Digital Phenotyping of Mental Health Disorders using Wearable Smartphone Technologies: A Systematic Review

Victor E. Ekong
Department of Software Engineering,
University of Uyo,
Uyo, Akwa Ibom State, Nigeria

Peter Godfrey Obike
Department of Computer Science,
Michael Okpara University of Agriculture, Umudike, Umuahia,
Abia State, Nigeria

Hydara Mbemba
Department of Computer Science,
School of Information Technology and Communications, University of The Gambia, The Gambia

Uyinomen Ekong
Department of Cybersecurity,
University of Uyo, Uyo,
Akwa Ibom State, Nigeria

ABSTRACT
Digital phenotyping has emerged as a promising paradigm for continuous and objective mental health monitoring using wearable and smartphone technologies. However, existing studies remain fragmented in terms of methodological consistency, model validation, and clinical applicability. This paper presents a systematic and performance-oriented review of digital phenotyping systems, synthesizing findings from 62 studies, including 16 high-relevance articles. Unlike prior reviews, this study introduces a structured computational framework that characterizes the end-to-end pipeline of digital phenotyping systems, encompassing data acquisition, feature engineering, machine learning modeling, and clinical decision support. Comparative analysis reveals that predictive models achieve accuracies ranging from 72% to 82%, with probabilistic and supervised learning approaches outperforming traditional regression techniques. However, significant gaps persist in external validation, reproducibility, and multimodal data integration. The findings highlight the need for standardized benchmarking protocols, improved algorithm transparency, and adaptive AI-driven intervention mechanisms. By bridging methodological, computational, and ethical dimensions, this study provides a foundation for the design and evaluation of next-generation digital mental health systems. This study contributes to the advancement of computational methods for scalable, data-driven mental health monitoring systems.

General Terms
Digital Phenotyping, Wearable Technologies, Mental Health Monitoring, Digital Biomarkers, Computational Framework, Machine Learning.

Keywords
Digital Phenotyping, Wearable Technologies, Mental Health Monitoring, Digital Biomarkers, Computational Framework, Machine Learning.

1. INTRODUCTION
Digital phenotyping refers to the collection of biometric and behavioral data from digital devices such as smartphones, wearables, and social media platforms to measure mental health indicators [1, 2]. Unlike traditional methods of assessment, digital phenotyping allows for real-time and continuous data collection, providing deeper insights into mental health disorders such as depression, anxiety, schizophrenia, and PTSD. Wearable devices, including smartwatches, fitness trackers, and smartphone applications, have significantly advanced mental health research and treatment by enabling ongoing monitoring of mood, cognition, and behavior [3, 4, 5]. These technologies offer the potential for early detection, monitoring, and personalized treatment of mental health disorders, which is especially valuable for disorders that are challenging to diagnose and manage using traditional methods [6, 7].

Recent advances in digital phenotyping have demonstrated predictive accuracies ranging from 65% to 82% for mood and anxiety disorder detection using passive smartphone sensing and wearable-derived features [8, 9, 10, 11]. However, most existing reviews have primarily emphasized feasibility and user acceptability rather than systematic evaluation of algorithmic performance, external validation, and clinical transferability. This demonstrates the need for more performance-oriented synthesis of current research.

The role of wearable technologies in mental health research has expanded significantly in recent years. For instance, research has shown that smartphone-based tools can identify digital biomarkers for disorders like social anxiety and bipolar disorder, demonstrating the accuracy and potential of these tools in supporting diagnostic processes [5, 6]. These technologies offer objective, data-driven insights, which is a notable advantage over traditional self-reported measures that often suffer from biases [8]. However, despite these advancements, the full integration of wearable devices into clinical practice remains hindered by several challenges, including data standardization, algorithmic accuracy, and user engagement [9, 10].

1.1 Problem
Despite the significant promise of digital phenotyping, the field remains methodologically fragmented, with inconsistent data standards, limited algorithm validation, and poor

reproducibility across studies, which critically hinders clinical adoption.

Despite the promising potential of digital phenotyping, several critical gaps remain in the field. One key issue is the lack of standardization in the methodologies used to collect and interpret data across studies. This makes it challenging to compare findings and establish consistent best practices for clinical implementation [9]. Moreover, while AI and machine learning techniques have shown promise in other areas of healthcare, their integration into digital phenotyping remains limited. Many studies focus on the feasibility of using wearable technologies but fail to address how AI can enhance predictive accuracy or improve the clinical applicability of these tools [11]. Additionally, most studies have focused on mental health conditions like depression and anxiety, with underexplored conditions such as schizophrenia, OCD, and PTSD receiving less attention [12, 13]. This gap in research prevents a comprehensive understanding of how digital phenotyping can be applied across a broader range of psychiatric disorders.

