BRI-MH: Behavioral Risk Index for Mental Health — An Interpretable Multimodal LLM-Augmented Framework (Student Abstract)

Mahi Mann[1], **Avinash Anand**[1], **Rajiv Ratn Shah**[1]

[1]Indraprastha Institute of Information Technology Delhi (IIIT-Delhi)
mahi22272@iiitd.ac.in, avinasha@iiitd.ac.in, rajivratn@iiitd.ac.in

Abstract

Mental health monitoring faces challenges from fragmented data and opaque risk scores. We present **BRI-MH**, an interpretable multimodal framework combining behavioral signals with cognitive features from large language models to produce a weekly Behavioral Risk Index. Unlike prior work with isolated or black-box scores, BRI-MH offers transparent, actionable insights and links continuous monitoring to adaptive feedback and therapeutic support, bridging digital phenotyping and clinical care (Jain et al. 2015).

Introduction and Related Work

Mental health disorders affect over a billion people worldwide (Roberts, Kim, and Anderson 2025), yet monitoring remains reactive and fragmented, relying on infrequent clinical visits and unreliable self-reports that often miss crucial behavioral changes (Thompson, Martinez, and Johnson 2024). Digital phenotyping uses continuous data from smartphones and wearables to capture behavioral signals such as mobility, sleep, and device usage (Zhang et al. 2024); however, current approaches typically focus on single modalities, produce opaque risk scores, and seldom integrate with actionable feedback or clinical workflows (Heckler and Smith 2025).

Large language models (LLMs) have been shown to extract rich cognitive markers including sentiment, empathy, and suicidality indicators from text data (Chen et al. 2024), but prior work often treats these features in isolation rather than combining them with behavioral data. Multimodal fusion techniques improve healthcare predictions by integrating diverse data modalities (Liu et al. 2024), but their application in comprehensive mental health monitoring remains limited. Importantly, most existing tools stop at risk scoring without closing the loop to provide personalized feedback or facilitate clinician action (Thompson, Martinez, and Johnson 2024).

To address these gaps, we propose **BRI-MH** (Behavioral Risk Index for Mental Health), an interpretable multimodal framework that integrates passive behavioral data, active mood self-reports, and LLM-derived cognitive features into a weekly composite risk score with modality-level explainability. Unlike prior methods, BRI-MH links continuous risk assessment with adaptive feedback and clinician dashboards, enabling actionable interventions and bridging research and clinical practice.

Methodology

Figure 1 shows the BRI-MH architecture comprising Data Streams, Feature Engineering, Risk Scoring, and Monitoring & Feedback.

Data Acquisition

BRI-MH leverages design principles and data schemas inspired by well-known multimodal mental-health datasets, including:

- **StudentLife** (Wang et al. 2014): passive smartphone sensing (mobility, sleep, app use) with PHQ-9 scores.
- **StudentSADD** (Tlachac et al. 2021b): text, scripted and unscripted voice recordings, and PHQ-9 labels.
- **EMU** (Tlachac et al. 2021a) **/ DepreST-CAT** (Tlachac et al. 2022): mobile, social media, and audio data paired with PHQ-9 and GAD-7 measures.
- **Weibo Depression** (Shen et al. 2017): social media text, engagement patterns, and depression labels.

These datasets collectively motivate the three modalities used in BRI-MH: (1) Behavioral signals (mobility, activity, call patterns, device use); (2) Self-reports (EMA mood logs (Shiffman, Stone, and Hufford 2008), PHQ-9 (Kroenke, Spitzer, and Williams 2001) and GAD-7 surveys (Spitzer et al. 2006)); and (3) Cognitive/LLM-derived features (sentiment, empathy, stressors, suicidality cues).

For each participant, weekly data from all three modalities are combined into a unified feature representation. All feature groups are standardized using z-normalization to reduce scale bias.

