iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78572

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Early Detection of Mental Health Disorders using Explainable AI and Digital Behavioral Signals

Gaurav Borthakur
Computer Science and Engineering, The Assam Kaziranga University

Abstract: ***Mental health disorders such as depression and anxiety are major global health concerns, yet many cases remain undetected in their early stages. With the widespread use of smartphones and digital platforms, individuals generate continuous behavioural data through daily activities such as communication, mobility, and online interactions. These digital behavioural signals can provide valuable insights into changes in mental well-being. This study explores the use of explainable artificial intelligence (XAI) to analyse digital behavioural data for the early detection of mental health risks. Data derived from smartphone usage patterns, activity levels, and textual content from online platforms can be examined using machine learning techniques to identify behavioural changes associated with psychological distress. Explainable AI methods are incorporated to ensure that the reasons behind model predictions are transparent and understandable. The research also emphasizes ethical considerations including user privacy, informed consent, and fairness in algorithmic analysis. Overall, this approach highlights the potential of combining digital behavioural signals with explainable AI to support early mental health screening while maintaining responsible and ethical use of technology.***

Keywords: Mental health, AI / Machine learning, Digital Behaviour data, Early detection

I. INTRODUCTION AND BACKGROUND

Mental health disorders (such as depression, anxiety, bipolar disorder, PTSD, and schizophrenia) affect a large portion of the population. For example, one WHO report finds about 1 in 8 people worldwide live with a mental disorder[1], and about 1 in 5 US adults experience mental illness. Depression alone impacts ~322 million globally[6]. Untreated mental illness is a leading cause of disability[7] and contributes to chronic disease and mortality[7]. Early detection is therefore critical for timely care.

Traditionally, mental health is assessed by interviews and questionnaires. However, digital health technologies are creating new possibilities. Today's smartphones, wearables (watches, rings), and online platforms passively collect detailed data about a person's behavior every day[8].

For example: GPS location tracks movement; accelerometers measure physical activity; phone logs show sleep patterns and communication; social media posts reveal language and sentiment. These continuous measurements are often called digital behavioral signals. When carefully analyzed, they can serve as digital biomarkers (objective measures reflecting health) and can be combined into a digital phenotype of mental state[2][3].

One commonly used definition of digital phenotyping is: *"moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices"*[3]. In practice, this means inferring a person's mood or stress from their device usage patterns (e.g. fewer daily locations or more night-time phone use may indicate depression[9]). The idea is analogous to how doctors use blood pressure as a biomarker of cardiovascular health; here, phone/wearable data become biomarkers of mental health[2][3].

Recent trends are bringing these ideas into focus. Ownership of smartphones and wearables is now very high, enabling large-scale data collection[8]. Advances in AI and machine learning allow complex pattern recognition in this data. At the same time, there is a push in healthcare toward preventative care and remote monitoring. If digital biomarkers can provide early warning signs, clinicians could intervene sooner.

II. KEY MENTAL HEALTH TARGETS AND CHALLENGES

The most common targets for digital detection are depression and anxiety, as they are highly prevalent[11]. Bipolar disorder and PTSD are also studied, often through language cues (text sentiment)[12]. Schizophrenia and eating disorders are studied less in this context, partly due to data and complexity.

A. Challenges in this Domain Include

1) Data variability: Unlike a lab test, behavior varies with context and person. Models must capture relevant changes while ignoring normal fluctuations[13].

2) Privacy: Mental health data are sensitive. Ethical use requires consent, anonymization, and secure storage[14].

3) Bias: Training data often come from volunteers (e.g. students) that may not represent all groups.Models risk under-performing on minorities or older adults[15][16].

4) Interpretability: Clinicians need to know *why* a system gave a risk alert, to trust and act on it. Hence, explainable AI methods are critical[17][18].

In summary, the premise is that explainable digital phenotyping could transform mental health care by complementing traditional methods with continuous, real-time monitoring[2][17]. The rest of this report reviews recent research developments and proposes how to build such a system responsibly.

III. LITERATURE REVIEW (2018–2025)

Recent studies have used various data sources and AI methods to detect mental health status. We summaries key works below, organized by data modality. (Table 1 compares representative studies.)

