulatory systems). It emphasizes neurobiological mechanisms (e.g., genes, molecules, and circuits) and argues that psychopathology differs only in degree from normal functioning. Although this approach is intended to guide research rather than clinical practice, assessment would involve collecting biological and behavioral data and integrating them to assign dimensional scores on domains and constructs (Insel 2014).

A third approach is that of the Hierarchical Taxonomy of Psychopathology (**HiTOP**; Kotov et al. 2017, Krueger et al. 2018), which organizes maladaptive traits and symptoms into a hierarchy of dimensions with varying degrees of specificity (e.g., performance anxiety is a specific symptom within the fear subfactor of the internalizing spectrum). HiTOP also argues that psychopathology differs only in degree from normal functioning; however, it adopts an empirical, quantitative method to define its dimensions and focuses more on the observed co-occurrence of symptoms than underlying mechanisms. Assessment under this approach involves measuring various traits and symptoms and then calculating a profile of dimensional scale scores at each level of the hierarchy (Simms et al. 2022).

Finally, the emerging **Network Theory** argues that, rather than being the effects of underlying diseases, psychiatric symptoms cause each other (Borsboom 2008, 2017, Cramer et al. 2010). Mental disorders, then, emerge from the interconnected network of causal interactions between symptoms (e.g., delusion causes paranoia, which causes hostility and withdrawal, which together cause social isolation). The structure and dynamics of such networks can be described using specialized statistical techniques and interrogated to derive clinical insights (e.g., which symptoms are central to the network). Assessment under this approach involves identifying which symptoms are present and which network interactions sustain them; the approach is agnostic to exactly how the symptoms are represented.

Table 1 characterizes each approach by answering the following questions. Was the framework developed in a primarily top-down (i.e., consensus-based) or bottom-up (i.e., data-driven) manner? Does the framework primarily seek to describe psychopathology or to uncover its underlying mechanisms? What type of conceptual units does it propose?

2.3. Assessment Methods

Numerous methods can be used to gather information during clinical assessment, each with its own strengths and limitations. In practice, the most informative evaluations draw on multiple sources of evidence (Spitzer 1983). Here, we highlight the core methods most frequently employed in everyday clinical practice across diverse populations.

First, assessment may include clinicians' **observational ratings** (Bunney & Hamburg 1963, Girard & Cohn 2016, Norris et al. 2016) of the client's expressive behavior (e.g., gestures, expressions, gaze, speech, posture, and motion). In inpatient and residential settings, clinicians may have the opportunity to observe such behaviors from afar and across different contexts. In consulting and outpatient settings, opportunities may be limited to specific contexts such as assessment interviews and therapy. Observation can yield rich naturalistic data, can easily incorporate contextual information, and does not require clients to have insight into their own conditions. However, it can be time-consuming and relies on clinical judgment, which is often unreliable without standardized rating scales and adequate training. Another challenge is that some constructs are difficult for clinicians to precisely quantify, even with training (e.g., tangential thinking or psychomotor slowing).

Second, assessment may include **questionnaire ratings** (Stone et al. 2000) completed by clients or informants (e.g., the clients' family or close others). Such questionnaires often include items about the presence, frequency, and severity of various symptoms but may also assess levels of functioning, disability, coping potential, and other relevant constructs. Brief questionnaires are commonly used in screening and monitoring for many forms of psychopathology, and longer questionnaires play a vital role in the diagnosis and prognosis of certain conditions (e.g., personality disorders, developmental disorders, and childhood behavioral problems). Questionnaires are relatively inexpensive and quick to administer and interpret, and they capture the valuable perspectives of clients and informants in a standardized format. However, questionnaire items can be misunderstood, responses may be influenced by various biases, and accurate completion often requires a high degree of insight from the respondent (Althubaiti 2016).

Third, assessment may include a **psychiatric interview** in which clients are directly asked questions and observed by a trained interviewer (Silberman et al. 2015). Interviews may be more structured, using pre-determined questions to ensure consistency across clients, or less structured, with flexible questions tailored to the individual. They may also vary in focus; some emphasize factual information such as biographical or medical history, while others explore the client's personal interpretations and emotional reactions. Structured, fact-oriented interviews often yield reliable quantitative data, such as diagnostic labels or severity scores, whereas less structured, feeling-oriented interviews tend to produce rich qualitative data, including chief complaints and personal histories. Interviews can easily incorporate clarifying and contextual information, integrating both observational ratings and client-reported experiences. However, they typically require more time and training to administer than questionnaires, making them harder to implement and scale in many clinical and research settings.

Finally, assessment may include **structured tasks** (Haynes & O'Brien 2000, Kipps & Hodges 2005) designed to reveal patterns, processes, or deficits in behavior within a controlled environment. These include cognitive tests that assess attention and memory, social interaction tasks that evaluate communication skills and empathy, emotion elicitation tasks that probe reactivity and avoidance, and projective tests intended to surface unconscious fears or desires. Some tasks (e.g., many cognitive tests) have regimented scoring procedures, while others (e.g., many projective tests) are interpreted in a more creative and subjective manner. Structured tasks are generally most effective when highly specific and closely tied to observable symptoms. However, their artificial nature can shape client behavior in ways that limit how well findings generalize to real-world settings.

Each assessment method offers distinct strengths and limitations, often requiring trade-offs between capturing behavior in naturalistic settings (ecological validity) and obtaining tightly controlled, reliable measurements (precision). Emerging computational approaches may help reconcile these trade-offs by increasing the volume and quality of data that can be captured in both structured and real-world environments.

3. Computational Analysis

Computational analysis refers broadly to the use of algorithms and data-driven models to identify, quantify, and interpret patterns in complex data. In the context of clinical assessment, it is increasingly used to analyze expressive behavior, such as facial expressions, vocal signals, and language use, with the goal of generating insights about psychological states and interpersonal dynamics. These methods complement traditional assessment by enabling scalable, objective, and fine-grained measurement of behavior. In this section, we describe the core goals of computational analysis, the frameworks used to operationalize these goals, and the methods by which expressive behavior is measured and modeled.

3.1. Computational Goals

As we did for clinical assessment, the primary goals of computational analysis can be usefully organized into four broad objectives: prediction, explanation, discovery, and generation. These goals guide how data are modeled, what types of features are prioritized, and how the outputs of computational systems are interpreted and used in practice.

First, **prediction** refers to the use of computational models to estimate a target variable (i.e., *label*) based on observed data, whether from the past, present, or future. Predicting categorical labels is called "classification" and predicting continuous ones is "regression." In the analysis of expressive behavior, prediction can involve estimating expressive signals from raw inputs or using expressive features to predict clinically relevant outcomes. For example, models may aim to infer an individual's emotional state, symptom severity, or risk of relapse based on their observed behavior. Predictive modeling in this context often prioritizes accuracy and generalizability, with success typically evaluated based on how well predictions match known labels in new or unseen data (Yarkoni & Westfall 2017).

Prediction: A computational goal focused on estimating unknown outcomes from observed data with an emphasis on accuracy and generalization.

Second, **explanation** refers to the use of computational models to clarify the mechanisms or processes that give rise to observed behavior. This goal prioritizes interpretability, theory testing, and causal insight over raw predictive accuracy (Gelman et al. 2014, Pearl et al. 2016). In the analysis of expressive behavior, explanatory models aim to shed light on why certain behaviors emerge, what psychological constructs they reflect, and how they vary across individuals or situations. Success is typically evaluated based on how well the model aligns with theoretical expectations, supports causal inference, or enhances conceptual understanding, rather than how accurately it predicts new observations. The goal is to produce meaningful, interpretable insights that advance scientific understanding, even if the models are less complex or less accurate than those built primarily for prediction.

Explanation: A computational goal focused on clarifying the processes that produce observed data, prioritizing interpretability and causal insight.

Third, **discovery** refers to the use of computational models to uncover unexpected patterns, structures, or relationships in data. Unlike prediction or explanation, discovery is often exploratory and hypothesis-generating, which is particularly useful in high-dimensional or temporally complex behavioral data (Box 1976). In the analysis of expressive behavior, this may involve identifying recurring patterns across individuals, time points, or contexts that had not been theorized in advance. These approaches rely on data structure rather than prior assumptions, enabling flexible exploration of variability within and across individuals. Discovery can reveal new behavioral dimensions that inform theory and nosology (Bzdok & Meyer-Lindenberg 2018).

Discovery: A computational goal focused on uncovering novel patterns or structures in data without predefined hypotheses.

Finally, **generation** refers to the use of computational models to create new content or simulate phenomena based on patterns learned from data. Rather than merely describing or predicting observed behavior, generative approaches aim to produce novel outputs that reflect or respond to real-world psychological, social, or communicative processes. Recent advances have enabled the highly realistic synthesis of expressive behaviors, including facial expressions, vocal signals, and language patterns (Ma et al. 2025). At the same time, generative models can also formalize psychological theories by simulating the mechanisms that give rise to observed behaviors (Haines et al. 2025). Generative models thus serve both practical and theoretical aims: they can synthesize expressive behaviors for use in clinical training, adaptive assessment, and decision support, and they can simulate psychological processes to support theory development (Shortliffe & Sepúlveda 2018).

