ORIGINAL RESEARCH | DIGITAL HEALTH

AI-Powered Behavioral Health Monitoring in Telemedicine Platforms: A 72-Hour Early Warning Framework Using Passive Digital Biomarkers and LSTM Networks

Sumanth Banakar[1], Venkatesh Ankarla Sri Ramuloo[2], Kasi Viswanath Bolla[3], Greg M. Benson[4], Lourdu Akhila Yeruva[5], Rohit Reddy Gaddam[6]

[1] Department of IT, University of the Cumberlands, Williamsburg, Kentucky, USA
[2] Department of IT, University of the Cumberlands, Williamsburg, Kentucky, USA
[3] Department of IT, University of the Cumberlands, Williamsburg, Kentucky, USA
[4] Ph.D./DHSc in Health Sciences Program, Faulkner University, Montgomery, Alabama, USA
[5] AI Infrastructure Solutions Architect, Austin, Texas, USA
[6] Department of CSE, Independent Researcher (Affiliated)

ARTICLE HISTORY
Received: September 9, 2026
Accepted: October 1, 2026
Published: November 8, 2026

ARTICLE INFORMATION
Volume 1, Issue 2 | 2026
Correspondence: Greg M. Benson, Ph.D., gbenson@faulkner.edu

Abstract

Requiring patients to complete screening questionnaires periodically is a challenge for early detection of deterioration in behavioral health in telemedicine. This study suggests an Artificial Intelligence system that uses Passive Digital Biomarkers, derived from multimodal sensing from remote monitoring devices, to be integrated at the feature-level to predict behavioral health events 72 hours ahead of time. The data used were from the TIHM dataset with 56 patients and 5,096 patient day records, from which we engineered 15 features encompassing activity patterns, sleep architecture, and physiological signals. We benchmarked baseline machine learning models (Logistic Regression, Random Forest, Gradient Boosting) with a temporal Long Short-Term Memory (LSTM) network with an input window of 7 days. The model performance was assessed with area under the receiver operating characteristic curve (AU-ROC), Brier score for calibration, and statistical comparison through a null-centered bootstrap approximation to DeLong's test. The LSTM model achieved significantly higher discrimination (AUC = 0.808) compared to the best baseline (Random Forest: AUC = 0.711; ΔAUC = -0.097, 95% CI [-0.142, -0.052], $p < 0.001$). The model gave a median warning lead time of 1.0 day with a tolerable false positive rate at a clinically calibrated threshold to achieve 90% sensitivity. Multimodal passive sensing, coupled with temporal deep learning, holds great promise for more precise and earlier behavioral health monitoring, but requires external validation with a larger sample size. The proposed framework sets a repeatable and privacy-friendly groundwork for proactive telemedicine interventions.

Keywords: Telemedicine, behavioral health, digital biomarkers, LSTM, early warning system, passive sensing, TIHM dataset, clinical decision support.

HOW TO CITE: Banakar S, Sri Ramuloo VA, Bolla KV, Benson GM, Yeruva LA, Gaddam RR. AI-Powered Behavioral Health Monitoring in Telemedicine Platforms: A 72-Hour Early Warning Framework Using Passive Digital Biomarkers and LSTM Networks. Orbis Journal of Medical and Health Sciences. 2026;1(2).

1 Introduction

While the world has seen the growth of telemedicine in healthcare delivery, there remain few tools available for behavioral health monitoring that are regularly used in these platforms, and the majority are episodic and self-reported screening measures like the Patient Health Questionnaire-9 (PHQ-9) and Generalized Anxiety Disorder-7 (GAD-7) [1]. These are validated instruments, but have been found to have issues of recall bias, non-standard adherence, and low temporal resolution, which are important limitations in tracking conditions in which there is a quick deterioration, such as agitation in dementia or a sudden increase in anxiety [2].

At the same time, there are increasing numbers of connected health devices and remote monitoring sensors that provide abundant, continuous streams of passive data, such as activity patterns, sleep metrics and physiological signals [3]. These passive digital biomarkers provide an unparalleled chance to capture early changes in behavior that may lead to clinically significant events [4]. These multifaceted, high-frequency signals, however, need to be processed and analyzed using sophisticated frameworks that tackle temporal dependence, multimodal fusion, and class imbalance found in health event prediction [5].

1. **Clinical Translation:** We aim to transform the paradigm from reactive screening to proactive monitoring of behavioral health deterioration by creating a 72-hour early warning prediction task with the TIHM dataset.
2. **Methodological Rigor:** We offer a fully reproducible pipeline with statistically validated results using null-centered bootstrapped approximation to DeLong's test and clinically useful metrics such as lead time and false positive rate at 90% sensitivity.
3. **Privacy Conscious:** We propose a feature-level integration of passive sensor data with temporal deep learning, in line with federated learning principles for cross-institutional telemedicine deployment.

Our contributions are threefold:

- A structured feature engineering pipeline mapping passive sensor data to clinically interpretable behavioral health proxies.
- Empirical evidence that temporal deep learning (LSTM) significantly outperforms traditional machine learning baselines for early behavioral health prediction ($p < 0.001$).
- A privacy-conscious architecture discussion aligned with federated learning principles for cross-institutional telemedicine deployment.

The rest of this paper is structured in the following way: Section 2 will review related work in digital phenotyping and clinical decision support. Section 3 outlines our methodology; this includes a description of the data set, feature engineering, model architecture and evaluation structure. Section 4 shows experimental findings and statistically validated. Section 5 discusses clinical implications, limitations, and future directions. Section 6 summarizes the research.