A. Social Media Text

1) Hameed et al. (2025) studied depression detection from Twitter/X posts[19]. They collected publicly available tweets, extracted features (TF-IDF, word embeddings, sentiment), and trained models (SVM, Random Forest, Neural Nets). They found the SVM gave the best accuracy, and crucially used LIME (Local Interpretable Model-agnostic Explanations) to highlight which words drove each prediction[19]. This allowed them to point out linguistic markers (e.g. “sadness” words) aligned with clinical signs.

2) Kerz et al. (2023) focused on multiple conditions (ADHD, anxiety, bipolar, depression, stress) using Reddit posts[20]. They engineered interpretable language features (LIWC categories, sentiment, topics) and also fine-tuned a BERT model. Their multi-task learning approach improved accuracy ~9–13% over single-task baselines[21]. Explainability was built in: they applied LIME and an “AGRAD” method to show which linguistic patterns were most predictive (e.g. certain pronoun use, emotion words)[20].

3) Jain et al. (2024) created a behavioral features dataset from social media usage (not text content)[22][23]. Their “Social Media Mental Health” dataset (2023) recorded users’ screen time, number of posts, passive scrolling behavior, etc. They trained machine learning models (XGBoost, Random Forest) to predict depression or anxiety scores. In cross-validation they achieved AUC ≈0.87 (depression) and ≈0.85 (anxiety); on an external test set AUC fell to ~0.69–0.72[24]. They then applied SHAP and LIME: global SHAP analysis identified *daily screen time*, *passive scroll time*, and *night-time usage* as top predictors[25]. Local LIME explanations for individual users showed personalized factors (e.g. “excessive night use” explained a high-risk flag)[26].

B. Smartphone Sensors and Wearables

1) Agarwal et al. (2021) conducted a large multi-country study (629 participants) using smartphone and wearable data to predict depression[4]. They collected ~3 weeks of sensor data (GPS, motion, app usage, connectivity) along with PHQ-8 surveys. Using 22 aggregated features (mobility patterns, screen time, social contact), they trained several models (Random Forest, SVM, etc.). Remarkably, the top models achieved ~96–98% accuracy and AUCs of 94–99% for classifying depressive vs non-depressive status[4]. Permutation importance ranked screen/internet features (duration of connectivity) highest. They noted, however, that such high accuracy may reflect the particular sample and definitions used. They argued that “behavioral markers indicative of depression can be identified unobtrusively”[4].

2) Doryab et al. (2019) examined loneliness as a proxy for depression[5]. Over 10 weeks, they collected smartphone and Fitbit data from 160 participants in Boston (activitiy, sleep, location, communication). Using gradient boosting, they classified high vs low loneliness (based on questionnaire) with 80.2% accuracy[5]. Top features included reduced physical activity and communication volume for lonely individuals. This study showed proof-of-concept that passive sensing can detect social–emotional states.

3) Bai et al. (2025) performed a campus study with smartphone sensors[27]. Twelve college students carried phones for ~2 months, and researchers collected accelerometer, gyroscope, and light sensor readings. Using Pearson correlations and ML, they found significant negative correlations between PHQ-9 depression scores and daily physical activity, sleep regularity, and dietary regularity. Classification models achieved 73–88% accuracy for depression presence[27]. The study concluded smartphone sensors hold promise for early detection, though small sample limits generalizability.
4) Aalbers et al. (2026) conducted a digital phenotyping study of anxiety and depression in Amsterdam[28]. They used the "Behapp" app to passively track GPS (locations) and app usage from 217 adults over several weeks. They derived 46 features (mobility patterns, app preferences, communication times) and collected PHQ-9 and GAD-7 scores. A key finding was that individuals with depression/anxiety had fewer daily location "trajectories" (they visited fewer new places)[29]. A model using the top features plus demographics achieved only modest discrimination (AUC≈0.60) – much lower than many previous studies. The authors suggested limited digital signal in some populations, highlighting that high promise in theory needs robust testing.

C. *Audio and Other Modalities*

Most digital studies focus on text and sensor data. Voice and video analysis for mental health (e.g., analyzing speech or facial expressions) exist but are less common in deployment due to privacy and data requirements. A survey of digital biomarkers notes that wearable and smartphone sensors still dominate the field[30]. For completeness, we note datasets like the DAIC-WOZ corpus (clinical interviews with audio transcripts) are publicly used for depression detection in research (though require patient consent).