Generation: A computational goal focused on creating synthetic data that mimic real-world behaviors or phenomena.

3.2. Computational Frameworks

Three computational frameworks are commonly used to quantify expressive behavior. Each is designed to operate on a distinct modality of input data (i.e., video, audio, or text) and applies specialized techniques to extract, model, and interpret behavioral signals. Before discussing specific methods or research findings when applied to clinical assessment, we provide a brief overview of each framework and the behavioral signals it captures.

Computer Vision: A computational framework for analyzing visual information from images or video, including the quantification of expressive behavior.

First, **computer vision** (CV) is a computational framework for analyzing the data captured by a camera in order to detect, track, and interpret visual patterns (Szeliski 2022). When applied to human behavior, these systems quantify nonverbal cues in the face and body that reflect affective valence, attentional focus, motor coordination, and other key psychological processes. Examples of such cues include: (1) *Facial behaviors,* such as smiling, frowning, and brow raising; (2) *Ocular cues,* such as gaze direction, pupil dilation, and tearing; (3) *Head and body movements,* such as nodding, fidgeting, and psychomotor slowing; (4) *Postural signals,* such as slumping, leaning away, and crossing the arms; and (5) *Gestures,* such as pointing, shrugging, and self-touching.

Speech Signal Processing: A computational framework for extracting and analyzing vocal features from audio recordings of speech.

Next, **speech signal processing** (SSP) is a computational framework for analyzing data captured by a microphone in order to detect and quantify vocal patterns (Tan & Jiang 2018). When applied to human behavior, these systems prioritize the manner of speech over its content, quantifying nonverbal cues that reflect affective intensity, cognitive load, physiological arousal, etc. Examples of such cues include: (1) *Prosodic features,* such as pitch, loudness, and intonation; (2) *Voice quality,* such as breathiness, tension, and hoarseness; (3) *Temporal dynamics,* such as speech rate, pauses, and response latency; (4) *Articulatory-phonetic features,* such as precision, disfluencies, and phonetic variation; and (5) *Nonverbal vocalizations,* such as sighing, laughing, and groaning.

Natural Language Processing: A computational framework for analyzing the content and structure of spoken or written language.

Finally, **natural language processing** (NLP) is a computational framework for analyzing data captured through writing or speech transcription in order to detect and interpret patterns of language use (Jurafsky & Martin 2025). When applied to human behavior, these systems quantify linguistic features that reflect emotional content, cognitive style, communicative intent, etc. Examples of such cues include: (1) *Lexical choice,* such as the use of affective words, self-referential terms, and absolutist expressions; (2) *Syntactic structure,* such as sentence complexity, fluency, and grammatical accuracy; (3) *Semantic coherence,* such as topic relevance, consistency, and conceptual linkage; (4) *Pragmatic markers,* such as emotional tone, sarcasm, and social appropriateness; and (5) *Discourse organization,* such as narrative structure, turn-taking patterns, and referential clarity.

These frameworks offer complementary insights into expressive behavior by analyzing distinct but interrelated channels of communication. Yet relying on a single modality may miss both valuable information available in other channels and the richer meanings that emerge when modalities interact. **Multimodal approaches** seek to combine information from multiple modalities to improve representation, address data gaps, and enable better learning (Liang et al. 2024). Interactions between modalities can be especially informative for interpreting mixed emotional states or resolving ambiguity in social intent.

Multimodal approaches: Modeling approaches that combine multiple modalities, such as visual, audio, and text signals.

3.3. Computational Methods

To extract meaningful information from records of expressive behavior, each computational framework relies on machine learning methods tailored to the properties of its input data. This section outlines how these methods are used across video, audio, and text data, distinguishing between traditional approaches based on theory-derived features and more recent approaches based on deep learning. We also highlight advances in multimodal fusion, where information from multiple channels is combined to support richer, more robust models.

Machine Learning: Computational methods that optimize a performance criterion, typically for prediction or pattern discovery, using example data from the domain.

3.3.1. Machine learning methods.

Across computational frameworks, expressive behaviors are analyzed using statistical and **machine learning** methods that learn patterns from data in service of the goals described in subsection 3.1. Statistical methods are often favored for explanation, given their emphasis on hypothesis testing and the quantification of uncertainty, whereas machine learning methods are increasingly preferred for tasks such as prediction, discovery, and generation (Breiman 2001). Although new "interpretable" machine learning techniques hold promise, the field remains in active development and is fraught with conceptual ambiguities and risks of misinterpretation (Henninger et al. 2025).

Traditional approaches to machine learning analyze "hand-crafted features," which are measurable properties of the data specified by researchers based on theory and domain knowledge. For the analysis of expressive behavior, these features might include landmark positions, pitch contours, or linguistic word counts. While researchers decide which features to include, their relative importance and interactions are learned from the data using various algorithms. When labeled examples are available during training (e.g., images annotated as "smile" or "non-smile" for classification tasks or patient records with symptom severity rated on a continuous scale for regression tasks), *supervised learning* algorithms such as support vector machines or random forests can learn to predict behavioral constructs or clinical outcomes in new data using only the features. In the absence of labeled examples, *unsupervised learning* algorithms such as k-means clustering and non-negative matrix factorization can uncover latent patterns or groupings in the features, supporting the discovery of novel subtypes or behavioral profiles in the data (Hastie et al. 2009).

Deep learning methods, which underlie most modern artificial intelligence (AI) tools, learn abstract, hierarchical representations of data directly from raw or minimally processed input (Goodfellow et al. 2016). Relying less on expert human knowledge and hand-crafted features, these models discover and optimize their own internal features during training, often capturing complex patterns that are difficult to specify manually (e.g., nonlinear dependencies across time or modality). Architectures such as convolutional neural networks, recurrent neural networks, and transformers have demonstrated strong performance across a range of relevant tasks (LeCun et al. 2015). These models are especially powerful for large-scale prediction and generation, and are increasingly used in discovery-oriented work. Yet

Deep learning: Computational methods using multilayer neural networks to learn progressively more complex representations from raw data, typically for prediction or generation.

Transfer learning: Computational methods that use knowledge from one task or domain to improve performance on another.

their reliance on large labeled datasets and high computational demands can be prohibitive, particularly for behavioral and clinical applications where such resources are limited.

One promising solution to these challenges is **transfer learning**, a general paradigm for leveraging knowledge from one *task* (i.e., computational goal) or *domain* (i.e., type or source of input data) to improve learning in another (Pan & Yang 2010, Weiss et al. 2016). This approach is especially useful in fields like psychology, where data and labels are often scarce or expensive to obtain. One common transfer learning strategy is to pre-train a model on a large, general-purpose dataset (e.g., millions of images labeled for whether they depict a person) and then refine that training using a smaller, task-specific dataset involving the same type of input (e.g., hundreds of images labeled for whether they depict a smiling face), a process known as *fine-tuning*. Another strategy is *multi-task learning*, in which a model is trained on several related tasks simultaneously (e.g., detecting both anxiety and depression from the same speech recordings). This approach can improve performance by helping the model identify features that are shared across conditions as well as those that are distinct, facilitating more accurate and nuanced differentiation. A third approach, *domain adaptation*, is used when the task remains the same but the domain differs, e.g., adapting a model trained to detect psychological distress during clinical interviews, where speech is formal and guided by a fixed protocol, to perform well on naturalistic conversations, which are more spontaneous and variable in language and tone.

3.3.2. Computer vision methods. Analysis of expressive behaviors in the visual modality is usually based on pixel-based representations of images (e.g., video frames). Over time, methods have evolved from simple and interpretable hand-crafted features, to intermediate representations based on models of facial or bodily structure, to more complex and powerful but less transparent deep learning approaches.

Hand-crafted visual features include properties such as shape, texture, and motion, derived using manually defined descriptors grounded in perceptual theory (Palmer 1999). Examples include the locations of landmark points such as the outlines of the mouth and eyes (Kazemi & Sullivan 2014), appearance descriptors like Gabor wavelets and local binary patterns (Tian et al. 2002, Zhao & Pietikainen 2007), and motion estimators such as optical flow (Horn & Schunck 1981). These approaches are effective in controlled settings but are sensitive to variation in lighting, head pose, and identity.

Model-based representations in CV use parametric models to track the shape and configuration of expressive anatomy over time. Techniques such as 3D morphable models (Blanz & Vetter 1999), active appearance models (Cootes et al. 2001), and skeletal pose estimators (Loper et al. 2015) fit deformable templates to faces, hands, or bodies. These methods incorporate domain knowledge (e.g., anatomical constraints) and can perform well on small datasets, but often struggle in less structured environments. Model-based approaches to facial expression analysis have been strongly influenced by Facial Action Coding System (Ekman et al. 2002), an anatomically grounded system for describing facial movements in terms of underlying muscle actions. It has guided both manual labeling and the development of model-driven features used in computational analysis (Cohn & Ekman 2005).