Feature Extraction

Behavioral features capture mobility regularity, sleep consistency, screen-use variability, and other indicators of routine stability. Self-report features include mood levels, mood volatility, and survey-derived mental-health indicators. Cognitive features are extracted from text or transcripts using a

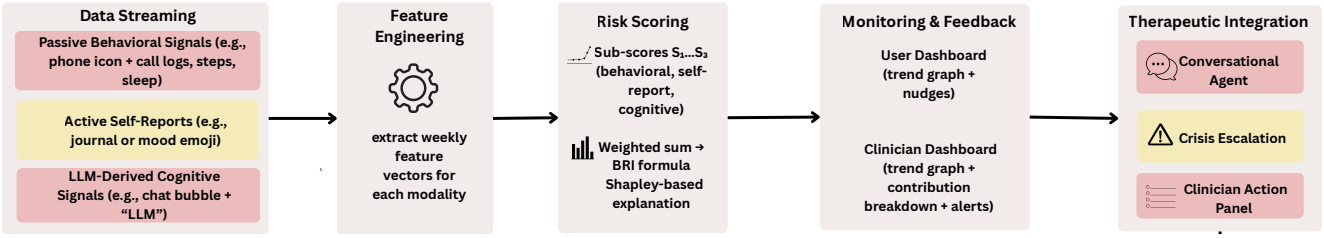

Figure 1: BRI-MH system architecture illustrating data ingestion, feature extraction, risk scoring with explainability, monitoring dashboards, and therapeutic integration.

mental-health–tuned LLM, producing sentiment, empathy, and emotional valence markers.

Each modality produces a normalized weekly risk component. The modular structure allows BRI-MH to incorporate additional modalities (e.g., voice or social media) while preserving interpretability.

Risk Modeling and Calibration

The final Behavioral Risk Index combines each modality's normalized risk component (derived from behavioral features, survey-derived severity, and LLM-extracted cognitive indicators) using a weighted fusion approach. Initial weights are estimated via logistic regression trained on historical PHQ-9 severity labels and then refined through clinician-informed calibration to align with real-world clinical reasoning and ensure interpretability.

Explainability is enabled through Shapley-based methods (Lundberg and Lee 2017), allowing clinicians to inspect which modalities or features contribute most to an elevated weekly risk score.

Adaptive Feedback and Online Learning

BRI-MH incorporates an adaptive feedback loop in which model updates are triggered by user engagement patterns and clinician feedback. The framework employs an online–federated learning approach (McMahan et al. 2017; Rieke et al. 2020) where local updates are computed on the user's device, securely aggregated across participants, and deployed only after stability and consistency checks. This design minimizes data exposure while enabling gradual personalization of individual risk thresholds and feature weights.

Monitoring and Validation

Users receive weekly summaries showing BRI trajectories, mood highlights, and personalized recommendations derived from detected risk drivers.Figure 2 shows an example of clinician dashboards, where clinicians can access synchronized views displaying weekly BRI scores and trends, along with modality-level attributions via SHAP (Lundberg and Lee 2017), and detected cognitive stressors and deviations from individual baselines. When a user's BRI surpasses personalized thresholds, the system triggers supportive messages, self-care recommendations, or clinician notifications, minimizing patient and clinician burden through automated triggers and integrated dashboards.

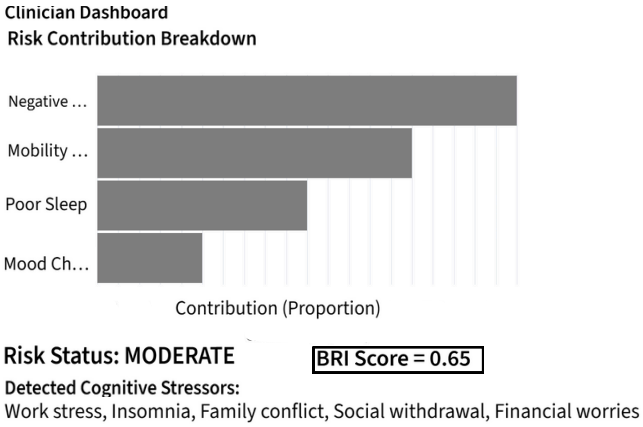

Figure 2: Prototype clinician dashboards illustrating the Behavioral Risk Index (BRI), modality contributions, and detected cognitive stressors, demonstrating integration of BRI-MH into clinical workflows.

Preliminary validation. As BRI-MH is presented as a conceptual framework, we conduct an initial feasibility assessment by examining cross-dataset consistency in feature definitions and modality relevance across StudentLife (Wang et al. 2014), StudentSADD (Tlachac et al. 2021b), EMU (Tlachac et al. 2021a), and Weibo Depression (Shen et al. 2017). These datasets demonstrate strong support for the three core modalities and validate the architectural compatibility of the framework.