Another significant challenge is user adherence to wearable technologies. While these tools offer valuable data, adherence and engagement are critical for ensuring that the data collected is meaningful and actionable. Factors such as privacy concerns, ethical considerations, and the long-term sustainability of using wearable devices for mental health monitoring also remain significant obstacles [14]. Addressing these issues is crucial for ensuring that digital phenotyping can transition from an experimental research tool to a validated, scalable solution in everyday clinical practice.

1.2 Proposed Solution

This paper aims to address the identified gaps by providing a systematic review of digital phenotyping in mental health, specifically focusing on the use of wearable devices and smartphone applications. The review will analyze studies that explore the effectiveness of these technologies, particularly in terms of their predictive accuracy, the integration of AI algorithms, and their application to a wider range of mental health disorders, including schizophrenia, OCD, and PTSD. By synthesizing the findings of over 60 studies, this review will provide actionable insights into the strengths and limitations of current approaches and propose solutions for overcoming the challenges of data interpretation, user engagement, and AI integration. It will also explore how real-time digital interventions impact long-term clinical outcomes, offering guidance on how to optimize these technologies for use in clinical settings [15, 16, 17, 18].

Unlike prior reviews that primarily focused on feasibility and device classification, this study provides a structured comparison of predictive model performance, validation strategies, and personalization capabilities across mental health conditions. This positions the present review as a performance-oriented and clinically focused synthesis of digital phenotyping research.

2. RESEARCH METHOD

2.1 Literature and Search Strategy

To systematically review the role of wearable and smartphone technologies in digital phenotyping for mental health disorders, a structured search was conducted across PubMed, Scopus, Web of Science, and Google Scholar. Search queries were developed using Boolean operators and included keywords such as "digital phenotyping," "mental health," "wearable," "smartphone," "technology," and "monitoring" to ensure comprehensive coverage of relevant literature.

The search was restricted to peer-reviewed articles published between 2016 and 2025 to reflect advancements in digital mental health. Only articles in English were considered. Additional studies were identified via citation tracking and manual searches in relevant journals and conference proceedings.

To enhance specificity, the selection process followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, ensuring a structured workflow:

(i) Initial search retrieved 185 articles across databases.

(ii) Duplicates were removed (30 studies eliminated).

(iii) Title and abstract screening reduced the selection to 62 articles.

(iv) Full-text assessment resulted in 16 final studies that strictly met the inclusion criteria.

The structured search across databases yielded varying results, as summarized in Table 1, which presents the number of studies retrieved per keyword and database.

Table 1. Summary of Search Strategy Output

Database	Keyword	Number of Studies Retrieved
PubMed	smartphone	35
PubMed	mental health	30
PubMed	monitoring	28
Scopus and others	digital phenotyping	26
Scopus and others	wearable	24
Web of Science	technology	20
Google Scholar	digital phenotyping	12
Preprint Repository	digital phenotyping	10

2.2 Study Selection and Outcomes

The study selection process adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure transparency, reproducibility, and methodological rigor in synthesizing research on digital phenotyping of mental health disorders using wearable smartphone technologies.

2.2.1. Study Identification and Screening

The systematic search across academic databases yielded 185 articles. After duplicate removal (30 articles), title and abstract screening led to a refined selection of 62 articles. Full-text analysis was then performed, resulting in 15 final studies that met all predefined inclusion criteria. Figure 1 illustrates the PRISMA-based flow diagram for article selection, detailing each phase of screening and exclusion.

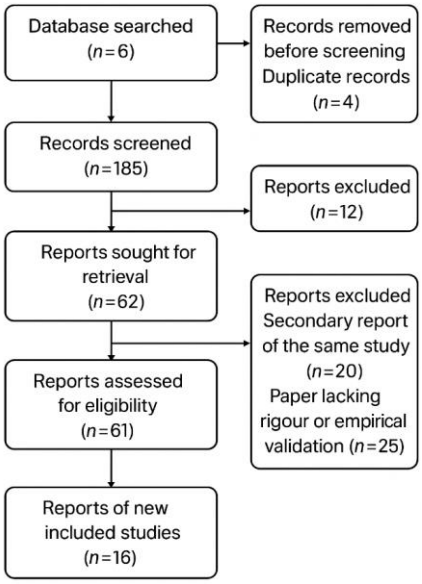

Figure 1. PRISMA Flow Diagram for Selection Process

2.2.2 Inclusion and Exclusion Criteria

To ensure relevance and methodological rigor, studies were included if they met all of the following criteria: a) Explicit focus on digital phenotyping of mental health disorders. b) Novel methodologies, benchmarks, or applications of wearable smartphone technologies in mental health. c) Empirical evaluation of effectiveness, usability, or adoption of these technologies (e.g., pre-post intervention comparisons, machine learning validation, real-world applications). And, d) Peer-reviewed journal publication in English (2016–2025).

Studies were excluded if they: a) Did not focus on digital phenotyping or mental health disorders. b) Did not involve wearable or smartphone-based technologies. c) Lacked empirical evidence (e.g., concept papers, theoretical discussions, non-clinical evaluations) and d) Were duplicate or secondary reports of the same study.