Deep learning approaches in CV learn visual patterns directly from raw image data. These models have shown strong performance in tasks such as expression recognition (Li & Deng 2022), gaze estimation (Cheng et al. 2024), and pose tracking (Zheng et al. 2023), often learning abstract features that generalize across varied conditions. However, their reliance on large labeled datasets presents a persistent challenge in behavioral research. To address this, recent work has leveraged motion-capture datasets and 3D model-based data synthesis to generate large-scale training data with automatically produced pose annotations (Zheng et al. 2023). These strategies reduce the burden of manual labeling and expand the diversity of training data. In parallel, advances in deep learning architectures' internal representations, such as graph convolutional neural networks and multi-task learning paradigms, have increased accuracy and robustness (e.g., Hu et al. 2025).

3.3.3. Speech signal processing methods. Analysis of expressive behavior in the audio modality is usually based on raw audio waveforms or derived time-frequency representations. Over time, these methods have evolved from simple and interpretable hand-crafted features, to intermediate models grounded in theories of speech production and prosody, to more flexible but less transparent deep learning approaches.

Hand-crafted vocal features aim to summarize variation in pitch, loudness, timing, and spectral characteristics of the voice. Commonly used acoustic-prosodic features include fundamental frequency, energy, speech rate, pause structure, jitter, shimmer, harmonics-to-noise ratio, formants, and mel-frequency cepstral coefficients (Eyben 2016, Eyben et al. 2016). These features are typically extracted from very short time windows, providing a moment-to-moment account of vocal dynamics. Known as "low-level descriptors" (LLDs), they are usually aggregated over time into summary statistics called "high-level descriptors" (HLDs), which serve as input to machine learning models. These features are widely used due to their efficiency and interpretability, but can be sensitive to differences in speaker identity, language, or recording quality (Low et al. 2020).

Model-based representations in SSP draw on theories of speech production and prosody to explain how acoustic signals arise from underlying physiological or phonetic and articulatory processes. Unlike hand-crafted features, which summarize surface-level signal properties, these models aim to capture the underlying generative mechanisms of speech. The source–filter model, for example, separates the speech signal into a glottal source and vocal tract filter, providing a general framework for analyzing speech production (Fant 1971). It supports a range of applications, including the analysis of voice quality characteristics such as breathiness or strain (Ladefoged & Johnson 2014). Other approaches use rule-based or statistical models to represent prosodic structures, such as pitch contours, rhythm, stress, and timing, that are linked to emphasis, emotion, or conversational dynamics (e.g., Xu 2013, Scherer 2003). These methods are grounded in theory and offer clear interpretability but can be difficult to scale and may require language-specific customization.

Deep learning approaches in SSP have advanced tasks such as emotion recognition and speaker state classification by learning complex patterns directly from raw audio or spectrogram inputs (Latif et al. 2023). Like their visual counterparts, these models require large labeled datasets and are often difficult to interpret. Speech signals also present unique challenges, including rapid temporal dynamics and variability across speakers and recording environments. To address these issues, recent models use architectures that explicitly capture temporal dependencies (e.g., recurrent or convolutional networks, and long short-term memory modules) or leverage pretraining on large-scale speech datasets to support transfer learning to smaller, behaviorally focused datasets (Chen & Rudnicky 2023). In parallel, deep learning has also driven major advances in automatic speech recognition, enabling high-accuracy transcription of spoken language. Tools such as Whisper (Radford et al. 2023) and Parakeet (Galvez et al. 2024) can efficiently generate transcripts from raw audio with minimal supervision, supporting downstream analysis of verbal content.

3.3.4. Natural language processing methods. Analysis of expressive behavior in the text modality focuses on patterns of language use, including word choice, grammar, meaning, and discourse structure. Over time, these methods have evolved from hand-crafted linguistic features, to semantic and discourse models, to deep learning approaches that derive high-dimensional language representations from large-scale data.

Hand-crafted linguistic features are derived from transcribed speech or written text using predefined rules grounded in linguistic and psychological theory. These features include word counts, part-of-speech tags, syntactic complexity, lexical diversity, sentiment polarity, and topic distributions. Tools such as LIWC (Boyd et al. 2022), Coh-Metrix (McNamara et al. 2014), and various affective lexica provide interpretable summary scores linked to psychological traits, emotional states, and social processes. These methods are efficient, theory-informed, and easy to interpret, but they often miss contextual nuance and depend on accurate transcription and language-specific resources (Eichstaedt et al. 2021).

Model-based representations in NLP focus on capturing meaning and structure at the levels of semantics and discourse. These include techniques such as topic modeling, semantic networks, and coherence models that track how ideas develop, shift, and relate across stretches of text (Churchill & Singh 2022, Segev 2022, Barzilay & Lapata 2008). Such representations enable the identification of narrative themes, relational framing, and discourse organization. Compared to hand-crafted features, they offer greater flexibility and contextual sensitivity, capturing meaning that emerges across sentences rather than at the word or phrase level. However, they often require careful tuning and domain adaptation, and their interpretability can vary. Outputs may be sensitive to text length, quality, and style, particularly in informal or noisy data, such as transcripts of spontaneous speech.

Deep learning approaches in NLP learn rich, contextualized representations of language. Transformer-based models such as BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and domain-specific variants (e.g., ClinicalBERT; Turchin et al. 2023) generate embeddings that capture subtle cues related to affect, tone, and intent. These representations support a range of downstream prediction and discovery tasks, including emotion classification, personality inference, identification of latent themes, and clustering of communication styles. Recently, **large language models** (LLMs) have enabled generative applications, including narrative summarization, conversational reasoning, and “few-shot” or “zero-shot” learning (i.e., making predictions with little or no labeled data; Brown et al. 2020). While these models offer impressive flexibility and performance, they are also resource-intensive and difficult to interpret. Moreover, most LLMs are developed and trained primarily in English, limiting their effectiveness in other languages without substantial adaptation. Although core NLP tasks (e.g., part-of-speech tagging) are increasingly supported in non-English languages, clinical applications require additional capabilities, such as recognizing medical terminology, abbreviations, and language-specific named entities, which remain underdeveloped in many contexts (Névéol et al. 2018). Based on a search of PubMed-indexed papers, non-English clinical NLP efforts have grown in the last decade, though important gaps persist (see Supplemental Figure S1). Since domain-specific tools (such as affect analysis) are also more advanced for English, texts in different languages are often automatically translated into English for further analysis (e.g., Halfon et al. 2021).

Large Language Model: A deep learning model trained on massive text datasets to analyze, predict, and generate language.

3.3.5. Multimodal Fusion. Multimodal fusion refers to the integration of multiple data streams to improve computational analysis. Each modality (ideally) provides unique and complementary information, and combining them can yield richer representations of psychological processes and more accurate, generalizable models. A comprehensive framework is provided by Baltrušaitis et al. (2018) and Liang et al. (2024), which outline theoretical foundations and practical implementations of multimodal fusion.

Multimodal systems differ in when and how they combine signals from each modality. In *early fusion*, features from each modality (such as pitch, word embeddings, and facial movements) are combined at the input stage, before any modeling occurs. This approach allows for learning cross-modal interactions from the start but can be limited by difficulties aligning signals in time or managing very different feature types. In contrast, *late fusion* involves analyzing each modality separately and then merging the resulting outputs, e.g., by averaging predictions or using voting rules. This strategy is modular and easier to interpret, but may overlook the way different cues influence each other. Finally, *hybrid fusion* sits between these extremes, combining partially processed representations from each modality. These models often incorporate mechanisms such as *attention* and *alignment* to capture relationships across modalities while preserving the distinct structure of each input. In this context, attention refers to a model's ability to weigh different parts of the input based on their relevance, while alignment involves mapping specific elements of the input to corresponding elements of the output (Bahdanau et al. 2015).

Recent advances in deep learning have made hybrid fusion increasingly effective, allowing systems to model the temporal, semantic, and emotional interplay among different modalities. This means that rather than treating each signal in isolation, the system learns how features interact to convey meaning over time. For example, it can capture how a sarcastic tone or raised eyebrow alters the interpretation of an otherwise neutral sentence. Modern architectures supporting this include memory-augmented recurrent networks (which track and integrate past information), co-attention transformers (which jointly focus on relevant parts of multiple modalities), and graph-based models (which represent relationships among modalities as structured, interconnected patterns) (Shen et al. 2024). These systems have been applied to a range of mental health tasks, including emotion recognition, depression detection, and therapeutic dialogue analysis (Sadeghi et al. 2024).

Despite these advances, multimodal systems still face several key challenges. Synchronizing data streams, managing computational demands, and modeling complex interactions across modalities remain difficult. Performance often degrades when input from one modality is missing or compromised (e.g., due to background noise in audio or occluded facial expressions in video). To address these issues, researchers have increasingly turned to *self-supervised learning*, a technique in which models learn useful patterns and representations from raw data by solving auxiliary tasks (e.g., predicting missing segments or aligning information across modalities) without requiring manual labels. This approach can reduce the need for large annotated datasets while improving the system's ability to generalize across real-world variability (Kaya et al. 2017).