Conclusion and Future Work

BRI-MH presents an interpretable multimodal risk index combining passive behavioral data, active self-reports, and LLM-derived cognitive features into a weekly composite score. The framework offers explainable insights and integrates therapeutic feedback, aiming to bridge AI risk assessment and clinical application.

Future work will focus on real-world pilot studies with consented college and clinical populations to validate the framework's effectiveness and refine personalization parameters. We will also investigate adaptive weight learning mechanisms to enable individual-level customization while maintaining safety and consistency across deployments.

References

Chen, R.; Wang, X.; Zhang, L.; and Liu, M. 2024. MentalLLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Nature Machine Intelligence*, 6(3): 287–301.

Heckler, T.; and Smith, R. 2025. Challenges and Opportunities in Digital Phenotyping for Mental Health. *Nature Digital Health*, 1(2): 100–110.

Jain, S. H.; Powers, B. W.; Hawkins, J. B.; and Brownstein, J. S. 2015. The Digital Phenotype. *Nature Biotechnology*, 33(5): 462–463.

Kroenke, K.; Spitzer, R. L.; and Williams, J. B. W. 2001. The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine*, 16(9): 606–613.

Liu, Y.; Wang, H.; Chen, Q.; and Zhang, B. 2024. Latent Space Data Fusion Outperforms Early Fusion in Multimodal Mental Health Assessment. *IEEE Transactions on Biomedical Engineering*, 71(8): 2245–2254.

Lundberg, S. M.; and Lee, S. I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 4765–4774.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and Arcas, B. A. Y. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 1273–1282.

Rieke, N.; Hancox, J.; Li, W.; Milletarì, F.; Roth, H. R.; Albarqouni, S.; Bakas, S.; Galtier, M. N.; Landman, B. A.; Maier-Hein, K.; Ourselin, S.; Sheller, M.; Summers, R. M.; Trask, A.; Xu, D.; Baust, M.; and Cardoso, M. J. 2020. The Future of Digital Health with Federated Learning. *npj Digital Medicine*, 3(1): 119.

Roberts, S.; Kim, J.; and Anderson, M. 2025. The Evolving Field of Digital Mental Health: Current Evidence and Future Directions. *Digital Medicine*, 8(2): 156–172.

Shen, G.; Jia, J.; Nie, L.; Feng, F.; Zhang, C.; Hu, T.; Chua, T. S.; and Zhu, W. 2017. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 3838–3844.

Shiffman, S.; Stone, A. A.; and Hufford, M. R. 2008. Ecological Momentary Assessment. *Annual Review of Clinical Psychology*, 4: 1–32.

Spitzer, R. L.; Kroenke, K.; Williams, J. B. W.; and Löwe, B. 2006. A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, 166(10): 1092–1097.

Thompson, K. L.; Martinez, A.; and Johnson, D. 2024. AI-Based Personalized Real-Time Risk Prediction for Behavioral Health Interventions. *Journal of Behavioral Health Services Research*, 51(2): 245–262.

Tlachac, M. L.; Flores, R.; Reisch, M.; Houskeeper, K.; and Rundensteiner, E. A. 2022. DepreST-CAT: Retrospective Smartphone Call and Text Logs Collected During the COVID-19 Pandemic to Screen for Mental Illnesses. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2): 1–32.

Tlachac, M. L.; Flores, R.; Toto, E.; and Rundensteiner, E. 2021a. EMU: Early Mental Health Uncovering Framework and Dataset. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1285–1292. IEEE.

Tlachac, M. L.; Sargent, E.; Toto, E.; Paffenroth, R.; and Rundensteiner, E. 2021b. StudentSADD: Rapid Mobile Depression and Suicidal Ideation Screening of College Students. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, 4536–4545. ACM.

Wang, R.; Chen, F.; Chen, Z.; Li, T.; Harari, G.; Tignor, S.; Zhou, X.; Ben-Zeev, D.; and Campbell, A. T. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 3–14. ACM.

Zhang, L.; Thompson, K.; Rodriguez, C.; and Kim, S. 2024. Multimodal Digital Biomarkers for Remote Monitoring of Mental Health Disorders. *npj Digital Medicine*, 7: 45.