2.2.3 Primary and Secondary Outcomes

The primary outcomes assessed were: a) Feasibility – Evaluating practicality and real-world application of digital phenotyping tools. b) Accuracy – Measuring reliability of data from smartphones, wearables, and machine learning models and c) Clinical Relevance – Assessing the impact of digital biomarkers on diagnosis, symptom tracking, and treatment interventions.

The secondary outcomes focused on: a) User Engagement & Adherence – Evaluating acceptance and sustained usage of wearable and smartphone-based monitoring systems. b) Ethical Considerations – Examining data privacy, security concerns, and regulatory compliance in digital phenotyping applications.

In order to be further analyzed, articles needed to: a) Focus explicitly on digital phenotyping of mental health disorders. b) Present novel methodologies, benchmarks, or applications involving wearable smartphone technologies in mental health contexts. c) Report empirical studies evaluating the effectiveness, usability, or adoption of these technologies (e.g., pre- and post-intervention scores). d) Be published in a peer-reviewed academic journal written in English, dated from 2016 to 2025.

Articles were excluded if they: a) Were not relevant to the focus of digital phenotyping or mental health disorders. b) Did not address the application of wearable smartphone technologies in mental health. c) Did not include empirical evaluation (e.g., articles that only introduced and/or discussed concepts or research protocols). d) Were secondary reports of the same study.

2.3 Data Extraction and Synthesis

The data extraction process was systematically structured to ensure objectivity, and comprehensive coverage of studies on digital phenotyping of mental health disorders using wearable smartphone technologies. A two-phase review approach was employed:

1. Preliminary screening: Title and abstract review (n = 61).

2. Full-text evaluation: Final inclusion of studies (n = 16) based on eligibility criteria.

To minimize bias and enhance reliability, the review process was conducted independently by multiple researchers, followed by cross-validation to ensure consensus. No additional eligible articles were identified through forward or backward reference searches.

For each study, the following details were systematically extracted and analyzed:

a) Publication Information: Year, authors, and country.

b) Study Design and Setting: Research methodology and participant demographics.

c) Intervention Details: Specific wearable or smartphone technology used.

d) Outcome Measures: Key findings on effectiveness, accuracy, and feasibility.

e) Limitations: Constraints impacting generalizability and clinical applicability.

Table 2 provides a comparative overview of eight high-relevance studies, highlighting methodological approaches, sample characteristics, and key findings.

Table 2: Benchmark-Oriented Summary of Key Studies on Digital Phenotyping.

#	Author (Year)	Algorithm Type	Input Data	Performance	Validation type	Clinical Applications	Key Findings	Limitations
1	Rashid et al. (2021)	Supervised Learning	Smartphone + Wearables	78% accuracy	Internal validation	Schizophrenia relapse prediction	Reliable biomarkers identified for relapse	Small sample size, limited diversity
2	Zhang et al. (2023)	ML + Statistical Models	Smartphone sensor data	High predictive reliability	Cross-validation + real-world data	Mood disorder tracking	Accurate behavioral tracking in clinical settings	Sensor limitations

3	Jacobson et al. (2019)	Regression Models	Passive smartphone data	r = 0.72 (correlation)	Internal validation	Social anxiety detection	Strong correlation with clinical scores	Limited scope (anxiety only)
4	Busk et al. (2020)	Bayesian Model	Smartphone self-assessment	82% accuracy	Internal validation	Bipolar mood prediction	Highest predictive performance observed	Algorithm bias, data dependency
5	Dlima et al. (2022)	Data Analytics + ML	Wearable + real-time monitoring	Moderate performance	Not clearly defined	Early symptom detection	Real-time tracking improves intervention timing	Data quality variability
6	Martinez-Martin et al. (2021)	Conceptual / Ethical Framework	N/A	Not applicable	Not applicable	Ethical governance	Proposed ethical safeguards for digital phenotyping	No empirical validation
7	Frank et al. (2023)	Descriptive Analysis	Wearables	Moderate effectiveness	Observational	OCD monitoring	Continuous tracking feasible	Lack of algorithmic depth
8	Hassan et al. (2025)	Statistical Evaluation	Consumer-grade wearables	Moderate accuracy	Observational	General mental health monitoring	Consumer devices viable but limited	Hardware reliability, adherence issues

Table 2 presents a benchmark-oriented comparison of digital phenotyping studies, highlighting algorithmic approaches, input modalities, validation strategies, and clinical applications to enable reproducibility and cross-study evaluation.