4. Illustrative Literature Review

While a comprehensive review of all interdisciplinary work applying computational analysis of expressive behavior to clinical assessment is possible, such a review would be prohibitively long and poorly suited to the goals and format of the current article. Instead, we adopt a two-part strategy that balances depth and breadth. First, we analyzed all peer-reviewed articles published in *IEEE Transactions on Affective Computing*, a flagship journal that consistently features state-of-the-art work at the intersection of computational modeling of human behavior and clinical science. While not exhaustive, this focused sample offers a representative snapshot of empirical developments over time. Second, we broaden our scope by summarizing key conclusions from selected reviews published across diverse venues. This dual approach enables us to identify emerging patterns and gaps in the field while drawing on the insights of prior syntheses to contextualize current trends.

4.1. Journal-Specific Review

Across the 1321 published and Early Access articles in the *IEEE Transactions on Affective Computing* journal from its inception in 2010 through June 6, 2025, a total of 101 articles were coded by J.M.G. or D.A.Y. as relevant to both computational analysis of expressive behavior and clinical assessment. Thus, this topic comprises 8% of the work in this journal.

Six of the relevant papers (6%) were review articles, and the remainder (94%) reported the results of novel empirical research. Figure 1a shows that the cumulative number of papers on this topic has increased exponentially over time, $F(1, 14) = 950.8$, $p < .001$.

Our coding of the clinical topics examined in these papers yielded nine main categories, listed here in descending order of frequency: *depression* and unipolar mood conditions, *autism* and related developmental conditions, *stress* not related to a specific condition, *mania* and bipolar mood conditions, *dementia* and neurodegenerative conditions, *suicide* and self-injurious thoughts and behaviors, *anxiety* and fear-related conditions, *trauma* and posttraumatic stress, and *psychosis* including schizophrenia and formal thought disorder. As shown in Figure 1b, these topics were not equally represented; the depression category alone accounted for over half of the empirical papers (54%). Two topics appeared only once and were grouped into an *other* category: distressed couples' level of conflict during discussion tasks and clients' level of engagement during motivational interviewing.

Unimodal studies, which examine a single modality of data (i.e., audio, video, or text), were more common (73%) than multimodal studies (27%). As shown in Figure 1c, unimodal video studies were most prevalent (39%), followed by text (18%) and audio (16%). Multimodal studies most often combined audio and video (11%) or all three modalities (9%), while combinations of audio and text (4%) or video and text (3%) were less frequent.

While caution is warranted in generalizing beyond this journal, certain clinical topics showed consistent patterns in modality use. The anxiety, dementia, trauma, and psychosis topics more often incorporated the audio modality (i.e., in over 33% of papers). The anxiety, autism, depression, and psychosis topics tended to favor the video modality, whereas the mania and suicide topics most often used the text modality. Notably, no papers in this journal have used the text modality to study trauma or psychosis. However, such work appears in other venues, especially for psychosis, as highlighted in subsection 4.4.

4.2. Computer Vision Reviews

Pampouchidou et al. (2019) systematically review over 60 studies that assess depression based on facial expressions, head movement, and postural cues. They highlight the promise of automated approaches for detecting depressive symptoms, but note that such systems remain far from clinical deployment. Key limitations include narrow sample diversity, a lack of standardized protocols, and insufficient validation in naturalistic environments. The authors recommend that future work prioritize large-scale, longitudinal studies, accommodate individual variability, and focus on building systems that are interpretable, reliable, and clinically grounded.

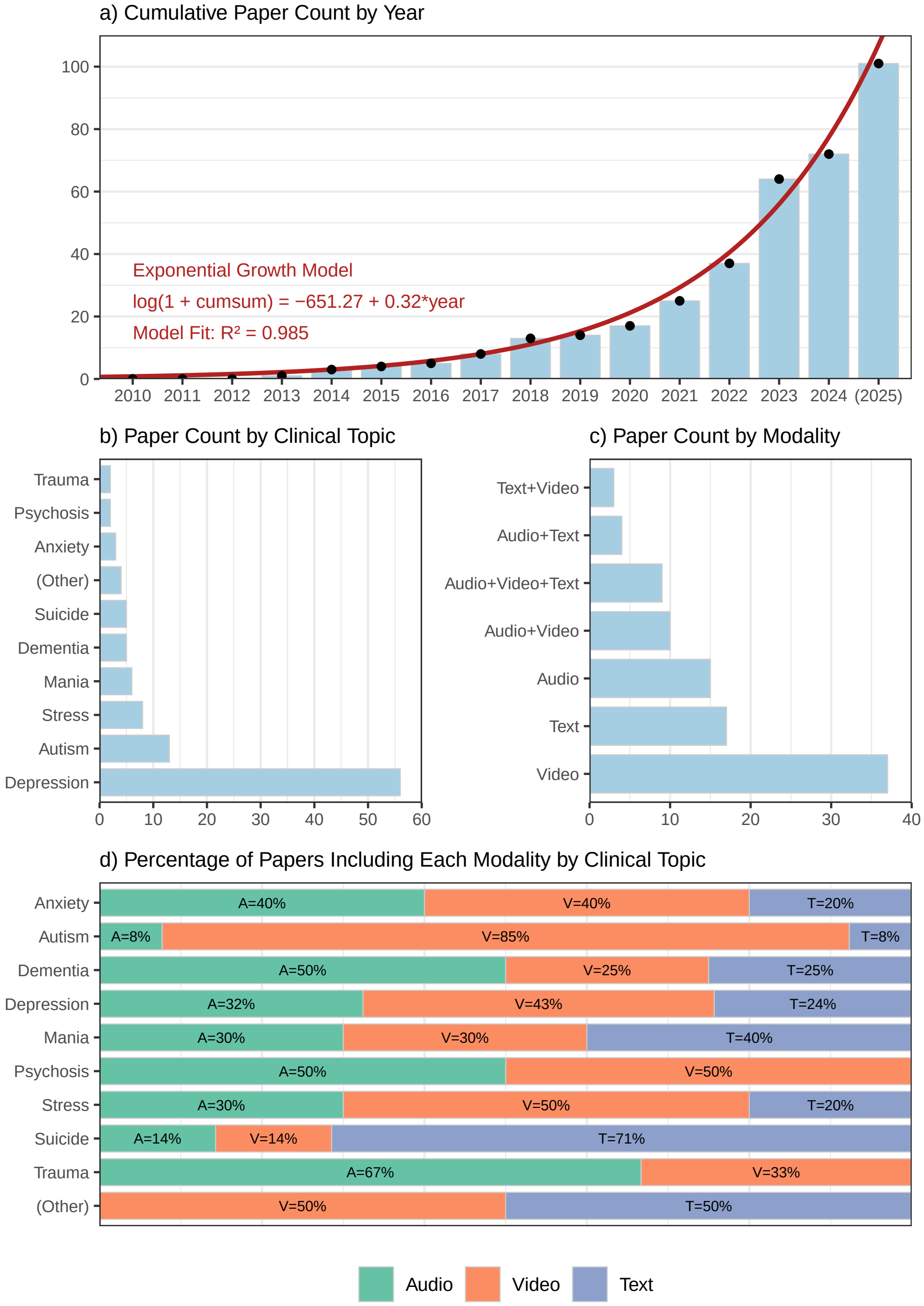

Figure 1

Review of Assessment-Related Papers in *IEEE Transactions on Affective Computing* (2010–2025) de Belen et al. (2020) systematically review 83 studies that assess autism using visual markers such as gaze patterns, facial expressions, body movements, and stereotyped behaviors, along with 11 additional studies using neuroimaging. These approaches show strong promise for identifying core features of autism, particularly in children. However, most studies to date have been conducted in artificial, highly controlled environments. The authors emphasize the need for benchmark datasets, more ecologically valid settings, and longitudinal designs to better capture developmental trajectories and individual variability.

Jiang et al. (2022) systematically review 17 studies that use facial behavior and head motion to assess schizophrenia, particularly negative symptoms such as blunted affect. They argue that these methods offer a valuable means of objectively quantifying subtle behavioral changes, but the field is still in its early stages. Small and homogeneous samples, inconsistent clinical labeling, and limited alignment with clinical outcomes limit generalizability. To address these gaps, the authors advocate for more diverse datasets, greater methodological consistency, and stronger integration with diagnostic and therapeutic frameworks.

4.3. Speech Signal Processing Reviews

Cummins et al. (2015) and Cummins et al. (2018) provide narrative reviews of studies that apply speech analysis to health-related outcomes, including autism, depression, and suicide risk. They emphasize that speech offers a non-invasive and accessible means of monitoring psychological states, and highlight how deep learning has expanded the modeling capabilities of such systems. Despite this promise, clinical application remains limited due to data sparsity, inconsistent protocols, and the opaqueness of many models. The authors underscore the need for multimodal approaches, standardized procedures, and validation in real-world settings to build more robust and clinically meaningful tools.