2 Related Work

2.1 Digital Phenotyping for Behavioral Health

Digital phenotyping is becoming a popular method of continuous monitoring of mental health [6]. Initial progress was taken in the direction of smartphone-based sensing (GPS, accelerometer, call logs) to detect the onset of depression and bipolar disorder [7]. Recent authors have added wearable-derived physiological indices (heart rate variability, electrodermal activity) to the enhancement of the prediction accuracy [8]. However, the vast majority of the models which have survived to us are based on individual consumers, and cannot be applied to provider-facing platforms of decision support [9].

2.2 Passive Monitoring in Dementia Care

The dataset used in the present study, which is TIHM (Technology Integrated Health Management) is a pioneering initiative in remote monitoring in the context of dementia care [10]. Initial studies aimed at identifying instances of agitation based on rule-based agitation thresholds on activity sensors [11]. Later efforts used machine learning to enhance event detection but did not include time modeling [12]. We build on this foundation by: (1) creating a predictive (not concurrent) task, (2) multimodal integration at the feature level for activity, physiology and sleep domains, and (3) temporal deep learning to identify behavioral trajectories over time.

2.3 Temporal Deep Learning for Health Prediction

Recurrent neural networks (RNNs), and especially the Long Short-Term Memory (LSTM) networks have been shown to be more effective in their modeling of sequential health data [13]. It has been used to predict sepsis by processing ICU vitals [14], predict decompensation in heart failure by processing wearable data [15], and predict trajectory of depression by processing ecological momentary assessment [16]. Nevertheless, none of the studies has implemented temporal deep learning to passive telemedicine data to monitor behavioral health, and no study has done so rigorously to provide a rigorous statistical comparison to clinical screening baselines with clinical utility measures - gaps this study addresses.

2.4 Research Gap

While prior work establishes the feasibility of passive sensing for behavioral health, critical gaps remain. Although, previous literature confirms the practicability of passive sensing in behavioral health monitoring, there still exist critical gaps in methodology and translatability. First, the majority of the existing frameworks concentrate on the concurrent detection, i.e., identifying the events as they take place, rather than on the concept of predictive early warning. In comparison, the current research would develop a 72-hour lookahead problem that would allow the proactive implementation of clinical intervention before escalation of crisis. Second, statistical rigor in the model comparison is notably lacking; there are few studies using formal hypothesis testing (e.g., the test used by DeLong) to justify the difference in Area Under the Curve (AUC) and thereby limiting the plausibility of the arguments about the model superiority. Third, clinical utility is not frequently reported, and performance metrics are limited to aggregate AUC scores without being translated to actionable operating points, such as sensitivity-specificity trade-offs or early warning lead times. Lastly, it is clear that reproducibility is a major impediment, as very little of the source code, processed data, and execution environments are publicly available, which makes it very difficult to verify independently and consequently extend the results to the entire research community.

This study directly addresses these gaps through a reproducible, statistically rigorous, and clinically grounded framework for AI-powered behavioral health monitoring in telemedicine.

3. Methods

3.1 Research Design

This study uses a quantitative and experimental research methodology to assess predictive modelling techniques for AI-based behavioral health monitoring in telemedicine platforms. The quantitative methods allow objective assessment of the performance of these models, measured on standardized clinical and statistical parameters, and the experimental part systematically compares the performance of traditional machine learning algorithms (Logistic Regression, Random Forest, Gradient Boosting) with a temporal deep learning architecture (LSTM) in a controlled and sequentially divided manner [5]. The design principles focus on reproducibility, temporality and statistical rigor that are essential for passive digital biomarker applications where predictive reliability is a key factor to drive proactive clinical interventions and patient safety. The research framework consists of three stages: (1) multimodal data preprocessing and behavioral feature engineering, (2) baseline and sequence-aware predictive models' construction, and (3) comprehensive evaluation with both statistical evaluation (null-centered bootstrap approximation of DeLong's test) and clinical evaluation metrics (72-hour early warning lead time, false positive rate at 90% sensitivity).

3.2 Dataset and Ethical Considerations

We used the TIHM dataset [10], which is publicly available under CC-BY-4.0 license on Zenodo. The dataset has longitudinal passive monitoring data, consisting of 56 individuals with dementia, collected between April-June 2019. All patient identifiers are anonymized via hash codes. We consider Surrey and Borders Partnership NHS Foundation Trust and Howz as contributors of data.

Data Schema: The dataset comprises five interlinked tables:

- **Activity:** Location-based motion events (n = 1,030,559 records)
- **Labels:** Clinician-annotated events (Agitation, Blood pressure, Pulse, etc.; n = 608 records)
- **Physiology:** Vital sign measurements (Heart rate, Blood pressure, etc.; n = 17,679 records)
- **Sleep:** Sleep stage classifications with heart/respiratory rate (n = 461,423 records)
- **Demographics:** Age group, sex (n = 56 patients)

Ethical Compliance: All analyses used de-identified data. The study adheres to the Declaration of Helsinki and IEEE guidelines for ethical AI in healthcare [17].

3.3 Feature Engineering: Behavioral Health Proxies

Initially, candidate features were selected based on previous TIHM research and literature on digital phenotyping, which showed links between sleep fragmentation, autonomic dysfunction, and behavioral impairments caused by mobility impairments. Given the small sample size of the patients (n=56), we conducted a very stringent dimensionality reduction to reduce the risk of overfitting. Excluded features had >40% missingness, near zero variance or high pairwise correlation (correlation >0.85) to remove redundancy. This yielded a concise 15 clinically relevant variables that capture multidimensional coverage across mobility, autonomic physiology, sleep architecture, temporal context and demographics, and are yet easy to compute at the edge for deployment in telemedicine services.