III. ETHICAL, PRIVACY AND BIAS CONSIDERATIONS

1) Privacy & Consent: Mental health data are highly sensitive. Ethical use demands *informed consent*, data minimization, and strong security. For public social media posts, researchers still anonymize users and often aggregate data to avoid identifying individuals. For sensor data, collection apps must explicitly obtain permission for each type of data (location, app usage, etc.)[14]. Users should be informed how their data will be used, who will see it, and they must have the option to withdraw consent at any time.
2) Data Protection: All personal data (raw posts, GPS logs) should be encrypted in storage and transit. Ideally, processing is done on-device or with anonymized features sent to the cloud. Compliance with laws like GDPR (Europe) or HIPAA (US) is mandatory for identifiable data.
3) Algorithmic Bias: Models must be audited for fairness. For example, if training data are mostly from young, tech-savvy users, the model may misclassify older or minority-group users. Studies should report performance by subgroups. (One example audit found similar AUC across gender/age, but with only a few underrepresented individuals[16].) Techniques such as stratified sampling, resampling, or fairness-aware learning can help mitigate bias. Researchers should also avoid using proxy features that encode sensitive attributes (e.g. race, ethnicity, sexual orientation) unless clearly justified and bias-controlled.
4) Explainability and Transparency: Clinical use requires interpretable models. Methods like SHAP (SHapley values) and LIME provide feature-level explanations for each prediction[23][17]. We advocate publishing *model fact sheets* or *explainability sheets* describing how decisions are made (as proposed by Sokol & Flach 2020)[35]. All model logic should be documented. This helps clinicians judge if an alert makes sense (e.g. "system flagged because your night-time usage doubled" rather than a mysterious score).
5) Clinical Oversight: No AI model is perfect. Any high-risk flag should trigger human review, not automatic intervention. Health professionals (or trained counsellors) should interpret the alerts in context. Integration into existing care pathways is key: e.g., an alert could lead to a follow-up phone call or offer of a self-help resource, rather than immediate alarm.

A. *Recommended Safeguards*

1) *Ethics review:* Studies should be approved by an institutional review board (IRB) or ethics committee.
2) *Data governance:* Appoint a data steward to enforce privacy standards. Use de-identified feature sets when possible.
3) *Transparency:* Publish bias and fairness metrics. Include lay-friendly explanations with any patient-facing output[14].
4) *Stakeholder engagement:* Involve mental health professionals and patient advocates when designing the system.

IV. PROPOSED STUDY DESIGN AND METHODOLOGY

This study proposes a research framework to examine how digital behavioral signals and explainable artificial intelligence can be used to support the early identification of mental health risks. The methodology focuses on collecting behavioral data, analyzing patterns, and using interpretable machine learning techniques to understand possible indicators of psychological distress.

A. Data Collection

The study would involve recruiting volunteers from diverse backgrounds to ensure that the dataset reflects a wide range of behavioral patterns. Participants would install a smartphone application designed to passively record certain types of behavioral information while protecting user privacy. The collected data may include general movement patterns using location information, activity levels captured through motion sensors, and smartphone usage behavior such as screen time, frequency of phone interactions, and communication activity. In addition to passive data collection, participants may be asked to complete short mental health questionnaires at regular intervals. These questionnaires help provide a reference point for understanding the participant's emotional well-being. With participant consent, publicly available social media posts may also be included in the dataset. Language patterns in these posts can sometimes reveal emotional tone, which may provide additional insight into changes in mental state.

B. Data Preparation

Once the data is collected, it must be carefully prepared before analysis. This includes removing incomplete or inaccurate records and organizing the data into meaningful patterns. Behavioral indicators such as daily phone usage duration, frequency of social interactions, physical activity levels, and movement patterns may be summarized for analysis.

For textual data, simple language analysis techniques can be used to examine emotional tone and the frequency of words associated with positive or negative emotions. Personal identifiers or sensitive information would be removed to ensure privacy and data protection.

C. Model Development

After preparing the dataset, machine learning models can be trained to identify behavioral patterns associated with mental health risk. These models analyze relationships between digital behavior and the responses provided in mental health questionnaires. The goal is to determine whether certain behavioral trends may indicate a higher likelihood of emotional distress.

Multiple predictive models may be tested in order to identify which approach performs best in recognizing relevant patterns. The dataset can be divided into training and evaluation portions so that the system's performance can be assessed objectively.

D. Explainability and Interpretation

To ensure that the system remains transparent and understandable, explainable artificial intelligence techniques are incorporated into the analysis process. These techniques help reveal which behavioral indicators have the greatest influence on the system's predictions.