Low et al. (2020) systematically review 127 studies that use acoustic features of speech to assess a wide range of psychiatric conditions, including depression, schizophrenia, bipolar disorder, posttraumatic stress disorder, anxiety, and eating disorders. They find strong potential for speech-based technologies to enable remote, scalable mental health assessment, but point to persistent barriers such as limited demographic and linguistic diversity, poor model generalizability, and challenges in interpreting system outputs. To improve reliability and impact, the authors advocate for transdiagnostic and longitudinal designs, attention to algorithmic fairness, and the adoption of reproducible research practices.

Ding & Zhang (2023) provide a narrative review of studies on prosodic features of speech in mental health, focusing on conditions such as depression, schizophrenia, bipolar disorder, and autism. They argue that prosody provides a rich but underutilized channel for detecting cognitive and emotional dysfunction. However, inconsistencies in how prosodic cues are defined, extracted, and interpreted, along with small sample sizes and limited use of longitudinal designs, have slowed progress. The authors call for stronger theoretical foundations, greater methodological rigor, and closer integration with other behavioral data streams to fully realize the clinical potential of prosodic analysis.

4.4. Natural Language Processing Reviews

Zhang et al. (2022) provide a narrative review of studies that apply NLP techniques to detect a broad range of mental health conditions, including depression, suicide risk, bipolar disorder, schizophrenia, and autism. These studies draw from diverse textual sources such as social media, clinical notes, and interviews. The authors note a clear trend toward deep learning methods but emphasize that the field still faces key challenges, including limited access to high-quality datasets, especially for non-English languages, inconsistent annotation standards, and a narrow focus on classification tasks. They call for increased investment in multilingual and multi-domain resources, greater emphasis on model interpretability, and methodological innovation that moves beyond current task conventions.

Harvey et al. (2022) provide a scoping review of 35 studies using NLP to assess bipolar disorder, focusing on how language use varies across manic, depressive, and euthymic states. They find that linguistic markers such as verbosity, syntactic complexity, and semantic coherence fluctuate with mood and may support automated monitoring. However, existing work is constrained by small and non-representative samples, inconsistent labeling practices, and unclear connections between linguistic features and clinical diagnoses. The authors emphasize the need for standardized data collection protocols, improved annotation of mood states, and longitudinal designs that can capture within-person change.

Deneault et al. (2024) provide a scoping review of 18 studies applying NLP techniques in schizophrenia research, identifying six main use cases: predictive modeling, analysis of linguistic features, coherence metrics, clinical decision support, interview analysis, and social media monitoring. They highlight the potential of NLP to quantify language disorganization and formal thought disorder, particularly through measures of semantic coherence and speech graph structure. However, progress has been limited by small sample sizes, diagnostic ambiguity, and limited clinical validation. The authors recommend more clinically grounded research designs, larger and more representative datasets, and shared benchmarks to support comparison and replication.

Although Alzheimer's disease is primarily a neurocognitive rather than psychiatric condition, it illustrates the broader potential of NLP methods for detecting subtle cognitive impairments. In a systematic review of 79 studies, Shakeri & Farmanbar (2025) examine how language features can serve as early indicators of decline in individuals with the disease. Lexical, syntactic, and semantic markers, such as vocabulary richness, coherence, and sentence structure, have shown promise as indicators of disease progression. However, the field continues to face major barriers, including fragmented datasets, limited cross-study comparability, and poor integration with clinical workflows. The authors emphasize the need for shared benchmarks, closer clinical collaboration, and models that are both robust and generalizable across populations and settings.

4.5. Multimodal Reviews

Recent reviews by Liang et al. (2019) and Khoo et al. (2024) synthesize research on multimodal digital phenotyping, which combines behavioral signals (e.g., facial expressions, vocal prosody, and language use) with passive sensing data from personal devices (e.g., GPS for location tracking and accelerometry for activity monitoring). Both reviews emphasize the promise of integrating video, audio, and text modalities with ambient data streams to enable remote, continuous, and context-aware assessment of mental health conditions such as depression, anxiety, and schizophrenia. Liang et al. (2019) offer a broad conceptual overview, arguing that fusing diverse data sources can yield richer behavioral representations and more scalable insights. However, they caution that methodological fragmentation and ethical concerns—particularly around privacy and consent—pose significant barriers to real-world deployment. Khoo et al. (2024), in a systematic review of 184 empirical studies, echo these themes but focus more on predictive performance, concluding that multimodal systems consistently outperform unimodal ones in both accuracy and ecological validity. Yet, they too highlight major limitations, including inconsistent data collection protocols, lack of standardized feature representations, and minimal integration with clinical workflows. Both reviews call for benchmark datasets, ethical safeguards, interdisciplinary collaboration, and alignment with clinical priorities to advance the field.

Al Sahili et al. (2025) offer a detailed technical survey of multimodal machine learning methods in mental health. They discuss a range of modalities (including text, audio, video, physiology, and neuroimaging) while particularly highlighting scalable, noninvasive signals such as text, audio, and video. Their analysis explores deep learning trends, especially multimodal transformers and attention-based fusion, showing how these architectures can compensate for weaknesses in individual modalities and capture complex intermodal dynamics. They also identify ongoing challenges: data heterogeneity, lack of cross-study comparability, and scarcity of large, annotated datasets, and they advocate for theoretical integration and ethical safeguards on privacy, consent, and algorithmic bias.

5. Discussion

The use of computational methods to analyze expressive behavior in clinical settings holds enormous promise, but also presents critical challenges. As this field matures, researchers must navigate conceptual, methodological, ethical, and practical tensions that span multiple disciplines. This section synthesizes key limitations of current approaches, including issues of data quality, labeling, bias, and integration with clinical needs. It also outlines opportunities for more inclusive, theory-driven, and collaborative frameworks that can improve both the scientific and clinical value of these tools.

5.1. Data and Sampling Issues

Many studies in this field rely on categorical diagnostic frameworks (e.g., DSM or ICD) to define inclusion criteria and compare individuals with a single diagnosis to healthy or typically developing controls. These sampling strategies aim to isolate disorder-specific features, but without clinical control groups or dimensional characterization, observed differences may instead reflect general psychological distress. The result is often limited insight into the specificity or generality of predictive features.

Samples are often drawn from university or urban clinical settings in high-income countries, limiting the cultural, linguistic, and socioeconomic diversity of participants. Yet expressive behavior varies across communities, shaped by differences in communication styles, emotional norms, and social expectations (Matsumoto et al. 2008). As a result, models developed in one context may not generalize well to others. This issue is particularly acute for language-based models, which are typically trained on English-language data and may not transfer effectively to other languages or dialects. These challenges are compounded by small and demographically narrow samples, which increase the risk of overfitting and limit generalizability across age groups, genders, and clinical subtypes.

Addressing these limitations will require greater coordination and shared infrastructure. The Audio/Visual Emotion Challenge (AVEC) has demonstrated the value of shared tasks in mental health prediction, helping to standardize preprocessing, evaluation, and benchmarking procedures (e.g., Valstar et al. 2013, Ringeval et al. 2019). Building on this foundation, future efforts should involve clinical scientists more centrally in task design, outcome selection, and interpretation. Shared datasets that span—and enable performance comparisons across—diagnostic categories, demographic groups, and recording conditions are essential for developing more robust, equitable, and clinically meaningful models.

5.2. Measurement Challenges

Even when samples are well chosen and adequately sized, the reliability and validity of outcome labels pose major obstacles to model development and evaluation (Flake & Fried 2020). Clinical labels are often noisy due to variability in assessment practices, diagnostic thresholds, and limited inter-rater agreement; even structured interviews can yield inconsistent results. Many studies instead rely on screening tools, using cutoffs to assign binary labels. While efficient, this approach introduces misclassification error, as screeners are not designed to distinguish true cases with clinical precision. False positives and negatives distort model training and evaluation, potentially inflating or masking observed effects. More reliable strategies, such as consensus ratings, multi-source validation, or continuous symptom measures, are needed to improve model validity.

Categorical systems like DSM and ICD provide operational clarity but may obscure individual variation and conflate shared features across conditions. Dimensional frameworks such as RDoC, HiTOP, and network models offer more nuanced or mechanistic views, better capturing symptom overlap and heterogeneity common in clinical populations. However, these alternatives bring their own challenges, including inconsistencies in how constructs are defined and measured, which can complicate model development and comparison. Without a clear gold standard, it is difficult to evaluate validity or determine which targets are most appropriate. Clarifying and justifying nosological choices will be essential as the field moves toward more flexible representations of mental health.

In addition to challenges with outcome labels, the behavioral features extracted from raw data also present significant measurement difficulties. Real-world data often include artifacts that challenge computational analysis. Visual occlusions, extreme head poses, poor lighting, and low-resolution video can interfere with the extraction of facial or bodily features. Likewise, background noise, low-quality audio, and overlapping speech reduce the reliability of acoustic features and transcription. These issues are especially common in naturalistic or remote settings and can substantially degrade model performance. Their impact may also be uneven, e.g., young children, older adults, or individuals with disabilities may be more likely to produce data with such artifacts. Addressing these challenges will require robust preprocessing, artifact-aware modeling strategies, and evaluation procedures that account for data quality.