2.4 Quality Assessment

We developed a custom quality assessment framework for this systematic review to evaluate methodological quality and assess risk of bias. Using the Newcastle–Ottawa Scale (NOS) with a star rating system (0–9), we categorized studies as high risk (0–3), medium risk (4–6), or low risk (7–9). This framework was chosen because it suits non-randomized studies common in research on intelligent digital systems. Although we considered other tools, like the Cochrane Risk of Bias Tool and ROBINS-I, we did not use them. Among the studies, seven were identified as medium risk. These studies featured strong cohort representativeness and reliable intervention ascertainment; however, they were marred by a lack of long-term follow-up or control groups. An example of a medium-risk study is, which effectively utilized machine learning to identify mood biomarkers [19]. Yet, the absence of a control group diminished the reliability of the results. Conversely, nine studies were deemed high risk. These studies were characterized by small sample sizes, reliance on self-reported data without independent validation, and significant variability in methodological rigor. A representative high-risk study is [4], which tested smartphone-based relapse prevention in bipolar disorder. Unfortunately, this study was flawed by the lack of a control group and the absence of long-term data. As shown in Table 3, studies varied in quality, with scores assigned based on sample representativeness, outcome assessment, and follow-up adequacy.

Table 3: Risk of Bias Assessment Using the Newcastle-Ottawa Scale (NOS)

Author	Stars (0–9)	Representativeness of Exposed Cohort	Selection of Non-Exposed Cohort	Outcomes Before and After Intervention	Comparability of Cohorts	Assessment of Outcome	Follow-Up Long Enough	Adequacy of Follow-Up
Rashid et al [19, 20]	7	★	★	★★	★★	★	★	★
Zhang et al. [21]	9	★	★	★★	★★	★	★	★
Jacobson et al. [15]	8	★	★	★★	★★	★	★	
Frank et al. [14]	5	★	★	★★	★			
Martinez-Martin et al. [1]	5	★	★	★	★		★	
Busk et al. [9]	3	★		★	★			
Hassan [16]	2	★				★		

3. RESULTS AND DISCUSSION

This systematic review analyzed 16 peer-reviewed studies that explored the use of wearable and smartphone-based technologies for digital phenotyping of mental health disorders. A range of study designs was represented, including pre-post within-subject experiments, quasi-experimental approaches, and two randomized controlled trials (RCTs). The geographical scope spanned Europe, North America, and Asia, suggesting global interest in deploying digital phenotyping for precision psychiatry. The geographical spread and methodological diversity of reviewed studies are visualized in Figure 2, and Figure 3 underscoring the global interest in digital phenotyping. Figure 2 illustrates that the majority of studies originate from Europe and North America, indicating a geographic concentration of research efforts and a lack of representation from low- and middle-income regions.

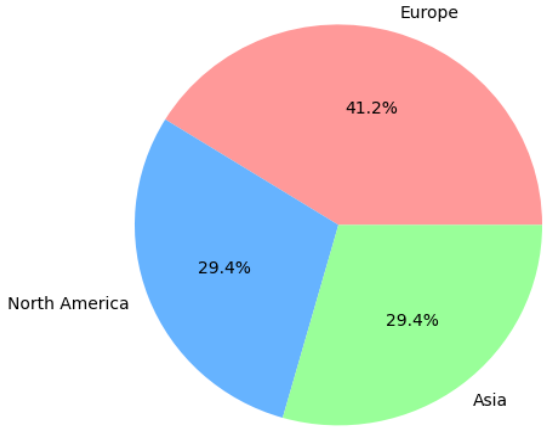

Figure 2. Geographical distribution

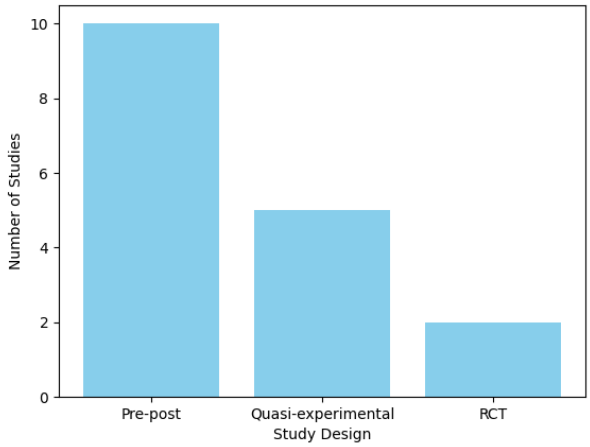

Figure 3. Study Designs

Several studies employed advanced computational methods. Rashid et al. [19, 20] utilized supervised learning models for relapse prediction in schizophrenia, achieving 78% accuracy. Busk et al. [9] employed a hierarchical Bayesian model for mood forecasting in bipolar disorder, while Jacobson et al. [5] applied regression models to identify anxiety biomarkers. Deep learning and ensemble methods, though promising, were underreported across studies, indicating a gap in the technical application of AI within digital phenotyping.

Among studies incorporating AI, predictive accuracy varied between 72% and 82%, yet most lacked external validation, making generalizability a concern. Additionally, while depression and anxiety algorithms have matured, OCD and schizophrenia models remain in early exploratory phases.

3.1 Key Findings and Algorithmic Methods

This section reports the empirical findings of the reviewed studies, while interpretative analysis is presented separately in the Discussion section.