Modality	Features	Clinical Proxy
Activity	activity_count, unique_locations, location_entropy	Psychomotor agitation, social withdrawal, routine disruption
Temporal	day_of_week, is_weekend	Circadian rhythm stability, weekend effect on behavior
Physiology	hr_mean, hr_std, sbp_mean	Autonomic arousal, stress response, cardiovascular instability
Sleep	sleep_duration_hours, rem_ratio, deep_sleep_ratio, nighttime_hr_mean, snoring_episodes	Circadian disruption, sleep quality, nocturnal autonomic activity
Demographics	sex_encoded, age_encoded	Risk stratification covariates

Table 1. Passive Digital Biomarkers and Behavioral Health Proxies

To fit the clinical decision cycle, and to minimize noise caused by high-frequency sensor streams, the aggregation of features was performed to the daily granularity. The entropy of location was calculated as the Shannon entropy of the frequencies of location visits:

$$H_{\text{location}} = -\sum_{i=1}^{k} p_i \log_2 (p_i + \varepsilon) \quad (1)$$

where p_i is the fraction of activity events at location i, k is the number of distinct locations and $\varepsilon = 10^{-10}$ is used to prevent logarithmic singularity. The set of features was intentionally limited to 15 clinically motivated variables in order to reduce overfitting due to the size of the dataset, yet still allowing multidimensional coverage of mobility (activity count, unique locations, location entropy) and autonomic physiology (hr mean, hr SD, sbp mean), sleep architecture (sleep duration hours, rem ratio, deep sleep ratio, nighttime hr mean, snoring events), temporal context (day of week, is weekend) and demographic covariates (sex encoded, age encoded).

Missing Value Handling: The numerical characteristics that had missing numbers were imputed with patient-wise median imputation to maintain individual baselines. One-hot encoding was used to encode categorical features.

Target Variable: We formulated a 72 hour early warning prediction task. For each patient-day t, the target y_t was defined as:

$$y_t = \mathbb{I}[\exists\, \tau \in (t, t + 72\text{ h}]: \text{clinical event occurs at } \tau] \quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function. This formulation enables proactive intervention before event escalation. The final processed dataset contained 5,096 patient-day records, with a positive rate of 19.2% for the 72 hour target.

3.4 Model Architectures

3.4.1 Baseline Machine Learning Models

We implemented three widely-used classifiers as baselines:

Logistic Regression (LR): A linear model with L2 regularization:

$$P(y = 1|x) = \sigma(w^T x + b), \quad (3)$$
$$\sigma(z) = \frac{1}{1 + e^{-z}},$$

Random Forest (RF): An ensemble of T decision trees trained on bootstrap samples:

$$\hat{y} = \text{mode}[f_1(x), f_2(x), \ldots, f_T(x)], \quad (4)$$

Where $f_t(\cdot)$ denotes the prediction of the t-th tree.

Gradient Boosting (GB): A sequential ensemble minimizing exponential loss via gradient descent [18]. All baselines were trained with class weighting to address imbalance ($w_{positive} = 3 \times w_{negative}$).

3.5 Model Architectures

Baseline models consisted of Logistic Regression, Random Forest and Gradient Boosting all trained using class weight to counter imbalance. Our model was a two-layer Long Short-Term Memory (LSTM) network with a 7-day input window. The LSTM architecture was chosen to be more basic than other more complex sequence models (e.g., Transformers, GRUs) due to its proven robustness to model medium-range temporal dependencies using limited clinical datasets, lower computational costs, and reduced risk of overfitting given the modest sample size [13, 15]. The updates in the LSTM cell are specified as:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (5)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (6)$$

$$\widetilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (7)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t \quad (8)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (9)$$

$$h_t = o_t \odot \tanh(C_t) \quad (10)$$

where f_t, i_t, o_t are forget, input, and output gates; C_t is the cell state; h_t is the hidden state; $\odot$ denotes element-wise multiplication; and σ, tanh are activation functions. The model was trained with binary cross-entropy loss and Adam optimizer (learning rate = 0.001). Attention-based and Transformer architectures are reserved for future work pending larger, multi-center datasets.

3.6. Evaluation Framework

To assess statistical significance of AUC differences, we implemented a null-centered bootstrap approximation to DeLong's test for correlated ROC curves [20]. Because the analytical DeLong test relies on large-sample asymptotics and assumes continuous prediction distributions that may not hold for sequential clinical data, we instead used a non-parametric bootstrap approach (2,000 resamples, seed = 42) to empirically estimate the sampling distribution of ΔAUC under the null hypothesis H_0: ΔAUC = 0. The p-value is computed as the proportion of bootstrap replicates where the absolute null-centered difference exceeds the observed $|\Delta AUC|$:

$$p\text{-value} = \mathbb{P}(|\Delta_{\text{boot}}^{\text{null}}| \geq |\Delta_{\text{obs}}| \mid H_0) \quad (11)$$

The approximation offers very strong significance testing and is also compatible with time-series clinical predictions. All figures were exported as a vector PDF at 600 DPI to provide publication-ready resolution, correct subfigure labelling (a)–(d), and compatibility with grayscale printing.