5.3. Interdisciplinary Challenges

Collaboration between clinical and computational communities is often hindered by differences in priorities, terminology, and methodological assumptions. Clinicians tend to emphasize interpretability, theoretical grounding, and clinical utility, while computational work often prioritizes predictive accuracy, scalability, and technical novelty. These differing goals can result in models optimized for metrics that lack clinical meaning. For example, a machine learning model may operationalize “engagement” in terms of observable behaviors like eye contact or speech duration, whereas clinicians typically require a more conceptually grounded definition, such as emotional involvement, therapeutic rapport, or responsiveness to treatment. Structural barriers further limit cross-pollination: researchers are typically trained in discipline-specific silos, publish in separate venues, and work on different timelines (e.g., computational fields often emphasize frequent conference papers, while clinical fields publish more slowly in journals). Even peer review itself poses challenges: it is often difficult to identify reviewers with expertise spanning both domains, and there are few incentives for providing interdisciplinary review service, whether for journal articles or grant proposals. Advancing the field will require greater cross-disciplinary training, shared language, and collaborative platforms that support both methodological rigor and clinical relevance.

5.4. Ethical Considerations

Behavioral data used in computational research, such as video, audio, and text, are highly sensitive and often personally identifiable. Unlike standard clinical measures, these data can reveal subtle patterns in how people express themselves, interact with others, and move through their environments. As a result, informed consent must go beyond the moment of data collection to address long-term issues, such as how data will be stored, reused, or shared. Even when identifying details are removed, linking datasets or using certain features can make it possible to re-identify individuals. Safeguards like limiting how much data is collected, restricting who can access it, and deleting it after a set time can reduce these risks. Technical strategies such as *federated learning* (training models across sites without moving the data) and *differential privacy* (adding noise to protect identity) offer promising solutions, though they may come with trade-offs in accuracy or transparency (Kairouz et al. 2021). Following ethical frameworks like the Unified Five Principles (Floridi & Cowls 2022), researchers must ensure that participants understand how their data will be used and what risks are involved (see Table 2 as well as the supplemental materials).

Models trained on limited or non-representative samples can reproduce or worsen existing social inequalities. If performance differences across demographic groups are not carefully examined, these systems may produce inaccurate or unfair results, particularly for groups already underserved by the healthcare system. In mental health settings, for instance, cultural norms around emotional expression or symptom reporting may affect how well a model works across different populations (Timmons et al. 2023). Bias can be introduced during data collection, in how data are labeled, or through the model itself. It may also vary based on combinations of characteristics like race, gender, age, or disability. Developers should routinely evaluate whether models perform equitably across groups and consider the specific harms of false positives and false negatives in different clinical settings (Sogancioglu et al. 2024). Ensuring fairness requires input from clinicians, community members, and people with lived experience.

As interest grows in using these tools in clinical practice, there is still little clear guidance on how to regulate, evaluate, or oversee them. Many systems are built and tested in research labs, without the clinical trials or ongoing monitoring typically required for medical tools. Yet once deployed, these models may influence decisions about diagnosis, treatment, or access to care. When errors occur, it is often unclear who is responsible: the clinician, the developer, or the organization deploying the system. These challenges are worsened by the opaque nature of many machine learning models, especially complex systems that offer little insight into how decisions are made. Building trust in clinical settings will require clearer rules about responsibility and systems that can explain their decisions in ways that make sense to patients, clinicians, and developers alike (Floridi & Cowls 2022).

Table 2 Ethical principles, major potential risks, and mitigation strategies

Ethical principle	Potential risks	Mitigation strategies
Respect for autonomy	Patient data used without the patient's informed consent	Raise awareness in the patients, obtain explicit consent
Non-maleficence	Autonomous systems of lower quality replace existing high-quality support.	Do not permit profits to dictate public health support.
	Data collected for training machine learning systems are repurposed.	To prevent function creep, provide legal safeguards, enforce data minimization, limit data storage time, ensure de-identification, permit users to access, delete, and correct their data records.
	Incorrect predictions about an individual's prognosis may affect their future care.	Increase the accuracy of computational assessment and conduct randomized controlled trials and real world effectiveness studies.
Beneficence	Automatic systems are used for marketing, political influence or for supporting individuals forced to perform tasks that run counter to human rights.	Implement policies that limit power of individual institutions and companies, support and empower non-governmental organizations for oversight.
Justice	Biases in the algorithms disadvantage minority groups. Intersectional biases are overlooked.	Explicitly test for performance on vulnerable and other subgroups. Make intersectional cases visible, test and report on issues.
Explicability	AI approaches are black-box or explanations are post hoc or fail to follow the logic of decision within the algorithm.	Prefer white-box methods, provide explainability at different levels (developer, caregiver, patient).

5.5. Future Directions

Addressing the conceptual and practical challenges outlined above will require coordinated efforts to build more inclusive, rigorous, and clinically grounded approaches. A central priority is the development of large, diverse, and multimodal datasets built on standardized protocols and shared benchmarks. Such resources can support more generalizable findings, enable stratified evaluation across subgroups, and reduce fragmentation in methods and outcomes. Just as important as the data themselves are the labels they carry. Improved labeling practices—such as consensus ratings, multi-source validation, and continuous symptom tracking—can help mitigate noise and better reflect the dimensional nature of mental health.

Future models must also be better aligned with both clinical theory and real-world practice. Many current systems prioritize predictive performance but lack grounding in established psychological constructs or relevance to the needs of end users. Integrating conceptual frameworks from psychology and psychiatry can improve interpretability and validity, while attention to clinical workflows, constraints, and decision-making contexts can enhance usability and trust. Technically sound models that are conceptually opaque or logistically impractical are unlikely to be adopted or sustained in clinical settings.

Rather than replacing clinicians, the most promising computational tools will augment clinical decision-making. When thoughtfully designed, these systems can improve consistency, reduce cognitive burden, and surface patterns that might elude human judgment. But achieving this vision requires more than accuracy; it demands models that are transparent, actionable, and tailored to specific use cases. Clinical utility should not be an afterthought but a central design goal.

Finally, progress in this field depends on sustained interdisciplinary collaboration. Training programs, research networks, and funding initiatives should promote integration across psychology, psychiatry, computer science, and ethics. Common standards for model development, validation, and reporting can improve transparency and comparability. And by involving clinicians, patients, and communities throughout the research process, computational assessment can become not only more sophisticated but also more meaningful, responsible, and equitable.

5.6. Conclusions

Computational analysis of expressive behavior offers a powerful set of tools for advancing mental health assessment, but its successful application depends on thoughtful alignment with clinical science. This paper aimed to provide an accessible conceptual framework for understanding these methods, bridging clinical and computational perspectives to clarify key ideas and highlight their points of intersection. We began by reviewing foundational goals and approaches in clinical assessment, then described how computational techniques from computer vision, speech signal processing, and natural language processing are being applied to study expressive behavior. Drawing on recent research, we illustrated how these tools are being used to address core assessment goals. Finally, we synthesized conceptual, methodological, and ethical challenges that must be addressed to ensure these systems are valid, equitable, and clinically meaningful. By prioritizing theory, context, and collaboration alongside technical performance, future work can help realize the potential of computational methods to support more precise, scalable, and person-centered mental health care.

SUMMARY POINTS

1. Clinical assessment relies on questionnaires, interviews, and observation, but each has well-documented limitations in reliability, validity, and scalability.
2. Computational methods (e.g., computer vision, speech signal processing, natural language processing, and machine learning) enable more objective, fine-grained, and scalable measurement of expressive behavior for clinical assessment.
3. Advances in multimodal approaches, deep learning, and generative AI allow richer integration of visual, vocal, and linguistic signals, enabling the capture of complex behaviors and the creation of realistic simulations for training and assessment.
4. A review of the literature shows rapid growth in the field, but most studies remain unimodal, concentrate on a narrow set of conditions such as depression, and have yet to fully integrate computational methods into clinical practice.
5. Key challenges include data diversity, measurement quality, interdisciplinary alignment, ethical safeguards, and the need for models that are both technically sound and clinically meaningful.

FUTURE ISSUES

1. Develop large, diverse, multimodal benchmark datasets with standardized protocols, enabling generalization across conditions, demographic groups, and contexts.
2. Improve labeling practices by adopting consensus ratings, multi-source validation, and dimensional symptom tracking to better capture clinical complexity.
3. Integrate clinical theory and workflow considerations into model design, ensuring tools are interpretable, actionable, and support rather than replace clinicians.
4. Promote interdisciplinary collaboration of the clinical and computational sciences through joint training, shared infrastructure, and common reporting standards.