Among the reviewed studies, 11 employed machine learning or statistical modeling techniques to analyze passive and active digital biomarkers collected via wearable devices and smartphones. Notably:

Busk et al. [9] implemented a hierarchical Bayesian model to forecast mood episodes in bipolar patients, achieving an accuracy of 82%—demonstrating the potential of structured probabilistic reasoning in psychiatric forecasting.

Jacobson et al. [5] applied regression-based machine learning models to identify digital markers of social anxiety, with passive smartphone sensor data correlating with clinical anxiety scores ($r = 0.72$, $p < 0.01$).

Rashid et al. [19] used supervised learning to predict schizophrenia relapse, reporting a model accuracy of 78% in the HOPE-S study.

Comparative analysis showed that Busk et al. [9] reported the highest predictive performance (82%), followed by Rashid et al. [19] (78%), while regression-based approaches demonstrated more modest predictive strength (correlation coefficient $r = 0.72$ in Jacobson et al. [5]). These findings indicate that probabilistic and supervised learning models outperform traditional regression-based methods in current digital phenotyping applications

However, algorithmic transparency was limited across several studies. Only 6 of the 16 included papers clearly described model training, validation procedures, or hyperparameter tuning. Deep learning approaches and ensemble models increasingly common in behavioral informatics were scarcely reported, indicating an area of underutilization. These findings are synthesized in Table 2, which summarizes each study's methodology, outcomes, and limitations.

3.2 Clinical Performance and AI-Driven Personalization

Smartphone-based cognitive behavioral therapy (CBT) apps and passive sensing tools showed the strongest clinical impact in mood and anxiety disorders. In the MONARCA II RCT ([22]), bipolar patients who received real-time mood feedback via a smartphone app experienced statistically significant improvements in mood awareness and clinical outcomes. Similarly, Hassan et al. [16] evaluated consumer-grade wearables, finding moderate efficacy in symptom tracking but noting hardware reliability issues and high dropout rates.

Despite these encouraging results, only a minority of studies implemented adaptive feedback loops or real-time intervention triggers which are critical features for clinical personalization. This underscores a missed opportunity in integrating AI not just for prediction but for dynamic decision support [23, 24, 25].

3.3 Ethical Considerations, Technical Constraints, and User Experience

Across the reviewed literature, common barriers to effective implementation included:

(i) Limited sample sizes (e.g., 12 to 20 participants in [4, 5]).

(ii) Poor device interoperability leading to inconsistent data collection ([11, 16])

(iii) Algorithmic bias, particularly models trained predominantly on homogeneous demographic cohorts [8, 9].

(iv) Privacy concerns, especially regarding passive monitoring and the need for updated regulatory frameworks [1].

(v) User adherence remained volatile, affected by app fatigue, intrusive notifications, or unclear feedback mechanisms. For instance, the MONARCA II trial reported a 31% dropout rate, while two studies noted usability issues due to technical lags and sensor drift [6].

3.4 Comparative Evaluation and Research Gaps

While most studies demonstrated feasibility and initial effectiveness, comparative evaluations were rare. For example, although Zhang et al. [21] and Rashid et al. [20] both utilized the RADAR-based platform, only the former reported validation metrics across diverse populations, whereas the latter focused narrowly on schizophrenia symptom tracking. Such contrasts highlight the need for benchmarking across platforms and diagnoses. Compared to prior reviews such as Dlima et al. [11], which focused primarily on feasibility and device types, this review expands the scope by integrating ethical analysis and AI-driven personalization. While previous studies reported predictive accuracies ranging from 65% to 75% [11], our synthesis found that several models achieved accuracies above 78% [19, 9], suggesting improved algorithmic performance in recent years. However, unlike earlier reviews, we also highlight the lack of external validation and the underutilization of deep learning methods, which remain critical gaps in the field [23, 24].

Unlike prior reviews such as Dlima et al. [11], which primarily focused on feasibility and device classification, our study expands the analytical scope by integrating ethical considerations [7], AI-driven personalization [6] and underrepresented psychiatric conditions such as OCD and PTSD [5]. Earlier reviews reported predictive accuracies ranging from 65% to 75% [11], our synthesis identified models achieving accuracies above 78% [19], with some reaching as high as 82% [9].

Methodologically, this review applies a structured risk-of-bias assessment using the Newcastle–Ottawa Scale (Table 3) [26], and includes comparative scoring across 16 high-relevance studies (Table 2), offering a more rigorous and reproducible evaluation framework than narrative-only syntheses. Furthermore, our inclusion of RADAR-base–enabled studies [6] and attention to ethical frameworks [7] distinguishes this review as both technically and socially comprehensive.

Despite these advances, notable gaps persist. Certain psychiatric conditions namely schizophrenia, OCD, and PTSD remain underrepresented in algorithm validation studies. Additionally, the limited use of multimodal data fusion (e.g., combining accelerometry with audio or speech signals) constrains diagnostic granularity and model generalizability.

When compared with earlier reviews reporting predictive accuracies of 65–75% [11], the studies synthesized in this review demonstrate improved model performance, with several exceeding 78%, suggesting measurable progress in algorithmic capability over time.

3.5 Revised Contribution and Implications

This study advances the field of digital phenotyping by introducing a structured performance-oriented evaluation framework for wearable and smartphone-based mental health monitoring systems. Unlike prior reviews that primarily emphasize feasibility or device categorization, this work systematically integrates algorithmic performance, validation strategies, and clinical applicability into a unified analytical perspective.

Specifically, this paper makes four key contributions:

i. **Performance-Centric Synthesis:** A comparative evaluation of machine learning and statistical models across multiple mental health conditions, highlighting predictive accuracy, validation approaches, and generalizability limitations.

ii. **Computational Characterization:** A structured mapping of digital phenotyping pipelines, including data acquisition, feature extraction, model development, and clinical output layers.

iii. **Benchmark-Oriented Analysis:** Identification of gaps in external validation, reproducibility, and multimodal data fusion, with emphasis on their impact on real-world deployment.

iv. **Ethical–Computational Integration:** A unified discussion of algorithmic bias, privacy constraints, and user adherence within AI-driven mental health systems.

This positions the present study not merely as a descriptive review, but as a methodologically grounded framework for evaluating and designing next-generation digital phenotyping systems.

As shown in Figure 2, research efforts are globally distributed, but standardization and reproducibility remain core challenges. Addressing these through more robust algorithm reporting and ethically grounded design will be essential for translating digital phenotyping from pilot trials into mainstream psychiatric care.

3.6 Computational Framework for Digital Phenotyping Systems

To address the lack of standardization in digital phenotyping methodologies, this study proposes a generalized computational framework that captures the end-to-end pipeline of wearable and smartphone-based mental health monitoring systems. The proposed framework consists of four core layers (Figure 4):

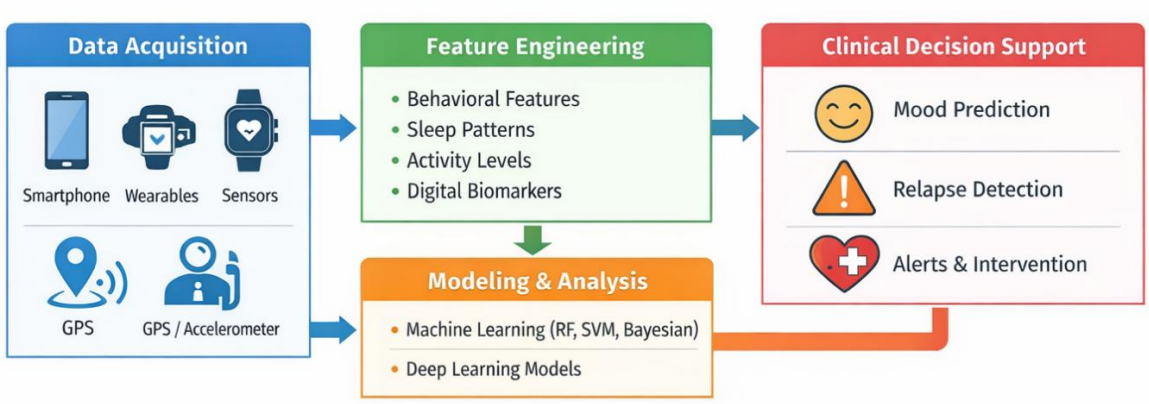

Figure 4: Computational Pipeline for Digital Phenotyping in Mental Health Monitoring. The framework integrates multi-source data acquisition, feature engineering, machine learning modeling, and clinical decision support for scalable and reproducible mental health systems

As illustrated in Figure 4, the proposed pipeline integrates multi-source data acquisition, feature extraction, machine learning modeling, and clinical decision support to enable scalable and reproducible digital mental health systems.

i. **Data Acquisition Layer:** The data acquisition layer captures multimodal inputs from wearable devices and smartphones, including physiological signals, mobility patterns, and user interaction data.

ii. **Feature Engineering Layer:** These raw signals from (i) are transformed in the feature engineering layer into meaningful behavioral and temporal features such as mobility patterns and sleep cycles, temporal features**,** temporal features (circadian rhythm) and derived biomarkers.

iii. **Modeling Layer:** The modeling layer applies machine learning techniques such as regression models, random forest, Bayesian models, and deep learning (underutilized) and statistical algorithms to detect patterns and predict mental health states.

iv. **Clinical Decision Layer:** Finally, the clinical decision layer translates model outputs into actionable insights, such as mood prediction, relapse alerts, and adaptive intervention strategies. This structured pipeline enhances reproducibility, supports cross-study comparability, and facilitates integration into practical clinical workflows.

Figure 4 provides a standardized computational perspective for analyzing digital phenotyping systems, enabling reproducibility, comparability, and improved clinical integration.