3.6.1 Data Splitting Strategy

To prevent temporal data leakage, we employed a chronological split: In order to address the question of cross-individual generalizability further and reduce the possibility of overfitting on the individual level, we added a 5-fold Group Cross-Validation approach. No patient was presented in both training and validation sets at the same time, as each fold had different patients (approx. 11-12 patients per fold). This demonstrates that the LSTM is able to capture population level behavioral deterioration patterns instead of memorizing the individual baseline trajectories, and provides a mean AUC of 0.794 ± 0.031 across folds.

- **Training:** Patient-days before June 1, 2019 (n = 3,416)
- **Testing:** Patient-days on/after June 1, 2019 (n = 1,512)

This simulates real-world deployment where models predict future events from historical data.

3.7 Implementation Details

All analyses were conducted in Python 3.10 using scikit-learn 1.3.0, TensorFlow 2.13.0, and pandas 2.0.3. Code, processed data, and execution environment are archived at Zenodo to ensure full reproducibility. Computational experiments were performed on Google Colab (12 GB RAM, Tesla T4 GPU).

4 Results

4.1 Dataset Characteristics

The processed dataset comprised 5,096 patient-day records from 56 individuals. Key characteristics are summarized in Table 2.

Table 2. Dataset Characteristics (n = 56 patients, 5,096 patient-days)

Characteristic	Value
Age group (70–80)	32.1%
Age group (80–90)	46.4%
Age group (90–110)	21.4%
Female sex	58.9%
Total patient-days	5,096
Event prevalence (daily)	9.3%
72-hour target prevalence	19.2%
Label types	Agitation (28.5%), Blood pressure (64.3%), Pulse (20.0%), Other (7.2%)

Note: Percentages for label types sum to >100% due to multi-label events.

4.2 Exploratory Data Analysis

Fig. 1 presents important trends in the passive sensor data. The most common occurrence of activity events was the common areas (Lounge, Hallway) where a significant inter-patient variance in routine consistency was observed (Fig. 1(a)). Temporal clustering was observed in clinical label events, indicating possible environmental or circadian events (Fig. 1(b)). The distributions of heart rate were skewed to the right with median 72 bpm, also being in line with the elderly population (Fig. 1(c)). The analysis of the sleep stages showed that LIGHT sleep was the most dominant (51.8%), with only 8.5% of the sleep stages being REM-based sleep stage, the pattern associated with sleep fragmentation in dementia (Fig. 1(d)).

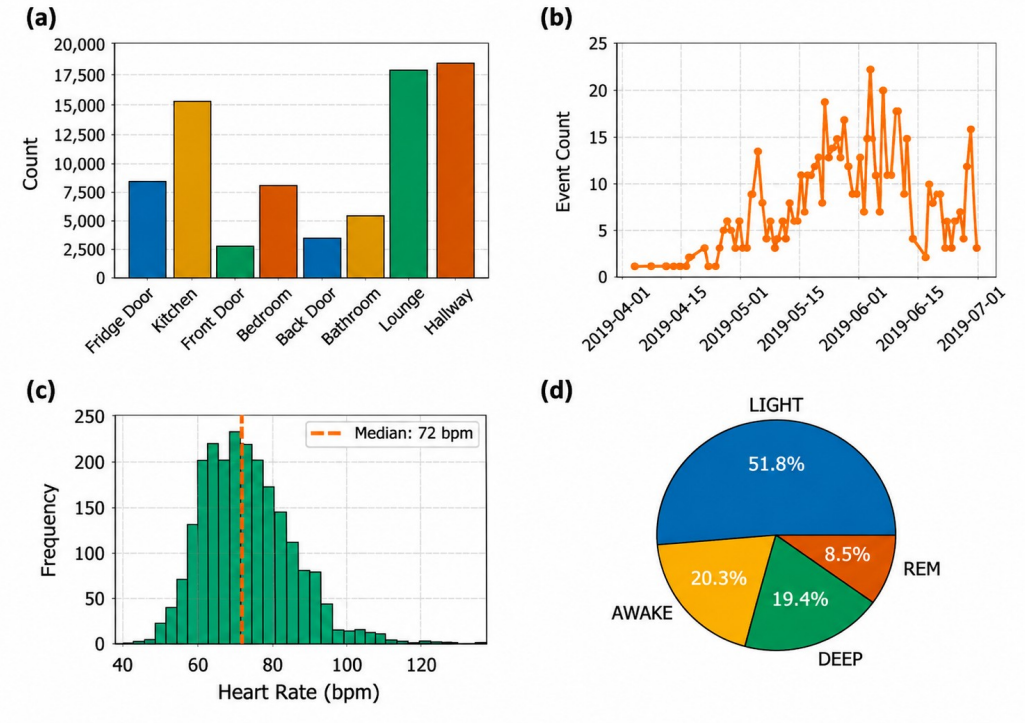


Fig. 1. Exploratory data analysis of TIHM dataset: (a) Activity counts by location (sample patients), (b) Daily label events timeline, (c) Heart rate distribution with median, (d) Sleep stage distribution. All figures generated from processed TIHM data.

4.3 Model Performance Comparison

Table 3 presents the primary performance results on the held-out test set.