DISCLOSURE STATEMENT

J.F.C. serves as Chief Scientist at Deliberate AI, a company whose interests are related to the subject matter of this paper. This affiliation is disclosed in the interest of transparency. The other authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported in part by the National Institute of Mental Health, under awards R01 MH125740 (to J.M.G.) and U01 MH136535 (to J.F.C.), and by the Netherlands Organisation for Scientific Research, under grant KICH1.MV03.23.009 (to A.A.S.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

LITERATURE CITED

Aftab A, Sharma M. 2021. How not to think about biomarkers in psychiatry: Challenges and conceptual pitfalls. *Biomarkers in Neuropsychiatry* 4:100031

Al Sahili Z, Patras I, Purver M. 2025. Multimodal machine learning in mental health: a survey of data, algorithms, and challenges

Althubaiti A. 2016. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare* 9:211–217

American Psychiatric Association. 2022. Diagnostic and statistical manual of mental disorders (DSM-5-TR). APA, 5th ed.

Bahdanau D, Cho KH, Bengio Y. 2015. Neural machine translation by jointly learning to align and translate, In *3rd International Conference on Learning Representations*

Baltrušaitis T, Ahuja C, Morency LP. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41(2):423–443

Barrett LF. 2022. Context reconsidered: Complex signal ensembles, relational meaning, and population thinking in psychological science. *American Psychologist* 77(8):894–920

Barzilay R, Lapata M. 2008. Modeling local coherence: an entity-based approach. *Computational Linguistics* 34(1):1–34

Blanz V, Vetter T. 1999. A morphable model for the synthesis of 3D faces, In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '99, pp. 187–194, USA: ACM Press/Addison-Wesley Publishing Co.

Borsboom D. 2008. Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology* 64(9):1089–1108

Borsboom D. 2017. A network theory of mental disorders. *World Psychiatry* 16(1):5–13

Boswell JF, Kraus DR, Miller SD, Lambert MJ. 2015. Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research* 25(1):6–19

Box GEP. 1976. Science and Statistics. *Journal of the American Statistical Association* 71(356):791–799

Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW. 2022. The development and psychometric properties of LIWC-22. Austin, TX: University of Texas at Austin

Breiman L. 2001. Statistical modeling: The two cultures. *Statistical Science* 16(3):199–215

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, et al. 2020. Language Models are Few-Shot Learners, In *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, Curran Associates, Inc.

Bunney Jr. WE, Hamburg DA. 1963. Methods for reliable longitudinal observation of behavior: development of a method for systematic observation of emotional behavior on psychiatric wards. *Archives of General Psychiatry* 9(3):280–294

Bzdok D, Meyer-Lindenberg A. 2018. Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3(3):223–230

Chen LW, Rudnicky A. 2023. Exploring Wav2vec 2.0 Fine Tuning for Improved Speech Emotion Recognition, In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. ISSN: 2379-190X

Cheng Y, Wang H, Bao Y, Lu F. 2024. Appearance-Based Gaze Estimation With Deep Learning: A Review and Benchmark. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 46(12):7509–7528

Churchill R, Singh L. 2022. The evolution of topic modeling. *ACM Computing Surveys* 54(10s):215:1–215:35

Cohn JF, Ekman P. 2005. Measuring facial action. In *The new handbook of nonverbal behavior research*, eds. JA Harrigan, R Rosenthal, KR Scherer. New York, NY: Oxford University Press, 9–64

Cootes TF, Edwards GJ, Taylor CJ. 2001. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(6):681–685

Cramer AOJ, Waldorp LJ, van der Maas HLJ, Borsboom D. 2010. Comorbidity: A network perspective. *Behavioral and Brain Sciences* 33(2-3):137–150

Croft P, Dinant GJ, Coventry P, Barraclough K. 2015. Looking to the future: should 'prognosis' be heard as often as 'diagnosis' in medical education? *Education for Primary Care* 26(6):367–371

Cummins N, Baird A, Schuller BW. 2018. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods* 151:41–54

Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication* 71:10–49

Cuthbert BN. 2022. Research Domain Criteria (RDoC): Progress and Potential. *Current Directions in Psychological Science* 31(2):107–114

Cuthbert BN, Insel TR. 2013. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine* 11(1):126

de Belen RAJ, Bednarz T, Sowmya A, Del Favero D. 2020. Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Translational Psychiatry* 10(1):333

Deneault A, Dumais A, Désilets M, Hudon A. 2024. Natural language processing and schizophrenia: a scoping review of uses and challenges. *Journal of Personalized Medicine* 14(7):744

Devlin J, Chang MW, Lee K, Toutanova K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs]

Ding H, Zhang Y. 2023. Speech prosody in mental disorders. *Annual Review of Linguistics* 9(1):335–355

Eichstaedt JC, Kern ML, Yaden DB, Schwartz HA, Giorgi S, et al. 2021. Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods* 26(4):398–427

Ekman P, Friesen WV, Hager J. 2002. Facial action coding system: A technique for the measurement of facial movement. Salt Lake City, UT: Research Nexus, 2nd ed.

Eyben F. 2016. Real-time speech and music classification by large audio feature space extraction. Springer Theses. Springer

Eyben F, Scherer KR, Schuller BW, Sundberg J, Andre E, et al. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7(2):190–202

Fant G. 1971. Acoustic theory of speech production. Mouton

Fernandes BS, Williams LM, Steiner J, Leboyer M, Carvalho AF, Berk M. 2017. The new field of 'precision psychiatry'. *BMC Medicine* 15(1):80

Finn SE. 2007. In our client's shoes: Theory and techniques of Therapeutic Assessment. Lawrence Erlbaum Associates

Finn SE, Tonsager ME. 1997. Information-gathering and therapeutic models of assessment: Complementary paradigms. *Psychological Assessment* 9(4):374–385

First MB. 2015. Psychiatric Classification. In *Psychiatry*, eds. A Tasman, J Kay, JA Lieberman, MB First, M Riba, vol. 4th. John Wiley & Sons, Inc., 657–671

Flake JK, Fried EI. 2020. Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science* 3(4):456–465

Floridi L, Cowls J. 2022. A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design* :535–545

Fusar-Poli P, Hijazi Z, Stahl D, Steyerberg EW. 2018. The Science of Prognosis in Psychiatry: A Review. *JAMA Psychiatry* 75(12):1289–1297

Galvez D, Bataev V, Xu H, Kaldewey T. 2024. Speed of light exact greedy decoding for RNN-T speech recognition models on GPU. ArXiv:2406.03791 [cs]

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2014. Bayesian data analysis. Boca Raton, FL: CRC Press, 3rd ed.

Girard JM, Cohn JF. 2016. A primer on observational measurement. *Assessment* 23(4):404–413

Goodfellow I, Bengio Y, Courville A. 2016. Deep learning. MIT Press

Haines N, Kvam PD, Irving L, Smith CT, Beauchaine TP, et al. 2025. A tutorial on using generative models to advance psychological science: Lessons from the reliability paradox. *Psychological Methods*

Halfon S, Doyran M, Türkmen B, Oktay EA, Salah AA. 2021. Multimodal affect analysis of psychodynamic play therapy. *Psychotherapy Research* 31(3):313–328

Harvey D, Lobban F, Rayson P, Warner A, Jones S. 2022. Natural language processing methods and bipolar disorder: scoping review. *JMIR Mental Health* 9(4):e35928

Hastie T, Tibshirani R, Friedman J. 2009. The elements of statistical learning: Data mining, inference, and prediction. Springer, 2nd ed.

Haynes SN, O'Brien WH. 2000. Principles and practice of behavioral assessment. Applied clinical psychology. New York: Kluwer Academic/Plenum

Henninger M, Debelak R, Rothacher Y, Strobl C. 2025. Interpretable machine learning for psychological research: Opportunities and pitfalls. *Psychological Methods* 30(2):271–305

Hilsenroth MJ, Peters EJ, Ackerman SJ. 2004. The Development of Therapeutic Alliance During Psychological Assessment: Patient and Therapist Perspectives Across Treatment. *Journal of Personality Assessment* 83(3):332–344

Horn BKP, Schunck BG. 1981. Determining optical flow. *Artificial Intelligence* 17(1):185–203

Hu J, Mathur L, Liang PP, Morency LP. 2025. Openface 3.0: A lightweight multitask system for comprehensive facial behavior analysis. *arXiv preprint arXiv:2506.02891*

Insel T, Cuthbert B, Garvey M, Heinssen R, Pine D, et al. 2010. Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry* 167(7):748–751

Insel TR. 2014. The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry. *American Journal of Psychiatry* 171(4):395–397

Jiang Z, Luskus M, Seyedi S, Griner EL, Rad AB, et al. 2022. Utilizing computer vision for facial behavior analysis in schizophrenia studies: A systematic review. *PLOS ONE* 17(4):e0266828

Jobes DA. 2023. Managing Suicidal Risk: A Collaborative Approach. Guilford Publications

Joseph RC, Hermann RC. 1998. Screening for Psychiatric Disorders in Primary Care Settings. *Harvard Review of Psychiatry* 6(3):165–170

Jurafsky D, Martin JH. 2025. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models. Prentice Hall, 3rd ed.

Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, et al. 2021. Advances and open problems in federated learning. *Foundations and trends in machine learning* 14(1–2):1–210

Kaya H, Gürpınar F, Salah AA. 2017. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing* 65:66–75

Kazemi V, Sullivan J. 2014. One millisecond face alignment with an ensemble of regression trees, In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874. ISSN: 1063-6919

Khoo LS, Lim MK, Chong CY, McNaney R. 2024. Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. *Sensors* 24(2):348

Kim J, Kim Dg, Kamphaus R. 2022. Early Detection of Mental Health Through Universal Screening at Schools. *Georgia Educational Researcher* 19(1):62

Kipps CM, Hodges JR. 2005. Cognitive assessment for clinicians. *Journal of Neurology, Neurosurgery and Psychiatry* 76(suppl 1):i22

Kirmayer LJ. 2005. Culture, context and experience in psychiatric diagnosis. *Psychopathology* 38(4):192–196

Kotov R, Krueger RF, Watson D, Achenbach TM, Althoff RR, et al. 2017. The hierarchical taxonomy of psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology* 126(4):454–477

Kraemer HC, Measelle JR, Ablow JC, Essex MJ, Boyce WT, Kupfer DJ. 2003. A New Approach to Integrating Data From Multiple Informants in Psychiatric Assessment and Research: Mixing and Matching Contexts and Perspectives. *American Journal of Psychiatry* 160(9):1566–1577

Krueger RF, Kotov R, Watson D, Forbes MK, Eaton NR, et al. 2018. Progress in achieving quantitative classification of psychopathology. *World Psychiatry*

Ladefoged P, Johnson K. 2014. A course in phonetics. Cengage Learning, 7th ed.

Latif S, Rana R, Khalifa S, Jurdak R, Qadir J, Schuller B. 2023. Survey of Deep Representation Learning for Speech Emotion Recognition. *IEEE Transactions on Affective Computing* 14(2):1634–1654

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–444

Lewis CC, Boyd M, Puspitasari A, Navarro E, Howard J, et al. 2019. Implementing Measurement-Based Care in Behavioral Health: A Review. *JAMA Psychiatry* 76(3):324–335

Li S, Deng W. 2022. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing* 13(3):1195–1215

Liang PP, Zadeh A, Morency LP. 2024. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *ACM Computing Surveys* 56:1–42

Liang Y, Zheng X, Zeng DD. 2019. A survey on big data-driven digital phenotyping of mental health. *Information Fusion* 52:290–307

Liu Y, Ott M, Goyal N, Du J, Joshi M, et al. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692 [cs]*

Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. 2015. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34(6):248:1–248:16

Low DM, Bentley KH, Ghosh SS. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology* 5(1):96–116

Ma F, Xie Y, Li Y, He Y, Zhang Y, et al. 2025. A Review of Human Emotion Synthesis Based on Generative Technology. *IEEE Transactions on Affective Computing* :1–20

Malasinghe LP, Ramzan N, Dahal K. 2019. Remote patient monitoring: a comprehensive study. *Journal of Ambient Intelligence and Humanized Computing* 10(1):57–76

Matsumoto D, Yoo SH, Fontaine J, Anguas-Wong AM, Arriola M, et al. 2008. Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism. *Journal of Cross-Cultural Psychology* 39(1):55–74

McNamara DS, Graesser AC, McCarthy P, Cai Z. 2014. Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press

Meyer GJ, Finn SE, Eyde LD, Kay GG, Moreland KL, et al. 2001. Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist* 56(2):128–165

Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. 2018. Clinical natural language processing in languages other than English: opportunities and challenges. *Journal of biomedical semantics* 9:1–13

Norris DR, Clark MS, Shipley S. 2016. The mental status examination. *American Family Physician* 94(8):635–641

Palmer SE. 1999. Vision science: Photons to phenomenology. MIT press

Pampouchidou A, Simos PG, Marias K, Meriaudeau F, Yang F, et al. 2019. Automatic assessment of depression based on visual cues: a systematic review. *IEEE Transactions on Affective Computing* 10(4):445–470

Pan SJ, Yang Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1345–1359

Pearl J, Glymour M, Jewell NP. 2016. Causal inference in statistics: A primer. Wiley

Radford A, Kim JW, Brockman G, McLeavey C, Sutskever I. 2023. Robust speech recognition via large-scale weak supervision, In *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, pp. 28492–28518

Rakel RE. 2016. Establishing rapport. In *Textbook of family medicine*, eds. RE Rakel, DP Rakel.

Elsevier, 9th ed., 141–156.e5

Rief W, Hofmann SG, Berg M, Forbes MK, Pizzagalli DA, et al. 2023. Do We Need a Novel Framework for Classifying Psychopathology? A Discussion Paper. *Clinical Psychology in Europe* 5(4):e11699

Ringeval F, Schuller B, Valstar M, Cummins N, Cowie R, et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition, In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, pp. 3–12, New York, NY, USA: Association for Computing Machinery

Ryder AG, Ban LM, Chentsova-Dutton YE. 2011. Towards a Cultural–Clinical Psychology. *Social and Personality Psychology Compass* 5(12):960–975

Sadeghi M, Richer R, Egger B, Schindler-Gmelch L, Rupp LH, et al. 2024. Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research* 3(66)

Scherer KR. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40(1):227–256

Segev E, ed. 2022. Semantic network analysis in social sciences. Routledge

Shakeri A, Farmanbar M. 2025. Natural language processing in Alzheimer's disease research: Systematic review of methods, data, and efficacy. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 17(1)

Shen H, Song S, Gunes H. 2024. Multi-modal human behaviour graph representation learning for automatic depression assessment, In *IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*

Shiffman S, Stone AA, Hufford MR. 2008. Ecological momentary assessment. *Annual Review of Clinical Psychology* 4:1–32

Shortliffe EH, Sepúlveda MJ. 2018. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* 320(21):2199–2200

Silberman EK, Certa K, Kay A. 2015. The psychiatric interview: Settings and techniques. In *Psychiatry*, eds. A Tasman, J Kay, JA Lieberman, MB First, M Riba. John Wiley & Sons, Inc., 4th ed., 34–55

Simms LJ, Wright AGC, Cicero D, Kotov R, Mullins-Sweatt SN, et al. 2022. Development of Measures for the Hierarchical Taxonomy of Psychopathology (HiTOP): A Collaborative Scale Development Project. *Assessment* 29(1):3–16

Singh JP, Grann M, Fazel S. 2011. A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review* 31(3):499–513

Sogancioglu G, Mosteiro P, Salah AA, Scheepers F, Kaya H. 2024. Fairness in ai-based mental health: Clinician perspectives and bias mitigation, In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, pp. 1390–1400

Spitzer RL. 1983. Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry* 24(5):399–411

Stone AA, Turkkan JS, Bachrach CA, Jobe JB, Kurtzman HS, Cain VS. 2000. The science of self-report: Implications for research and practice. Mahwah, NJ: Lawrence Erlbaum Associates

Szeliski R. 2022. Computer vision: Algorithms and applications. Texts in computer science. Springer, 2nd ed.

Søvold LE, Naslund JA, Kousoulis AA, Saxena S, Qoronfleh MW, et al. 2021. Prioritizing the Mental Health and Well-Being of Healthcare Workers: An Urgent Global Public Health Priority. *Frontiers in Public Health* 9

Tan L, Jiang J. 2018. Digital signal processing: Fundamentals and applications. Academic Press

Tedeschi RG, Kilmer RP. 2005. Assessing Strengths, Resilience, and Growth to Guide Clinical Interventions. *Professional Psychology: Research and Practice* 36(3):230–237

Tian Yl, Kanade T, Cohn JF. 2002. Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. *IEEE International Conference on Automatic Face and Gesture Recognition* :229–234

Timmons AC, Duong JB, Simo Fiallo N, Lee T, Vo HPQ, et al. 2023. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science* 18(5):1062–1096

Turchin A, Masharsky S, Zitnik M. 2023. Comparison of BERT implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked* 36:101139

Valstar MF, Schuller BW, Smith K, Eyben F, Jiang B, et al. 2013. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge, In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pp. 3–10

Venkatasubramanian G, Keshavan MS. 2016. Biomarkers in psychiatry – A critique. *Annals of Neurosciences* 23(1):3–5

Weiss K, Khoshgoftaar TM, Wang D. 2016. A survey of transfer learning. *Journal of Big Data* 3(1):9

World Health Organization. 2019. International statistical classification of diseases and related health problems. World Health Organization, 11th ed.

Xu Y. 2013. ProsodyPro - A tool for large-scale systematic prosody analysis, In *Proceedings of the Tools and Resources for the Analysis of Speech Prosody Workshop*, pp. 7–10, Aix-en-Provence, France

Yarkoni T, Westfall J. 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science* 12(6):1100–1122

Zhang T, Schoene AM, Ji S, Ananiadou S. 2022. Natural language processing applied to mental illness detection: a narrative review. *npj Digital Medicine* 5(1):46

Zhao G, Pietikainen M. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(6):915–928

Zheng C, Wu W, Chen C, Yang T, Zhu S, et al. 2023. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys* 56(1):1–37