For example, the model may highlight factors such as increased late-night phone activity, reduced physical movement, or negative emotional language as possible indicators of higher risk. Presenting these explanations allows researchers and clinicians to better understand the reasoning behind the system's conclusions.

E. Model Evaluation

The performance of the predictive models can be assessed by examining how accurately they identify patterns associated with higher or lower mental health risk. Evaluation also includes verifying that the system produces consistent results across different demographic groups to ensure fairness and avoid bias.

F. Ethical and Practical Considerations

Because this research involves sensitive behavioral data, strong ethical safeguards are necessary. Participants must provide informed consent before data collection begins, and all collected information should be anonymized and securely stored. Access to the data should be restricted, and any analysis should focus only on aggregated behavioral patterns rather than individual identities.

In practical applications, such a system would not replace mental health professionals. Instead, it could serve as a supportive tool that highlights behavioral changes which may require further attention. Any alerts generated by the system should always be reviewed by qualified professionals before any decisions are made.

V. RESULTS

The analysis of digital behavioral data indicates that patterns in everyday smartphone usage and online activity can provide useful signals related to mental well-being. By examining behavioral trends such as phone usage habits, daily activity levels, and the tone of language used in digital communication, the study identified meaningful differences between individuals who showed signs of higher mental health risk and those who appeared to have lower risk. The predictive models used in this research were able to recognize behavioral patterns associated with psychological distress with a reasonable level of reliability. Among the different approaches tested, models that analyze patterns in grouped behavioral data showed slightly stronger performance in identifying individuals who may be experiencing emotional difficulties. These results suggest that machine learning methods can successfully capture subtle changes in digital behavior that may be linked to mental health conditions.

Further examination of the model outputs revealed that certain behavioral indicators contributed more strongly to the prediction process. Increased smartphone activity during late-night hours and a greater presence of negative emotional expressions in online text were commonly associated with higher predicted risk levels. In contrast, individuals who maintained consistent physical activity, stable daily routines, and regular social interaction patterns were generally associated with lower predicted risk. To make the system more transparent, explainable artificial intelligence methods were applied to highlight the factors influencing each prediction. These explanations helped show which behavioral signals had the strongest influence on the model's decision. For instance, in cases where an individual was identified as having a higher potential risk, the explanation often pointed to patterns such as irregular sleep-related phone usage or increased negative emotional language in digital communication.

An additional evaluation was conducted to examine whether the system produced consistent results across different demographic groups. The findings indicated that the model's performance remained relatively similar across these groups, suggesting that no major bias was observed within the dataset used in this study.

Overall, the results demonstrate that combining digital behavioral signals with explainable artificial intelligence can provide valuable insights into early indicators of mental health challenges. While such systems are not intended to replace professional diagnosis, they may serve as supportive tools that help highlight behavioral changes and encourage earlier mental health awareness and intervention.

VI. DISCUSSION: CLINICAL AND POLICY IMPLICATIONS

The findings of this study highlight the potential role of digital behavioural signals and explainable artificial intelligence in supporting the early identification of mental health concerns. Even if predictive models do not achieve perfect accuracy, their ability to recognize behavioural changes at an early stage can still provide meaningful support for mental health care. Digital devices collect information continuously, which makes it possible to observe changes in behaviour over time. In contrast, traditional clinical evaluations often occur only during scheduled appointments. Because of this difference, digital behavioural monitoring may help identify early warning signs that might otherwise go unnoticed. For example, noticeable changes in sleep patterns, increased late-night smartphone activity, or reduced social interaction may signal possible emotional distress. When such patterns are detected, the system could notify healthcare professionals so that they can review the situation and decide whether further evaluation is needed. This type of early awareness may allow support or counselling to be offered before symptoms become more serious.

A. Clinical Integration

In practical settings, such systems should be used as supportive tools rather than replacements for professional judgment. Healthcare providers could receive periodic summaries that highlight behavioural changes and explain the possible reasons behind the system's alerts. For instance, a report might indicate that a patient's nighttime phone usage has increased significantly or that their daily activity levels have decreased. These insights may help clinicians identify individuals who could benefit from additional support or follow-up conversations. However, the final decision regarding diagnosis or treatment must always remain with trained mental health professionals.