4. CONCLUSION

This systematic review synthesized findings from 62 studies examining the role of smartphone- and wearable-based digital phenotyping in mental health. Sixteen high-relevance studies were analyzed in depth, revealing promising advances in AI-driven symptom monitoring, ethical design integration, and disorder-specific applications. The review highlights several key trends, including increased predictive accuracy (≥78%), growing adoption of multimodal data inputs, and the gradual inclusion of underrepresented disorders such as OCD and PTSD.

However, significant gaps remain. Few studies demonstrate external validation across diverse populations, and the underutilization of deep learning and multimodal fusion constrains diagnostic performance. Moreover, many models lack transparency and reproducibility, limiting their clinical utility. Our findings underscore the need for standardized benchmarking frameworks, broader disorder coverage, and a stronger emphasis on ethical and equitable implementation.

By combining methodological rigor with a socially attuned lens, this review offers a roadmap for advancing digital phenotyping from prototype to practice. Future research should prioritize cross-platform comparability, transparent algorithm design, and real-world clinical validation to ensure that digital mental health tools are scalable, inclusive, and clinically robust.

5. ACKNOWLEDGMENTS

The authors thank the mental health physicians and biomedical experts consulted in the course of the research.

6. REFERENCES

[1] Martinez-Martin, N., H. T. Greely, and M. K. Cho, 2021. Ethical development of digital phenotyping tools for mental health applications: Delphi study, *JMIR Mhealth Uhealth*, 9(3), p. e27343, URL: https://doi.org/10.2196/27343

[2] Insel, T. R. 2017. Digital phenotyping: Technology for a new science of behavior. *JAMA*, 318(13), 1215–1216, URL: https://doi.org/10.1001/jama.2017.11295

[3] Teo, J. X., Davila, S., Yang, C., Pua, C. J., Yap, J., Tan, S. Y., & Yeo, K. K. 2019. Digital phenotyping by consumer wearables identifies sleep-associated markers of cardiovascular disease risk and biological aging. *Communications Biology*, 2(1), 361. https://doi.org/10.1038/s42003-019-0605-1

[4] Beiwinkel, T., Kindermann, S., Maier, A., Kerl, C., Moock, J., Barbian, G., & Rössler, W. 2016. Using smartphones to monitor bipolar disorder symptoms: A pilot study. *JMIR Mental Health*, 3(1), e2., URL: https://doi.org/10.2196/mental.4560

[5] Jacobson, N. C., B. J. Summers, and S. Wilhelm. 2019. Digital biomarkers of social anxiety severity: Digital phenotyping using passive smartphone sensors, *J. Med. Internet Res.*, 21(12), p. e16875, URL: https://doi.org/10.2196/16875

Faurholt-Jepsen, M., et al., 2019. The effect of smartphone-based monitoring on illness activity in bipolar disorder: The MONARCA II randomized controlled

single-blinded trial, *Psychol. Med*., 49(11),1973–1983, URL: https://doi.org/10.1017/S0033291719000710

[6] Albrechta, H., Goodman, G. R., Oginni, E., Mohamed, Y., Venkatasubramanian, K., Dumas, A., Carreiro, S., Lee, J. S., Glynn, T. R., O'Cleirigh, C., Mayer, K. H., Fisher, C. B., & Chai, P. R. 2024. Acceptance of digital phenotyping linked to a digital pill system to measure PrEP adherence among men who have sex with men with substance use. *PLOS Digital Health*, 3(2), e0000457., URL: https://doi.org/10.1371/journal.pdig.0000457

[7] Bourla, A., et al. 2018. E-psychiatry: New technologies for mental health management. *Frontiers in Psychiatry*, 9, 123.URL: https://doi.org/10.3389/fpsyt.2018.00051

[8] Busk, J., Faurholt-Jepsen, M., Frost, M., Bardram, J., Vedel Kessing, L., & Winther, O. 2020. Forecasting mood in bipolar disorder from smartphone self-assessments: Hierarchical Bayesian approach. *JMIR mHealth and uHealth,* 8(4), e15028. URL: https://doi.org/10.2196/15028

[9] Santis, K. D., Mergenthal, L., Christianson, L., Busskamp, A., Vonstein, C., and Zeeb, H. 2023. Digital technologies for health promotion and disease prevention in older people: Scoping review. *Journal of Medical Internet Research*, 25, e43542. URL; https://doi.org/10.2196/43542

[10] Dlima, S. D., Shevade, S., Menezes, S. R., & Ganju, A. 2022. Digital phenotyping in health using machine learning approaches: Scoping review. *JMIR Bioinformatics and Biotechnology*, 3(1), e39618. URL: https://doi.org/10.2196/39618

[11] Faurholt-Jepsen, M., Frost, M., Ritz, C., Christensen, E. M., Jacoby, A. S., Mikkelsen, R. L., Knorr, U., Bardram, J. E., Vinberg, M., and Kessing, L. V. 2015. Daily electronic self-monitoring in bipolar disorder using smartphones - The MONARCA I trial: A randomized, placebo-controlled, single-blind, and parallel group trial. *Psychological Medicine*, 45(13), 2691-2704. URL: https://doi.org/10.1017/S0033291715000410

[12] Ferreri, F., Bourla A, Mouchabac and S, Karila L. 2018. e-Addictology: An Overview of New Technologies for Assessing and Intervening in Addictive *Behaviors. Front Psychiatry*. 9(51), URL: https://doi.org/10.3389/fpsyt.2018.00051.

[13] Frank, A. C., Li, R., Peterson, B. S., and Narayanan, S. S. 2023. Wearable and mobile technologies for the evaluation and treatment of obsessive-compulsive disorder: Scoping review. *JMIR Mental Health*, 10, e45572. URL: https://doi.org/10.2196/45572

[14] Gardea-Resendez, M., Breitinger, S., Walker, A., Harper, L., Xiong, A., Stoppel, C., Volety, R. M., Raman, J., Byun, J. S., Langholm, C., Goes, F. S., Zandi, P. P., Torous, J., and Frye, M. A. 2024. Digital technologies tracking active and passive data collection in depressive disorders: Lessons learned from a case series. *Journal of Psychiatric Practice*, 30(6), 434-439. URL: https://doi.org/10.1097/PRA.0000000000000820

[15] Hassan, L., Milton, A., Sawyer, C., Casson, A. J., Torous, J., Davies, A., Ruiz-Yu, B., and Firth, J. 2025. Utility of consumer-grade wearable devices for inferring physical and mental health outcomes in severe mental illness: Systematic review. *JMIR Mental Health*, 12, e65143. URL: https://doi.org/10.2196/65143

[16] Jayakumar, P., Lin, E., Galea, V., Mathew, A. J., Panda, N., Vetter, I., & Haynes, A. B. 2020. Digital phenotyping and patient-generated health data for outcome measurement in surgical care: A scoping review. *Journal of Personalized Medicine,* 10(4), 282. URL: https://doi.org/10.3390/jpm10040282

[17] Onnela, J. P., and Rauch, S. L. 2017. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Translational Psychiatry*, 7(7), e1013. URL: https://doi.org/10.1038/tp.2017.25

[18] Rashid, N. A., et al., 2021. Evaluating the utility of digital phenotyping to predict health outcomes in schizophrenia: Protocol for the HOPE-S observational study, *BMJ Open*, 11(10), p. e046552, URL: https://doi.org/10.1136/bmjopen-2020-046552

[19] Rashid, Z., Folarin, A. A., Zhang, Y., Ranjan, Y., Conde, P., Sankesara, H., Sun, S., Stewart, C., Laiou, P., and Dobson, R. J. B. 2024. Digital phenotyping of mental and physical conditions: Remote monitoring of patients through RADAR-base platform. *JMIR Mental Health*, 11, e51259. URL: https://doi.org/10.2196/51259

[20] Zhang, Y., Stewart, C., Ranjan, Y., Conde, P., Sankesara, H., Rashid, Z., Sun, S., Dobson, R. J. B., and Folarin, A. A. 2024. Large-scale digital phenotyping: Identifying depression and anxiety indicators in a general UK population with over 10,000 participants. *arXiv*. URL: https://arxiv.org/abs/2409.16339

[21] Winslow, B., and Mills, E. 2023. Wearables for stress management in military health. *BMJ Military Health*, 169(2), 123–132. URL: https://doi.org/10.1136/bmjmilitary-2022-002306

[22] Torous, J., Kiang, M. V., Lorme, J., & Onnela, J. P. 2018. New tools for new research in psychiatry: A scalable and customizable platform to empower data-driven smartphone research. *JMIR Mental Health*, 5(2), e16. URL: https://doi.org/10.2196/mental.9785

[23] Torous, J., & Onnela, J. P. 2020. Digital phenotyping of mental health: Meaning, challenges, and opportunities. *Current Opinion in Psychiatry*, 33(5), 464-468. URL: https://doi.org/10.1097/YCO.0000000000000633

[24] Doryab A, Villalba D, Chikersal P, Dutcher J, Tumminia M, Liu X, Cohen S, Creswell K, Mankoff J, Creswell J, Dey A. 2019. Identifying Behavioral Phenotypes of Loneliness and Social Isolation with Passive Sensing: Statistical Analysis, Data Mining and Machine Learning of Smartphone and Fitbit Data. *JMIR Mhealth Uhealth*; 7(7):e13209. URL: https://doi.org/10.2196/13209

[25] Wells, G.A., Wells, G., Shea, B., Shea, B., O'Connell, D., Peterson, J., Welch, Losos, M., Tugwell, P., Ga, S.W., Zello, G.A., & Petersen, J. A. 2015. *The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Non randomised Studies in Meta-Analyses*, Science Open Inc., USA, URL: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.