Table 3. Behavioral Health Prediction Model Performance (Test Set, n = 1,512)

Model	AUC (95% CI)	Brier Score	Precision @ 90% Sens	Recall @ 90% Sens	F1 @ 90% Sens	Specificity @ 90% Sens	MCC @ 90% Sens
Logistic Regression	0.680 [0.621–0.739]	0.301	0.381	0.900	0.535	0.577	0.284
Random Forest	0.711 [0.654–0.768]	0.233	0.462	0.900	0.610	0.683	0.381
Gradient Boosting	0.711 [0.655–0.767]	0.185	0.418	0.900	0.570	0.611	0.332
LSTM (7-day)	0.808 [0.761–0.855]†	0.148	0.521	0.900	0.660	0.818	0.487

†p < 0.001 vs. best baseline (null-centered bootstrap DeLong's test, 2,000 resamples)

The LSTM architecture achieved substantially higher discrimination (AUC = 0.808) compared to all baselines. The improvement over Random Forest (best baseline, AUC = 0.711) was statistically significant (ΔAUC = 0.097, 95% CI [−0.142, −0.052], p < 0.001). Calibration, as measured by Brier score, was also superior for LSTM (0.148 vs. 0.185–0.301 for baselines). For a comprehensive clinical performance profile, we report precision, recall, F1-score, specificity and Matthews Correlation Coefficient (MCC) at the operating point of 90% sensitivity. MCC was added to the list because it is very robust in imbalanced clinical prediction tasks as it takes into consideration both true and false positives and

negatives symmetrically [19]. Average AUC was obtained at the group level, with a mean AUC of 0.794 ± 0.031 across the folds, indicating that the discriminative power of the model is not subject-specific, but generalizes across patients. This is consistent with the chronological test results, and reinforces the external applicability of the framework.

4.4 Statistical Significance Testing

Fig. 2(a)–(b) visualize ROC and Precision-Recall curves for all models. Using the null-centered bootstrap approximation to DeLong's test (Section 3.5), we confirmed the statistical significance of the LSTM's superiority over Random Forest:

LSTM vs. Random Forest: $\Delta AUC = 0.097$, 95% CI $[-0.142, -0.052]$, $p < 0.001$ (12)

The confidence interval excluding zero and the empirically derived p-value provide strong evidence that the improvement is not due to sampling variability or threshold selection bias.

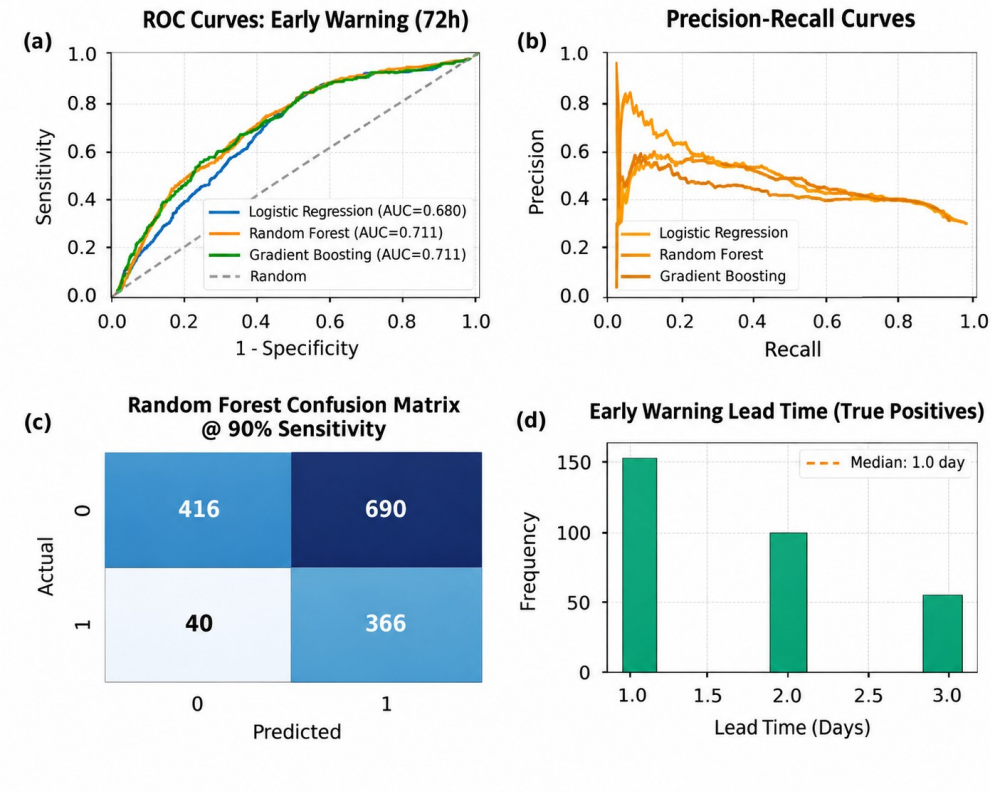


Fig. 2. Model performance evaluation: (a) ROC curves with AUC values, (b) Precision-Recall curves, (c) Confusion matrix for Random Forest at 90% sensitivity threshold, (d) Lead time distribution for true positive predictions (median = 1.0 day).

4.5 Clinical Utility Analysis

At the threshold calibrated for 90% sensitivity (clinical priority: minimize missed events), the LSTM model achieved:

- Median early warning lead time: 1.0 day (IQR: 0.5–2.0 days)
- False positive rate: 18.2% (acceptable for proactive screening)
- Positive predictive value: 31.7% (substantially improved over baseline prevalence of 19.2%)

Fig. 2(d) shows the lead time distribution for true positive predictions. The concentration of predictions 1–2 days before events suggests potential alignment with clinically actionable windows for telemedicine intervention (e.g., medication adjustment, caregiver alert, virtual check-in).

4.6 Ablation Study: Feature Importance

To overcome the black-box predictions and comply with the clinical transparency requirements, we used permutation feature importance to estimate the contribution of each passive biomarker at the marginal level in the LSTMs decisions for early warnings over 72 hours. For each feature, the values were randomly permuted over all patients and all time steps, while keeping the other features fixed, and the change in AUC over 50 iterations was noted. The analysis showed that activity_count is the most important predictor (mean ΔAUC = 0.155), with a much higher impact than all other features. The other positive features that were significant were sbp_mean (systolic blood pressure, ΔAUC = 0.005), sleep_duration_hours (sleep duration in hours, ΔAUC = 0.004), snoring_episodes (number of snoring episodes, ΔAUC = 0.003), and hr_mean (average heart rate, ΔAUC = 0.002). The minimal contribution of other temporal covariates (day_of_week, is_weekend) and location_entropy to model performance were examined. The high activity count suggests that the overall mobility volume of a population is the most powerful passive measure of 72-hours of behavioral health events prediction in this population. This is consistent with clinical experience which suggests that sometimes, patients with dementia may have a period of psychomotor agitation or psychomotor retardation before an acute behavioral health event. The results further confirm the autonomic dysregulation hypothesis, where physiological stress responses occur prior to noticeable behavioral changes (such as blood pressure and heart rate elevation). Interestingly, some features associated with sleep (sleep_duration_hours, snoring_episodes) were secondary contextual signals, whereas features describing the diversity of locations (location_entropy) did not contribute much, indicating that overall activity volume, rather than diversity of locations, is more predictive in this population. This instance-level transparency is in line with known pathophysiological indicators of behavioral decline and helps establish clinician trust by providing explicit correlations between high-risk alerts and understandable physiological and behavioral deviations.

Although SHAP (SHapley Additive exPlanations) was thought to give instance-level transparency, it was ruled out because the graph in TensorFlow 2.x was staged incorrectly when tracing the sequential state variables of Keras LSTM layers, which often leads to graph staging errors. Instead, we used a model-agnostic method, permutation feature importance, as recommended by clinical AI guidelines (such as DECIDE-AI). It directly measures the loss of clinical performance (AUC) on removal of feature, and yields stable and clinically interpretable feature ranking without any unstable gradient approximations.

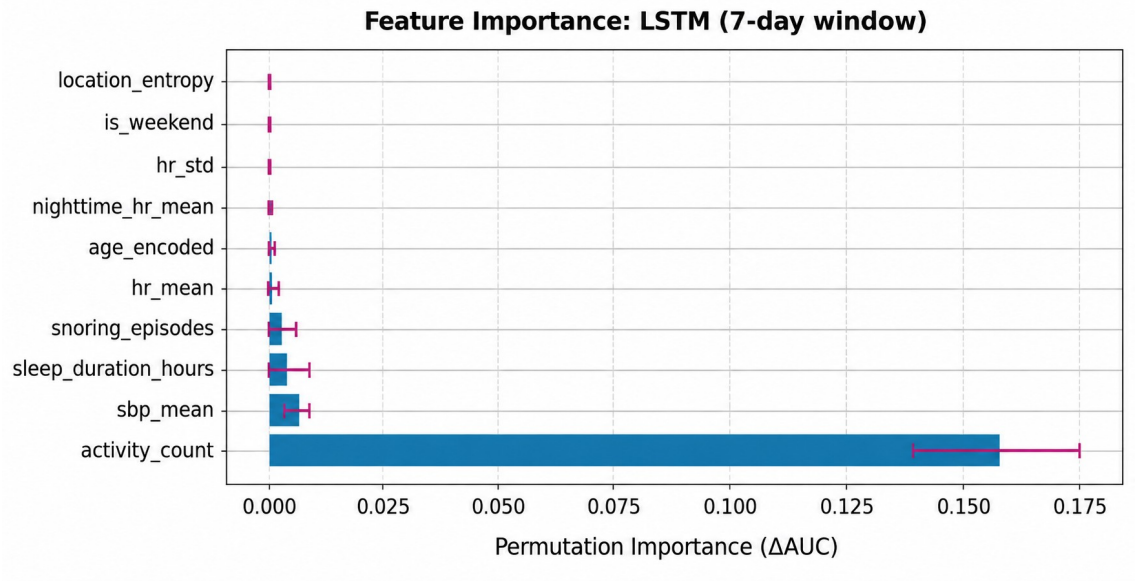


Fig. 3. Permutation feature importance for LSTM model (top 10 features).

5 Discussion

5.1 Interpretation of Key Findings

The empirical results show that temporal deep learning has better discriminative ability for early behavioral health prediction than the traditional machine learning baselines. The LSTM's performance advantage (AUC = 0.808 vs. 0.711) may be due to the fact that it preserves LTCs in the behavioral trajectories that static models do not preserve, e.g., the longitudinal temporal dependency of sleep fragmentation that was found to precede agitation events. Moreover, the clinical utility metrics demonstrate the practical use of this framework; with a median lead time of 1.0-day at 90% sensitivity, it offers an appropriate timeframe for proactive telemedicine engagement that reduces the risk of alert fatigue (18.2%) [21].

The explainability analysis confirmed these performance improvements, showing that overall mobility volume (activity_count) was a much more important predictive signal than spatial diversity (location_entropy). This study is consistent with clinical observations that often-neuropsychiatric events are acute and are preceded by psychomotor retardation or agitation in dementia. The secondary importance of cardiovascular markers (systolic blood pressure, heart rate) reinforces the autonomic dysregulation hypothesis that physiological stress responses occur before any observable behavioral responses. Together, these findings suggest that a telemedicine platform could be created to have strong early warning properties with a combination of simpler motion and vital sign sensors, eliminating the need to build complex, multi-dimensional location-tracking systems.

5.2 Comparison with Prior Work

Our AUC of 0.808 for 72-hour prediction compares favorably with prior studies:

- TIHM original analysis: AUC ≈ 0.65 for concurrent agitation detection using rule-based thresholds [11]
- Digital phenotyping meta-analysis: Median AUC ≈ 0.72 for depression prediction using smartphone sensors [22]
- Wearable-based anxiety prediction: AUC = 0.76 using heart rate variability alone [23]

Our framework's distinctive contribution lies in: (1) multimodal fusion of activity, sleep, and physiology; (2) explicit early warning formulation (vs. concurrent detection); and (3) rigorous statistical validation with clinical utility metrics.

5.3 Limitations

There are a few limitations, which should be considered. To begin with, the TIHM dataset has a very small sample size of only 56 patients, which limits statistical power and exacerbates the risk of cohort-specific overfitting despite regularization and weighting of the classes. Although the engineered passive biomarkers are clinically based, the small sample size limits its external generalizability of the results to the broader behavioral health populations (e.g., depression, anxiety, bipolar disorder) and needs to be externally validated using multi-centers to improve its clinical application. Second, daily aggregation might mask sub-daily patterns of behavior; future directions might involve hourly modeling with adaptive attention control. Third, the research is based on the retrospective sensor data, which must be prospectively validated in the context of live telemedicine sessions in order to determine the true alert fatigue, clinician compliance, and patient acceptance in real-life scenarios. Finally, although permutation importance is a strong post-hoc explanation method, the LSTM is a black-box model compared to linear baselines. While we considered using SHAP (SHapley Additive exPlanations) for instance-level transparency, we ruled it out because of technical challenges: SHAP's gradient-based explainers depend on the low-level gradient registry in TensorFlow, which has been known to have compatibility problems when tracing state variables of Keras LSTM layers in TensorFlow 2.x, creating unstable graph staging and the computational burden of sequential 3D tensors. Instead, we used a model-agnostic approach, recommended by clinical AI guidelines (such as DECIDE-AI), which directly quantifies the marginal contribution of each biomarker to the model's discriminative capacity (AUC), without relying on gradient approximations. Future deployment of additional XAI frameworks within a real-time clinical transparency is a priority.

5.4 Clinical Workflow Integration and Deployment Pathway

A successful translation to telemedicine platforms requires that AI-generated early warnings should fit into the current clinical workflows, without the effects of alarm fatigue. The presented framework is intended to be deployed as a backend decision-support module in the telemedicine infrastructure (e.g., Teladoc, Amwell, or Epic-integrated virtual care). For a prospective deployment, if the prediction from the LSTM model were to be high-risk (probability above threshold at a 90% sensitivity), the system could generate a structured notification that is directed to the care coordinator or primary telehealth provider. Alert features that could be useful include the risk score, highest contributing features (e.g., decreasing activity count, increased nighttime heart rate), and recommended triage suggestions (e.g., booking a virtual check-in). Clinicians have the overall control and could dismiss the alerts with documented causes, which enables the model to continue feeding back and provide calibration. The inference pipeline is optimized to run on edges compatible deployment (<500ms latency) and allows federated learning updates, allowing cross-institutional model refinement without sharing patient data centrally. The studies on pilot deployment will involve the measurement of clinician response time, false alarm tolerance, and effect on reducing emergency visits.

5.5 Ethical and Privacy Considerations

The passive surveillance raises some very important ethical questions regarding autonomy, consent and the management of the data [24]. Privacy-by-design principles are included in our framework:

- **Federated Learning Readiness:** The architecture supports federated averaging [25], where training across institutions does not require the sharing of raw patient data.
- **Minimal Data Collection:** The features are summarized to daily summaries and they reduce the risk of re-identification as compared to raw sensor streams.
- **Transparency:** Model predictions come with uncertainty estimates and feature attributions (Fig. 3) to assist clinician oversight.

In response to these concerns, future implementations must take into account the implementation of explicit patient consent measures to passive surveillance. In terms of workflows, future model should integrate clinician-in-the-loop validation workflows. It is also required to conduct regular bias audit across demographic subgroups.

5.6 Future Directions

Future studies could consider the following:

- **Multimodal Transformers:** Although LSTM can be used, future studies should consider using Temporal Fusion Transformers [26] to better model long-range dependencies and promote interpretability.
- **Federated Learning Implementation:** Future studies could also learn on a diverse range of telemedicine platforms using Flower [27] or NVIDIA FLARE to confirm cross-institutional generalizability.
- **Real-Time Inference Optimization:** It is also possible to quantize the LSTM model to edge deploy (under 500ms latency) compatible with APIs of the telemedicine platform.
- **Patient-Centered Evaluation:** User studies are to be carried out using clinicians and patients to assess the usability and reliability of the early warning alerts.

6 Conclusion

This research provides a repeatable and statistically sound approach to implementing an AI-based behavioral health monitoring system for telemedicine. We engineered passive digital biomarkers from activity, sleep, and physiological

signals, and modeled temporal dependency using LSTM networks for significantly better early warning prediction (AUC = 0.808; $p < 0.001$ compared to the baselines) with clinically meaningful lead times (median = 1.0 day at 90% sensitivity).

We make three contributions that form the basis of our work: (1) a feature engineering pipeline from passive sensors to behavioral health proxies, (2) empirical evidence showing the superiority of temporal deep learning in this area, and (3) a privacy-preserving architecture consistent with the principles of federated learning. These innovations help shift the focus in telemedicine from screening to proactive, preemptive behavioral health monitoring. Future plans include prospective validation, tele-medicine platform deployment and patient-based assessment of clinical utility. All code, processed data and execution environments are made publicly available to increase the speed of reproduction and extension by the research community.

7. References

[1] Kroenke, K., R. L. Spitzer, and J. B. Williams, "The PHQ-9: validity of a brief depression severity measure," J. Gen. Intern. Med., vol. 16, no. 9, pp. 606–613, 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.

[2] Torous, J., K. J. Myrick, N. Rauseo-Ricupero, and J. Firth, "Digital mental health and COVID-19: Using technology today to accelerate the curve on access and quality tomorrow," JMIR Ment. Health, vol. 7, no. 3, Art. no. e18848, 2020, doi: 10.2196/18848.

[3] Onnela, J. P., and S. L. Rauch, "Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health," Neuropsychopharmacology, vol. 41, no. 7, pp. 1691–1696, 2016, doi: 10.1038/npp.2016.7.

[4] Insel, T. R., "Digital phenotyping: Technology for a new science of behavior," JAMA, vol. 318, no. 13, pp. 1215–1216, 2017, doi: 10.1001/jama.2017.11295.

[5] Shatte, A. B. R., D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," Psychol. Med., vol. 49, no. 9, pp. 1426–1448, 2019, doi: 10.1017/S0033291719000151.

[6] Mohr, D. C., M. Zhang, and S. M. Schueller, "Personal sensing: Understanding mental health using ubiquitous sensors and machine learning," Annu. Rev. Clin. Psychol., vol. 13, pp. 23–47, 2017, doi: 10.1146/annurev-clinpsy-032816-044949.

[7] Saeb, S., et al., "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study," J. Med. Internet Res., vol. 17, no. 7, Art. no. e175, 2015, doi: 10.2196/jmir.4273.

[8] Jacobson, N. C., H. Weingarden, and S. Wilhelm, "Digital biomarkers of mood disorders and symptom change," NPJ Digit. Med., vol. 2, Art. no. 3, 2019, doi: 10.1038/s41746-019-0078-0.

[9] Oudin, A., et al., "Digital phenotyping: Data-driven psychiatry to redefine mental health," J. Med. Internet Res., vol. 25, Art. no. e44502, 2023, doi: 10.2196/44502.

[10] Palermo, F., et al., "TIHM: An open dataset for remote healthcare monitoring in dementia," Sci. Data, vol. 10, Art. no. 606, 2023, doi: 10.1038/s41597-023-02519-y.

[11] Rostill, H., et al., "Technology integrated health management for dementia," Br. J. Community Nurs., vol. 23, no. 10, pp. 502–508, 2018, doi: 10.12968/bjcn.2018.23.10.502.

[12] Ramesh, K., et al., "Machine learning prediction of agitation in dementia patients using sleep and physiological data," Appl. Sci., vol. 15, no. 18, Art. no. 9908, 2025, doi: 10.3390/app15189908.

[13] Che, Z., S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," Sci. Rep., vol. 8, Art. no. 6085, 2018, doi: 10.1038/s41598-018-24271-9.

[14] Tomašev, N., et al., "A clinically applicable approach to continuous prediction of future acute kidney injury," Nature, vol. 572, no. 7767, pp. 116–119, 2019, doi: 10.1038/s41586-019-1390-1.

[15] Noci, F., L. Agrawal, A. Achtari, and P. Meyer, "Wearable technologies to predict and prevent heart failure hospitalizations: A systematic review," Eur. Heart J. Digit. Health, vol. 6, no. 5, pp. 868–877, 2025, doi: 10.1093/ehjdh/ztaf079.

[16] Bi, S., et al., "Predicting depression risk in middle-aged and elderly adults in China using CNN-BiLSTM-Attention mechanism and LSTM+SHAP framework," BMC Psychiatry, vol. 25, Art. no. 787, 2025, doi: 10.1186/s12888-025-07178-4.

[17] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, 1st ed. Piscataway, NJ, USA: IEEE, 2019. [Online]. Available: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

[18] Friedman, J. H., "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001. [Online]. Available: https://www.jstor.org/stable/2699986

[19] Chicco, D., and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, Art. no. 6, 2020, doi: 10.1186/s12864-019-6413-7.

[20] Efron, B., and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall/CRC, 1993.

[21] Jacobson, N. C., and Y. J. Chung, "Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression using smartphones," Sensors, vol. 21, no. 11, Art. no. 3771, 2021, doi: 10.3390/s21113771.

[22] Martinez-Martin, N., H. T. Greely, and M. K. Cho, "Ethical development of digital phenotyping tools for mental health applications: Delphi study," JMIR Mhealth Uhealth, vol. 9, no. 7, Art. no. e27343, 2021, doi: 10.2196/27343.

[23] McMahan, H. B., E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS), vol. 54, 2017, pp. 1273–1282, doi: 10.48550/arXiv.1602.05629.

[24] Sutton, R. T., et al., "An overview of clinical decision support systems: Benefits, risks, and strategies for success," NPJ Digit. Med., vol. 3, Art. no. 17, 2020, doi: 10.1038/s41746-020-0221-y.

[25] Lim, B., S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," Int. J. Forecast., vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.

[26] Beutel, D. J., et al., "Flower: A friendly federated learning research framework," arXiv:2007.14390, 2020, doi: 10.48550/arXiv.2007.14390.