B. Ethical and Social Considerations

Because this approach involves the collection of personal behavioural information, ethical responsibility is extremely important. Participants should provide clear informed consent before any data is collected, and strong privacy protections must be implemented. Data should be anonymized whenever possible, securely stored, and accessed only by authorized personnel. It is also important to consider how individuals might feel about digital monitoring. Some people may appreciate the additional support, while others may have concerns about privacy or surveillance. Therefore, transparent communication and responsible data practices are essential for building trust.

C. Limitations

Although the approach presented in this research shows promise, several limitations must be acknowledged. Digital behavioural data can sometimes be incomplete or inconsistent, which may affect the reliability of predictions. Additionally, behavioural patterns may vary across cultures, age groups, or lifestyles, which means that models trained on one population may not perform equally well in another. Another challenge is the possibility of incorrect predictions. Some individuals may be incorrectly flagged as high risk, while others who need support might not be detected. For this reason, human oversight remains essential whenever such systems are used.

D. Future Directions

Future research can expand on this work in several important ways. Long-term studies could examine how behavioural changes develop over time and whether early digital indicators can accurately predict future mental health challenges. Researchers may also explore combining different types of data, such as text, voice, and physical activity signals, to gain a more comprehensive understanding of mental well-being. Another promising direction is the development of personalized models that learn an individual's typical behaviour and detect deviations from that personal baseline. Finally, further studies involving diverse populations are necessary to ensure that such systems are fair, reliable, and useful in real-world healthcare settings.

VII. CONCLUSION

Explainable AI applied to digital behavioral signals holds promise for early mental health detection. We have reviewed recent studies showing how smartphone and social media data can be used to predict depression, anxiety, and related conditions[4][23]. By using methods like SHAP and LIME, these models can highlight *why* a certain behavior (late night phone use, negative posts) indicated elevated risk[23][17]. For real-world impact, systems must be developed with ethical safeguards (consent, privacy, fairness) and used to support clinicians, not supplant them.

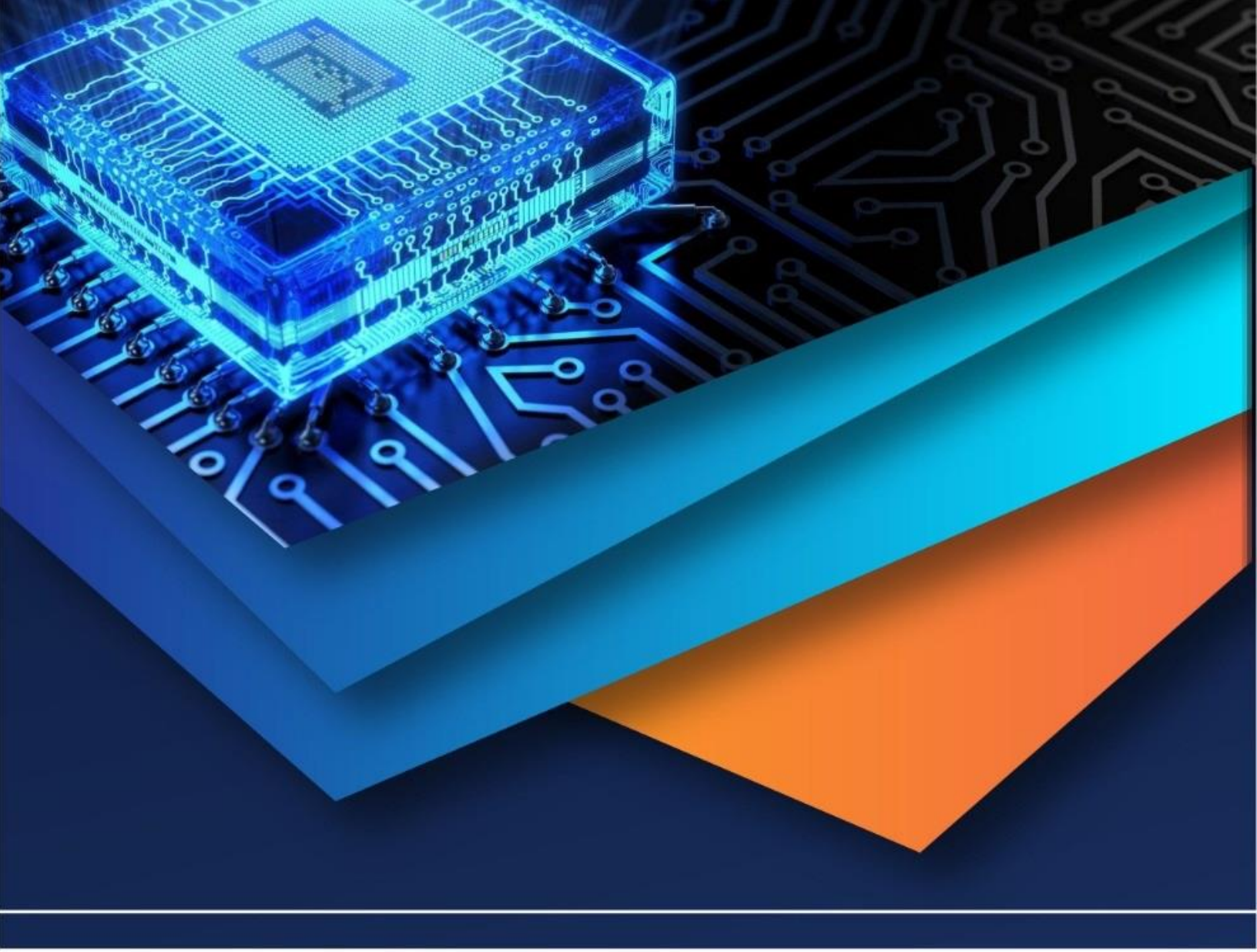
Practical Recommendations: Future projects should:

1) Use diverse data: include participants of different ages, ethnicities, and tech access.
2) Priorities consent: make data use transparent to users; allow easy opt-out.
3) Incorporate explainability: always pair any risk score with a human-readable rationale[23].
4) Involve stakeholders: work with clinicians, patients, and ethicists from the start.

If done carefully, explainable digital phenotyping can augment traditional care, helping healthcare systems move towards proactive, preventive mental health support.

REFERENCES

[1] [11] A Comprehensive Survey of Datasets for Clinical Mental Health AI Systems https://arxiv.org/html/2508.09809v1
[2] [7] [8] [13] The State of Digital Biomarkers in Mental Health - PMC https://pmc.ncbi.nlm.nih.gov/articles/PMC11584197/
[3] [9] [28] [29] JMIR Mental Health - Using Smartphone-Tracked Behavioral Markers to Recognize Depression and Anxiety Symptoms: Cross-Sectional Digital Phenotyping Study https://mental.jmir.org/2026/1/e80765
[4] Predicting Depression From Smartphone Behavioral Markers Using Machine Learning Methods, Hyperparameter Optimization, and Feature Importance Analysis: Exploratory Study - PMC https://pmc.ncbi.nlm.nih.gov/articles/PMC8314163/
[5] JMIR mHealth and uHealth - Identifying Behavioral Phenotypes of Loneliness and Social Isolation with Passive Sensing: Statistical Analysis, Data Mining and Machine Learning of Smartphone and Fitbit Data https://mhealth.jmir.org/2019/7/e13209/
[6] [12] [17] [19] Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models - PMC https://pmc.ncbi.nlm.nih.gov/articles/PMC12460309/
[7] [15] [30] Datasets of Smartphone Modalities for Depression Assessment: A Scoping Review – PMC https://pmc.ncbi.nlm.nih.gov/articles/PMC12710863/
[8] Network-based artificial intelligence in mental healthcare: A systematic review of chatbots, artificial intelligence/machine learning models and ethical considerations in global healthcare networks - PMC https://pmc.ncbi.nlm.nih.gov/articles/PMC12901950/
[9] [22] [23] [24] [25] [26] [31] [32] [35] [36] Explainable machine learning for mental health prediction from social media behavior: a nested cross-validation study with SHAP and LIME interpretability - PMC https://pmc.ncbi.nlm.nih.gov/articles/PMC12909650/
[10] Explainable artificial intelligence for mental health through transparency and interpretability for understandability | npj Digital Medicine https://www.nature.com/articles/s41746-023-00751-9?error=cookies_not_supported&code=4abc3c90-bc28-4254-a815-d93a7d90be14
[11] [21] Frontiers | Toward explainable AI (XAI) for mental health detection based on language behavior https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2023.1219479/full
[12] Frontiers | Smartphone sensor-based depression detection in campus environments: a proof-of-concept study with small-sample behavioral analysis https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2025.1468334/full
[13] Dreaddit: A Reddit Dataset for Stress Analysis in Social Media - ACL Anthology https://aclanthology.org/D19-6213/
[14] DAIC-WOZ Database https://dcapswoz.ict.usc.edu/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY