

Article

From Patterns to Deviations: Detecting Behavioural Drift for Mental Health Monitoring Using Smartphone and Wearable Data

Eden Tolosa, Isibor Kennedy Ihianle * , Pedro Machado , Salisu Wada Yahaya  and Ahmad Lotfi 

Department of Computer Science, Nottingham Trent University, Nottingham NG1 4FQ, UK; eden.tolosa2024@my.ntu.ac.uk (E.T.); pedro.machado@ntu.ac.uk (P.M.); salisu.yahaya@ntu.ac.uk (S.W.Y.); ahmad.lotfi@ntu.ac.uk (A.L.)

* Correspondence: isibor.ihianle@ntu.ac.uk

Abstract

Monitoring behavioural drift, a sustained shift in an individual's daily activity, sleep, or social patterns, offers a significant lens for early mental health intervention. However, detecting these drifts in free-living settings remains challenging due to the absence of ground-truth labels, the temporal complexity of human behaviour, and fragmentation across heterogeneous sensing modalities. This paper proposes a multimodal approach to quantify and detect behavioural drift using longitudinal data from over 500 university students in the NetHealth cohort. We extract personalised, longitudinal features spanning three behavioural domains: physical activity, sleep hygiene, and communication diversity and model deviations relative to rolling, individual-specific statistical baselines. To differentiate transient anomalies from meaningful behavioural change, we introduce a sustained streak mechanism that identifies persistent drift episodes. We evaluate three temporal modelling strategies: Isolation Forest, Convolutional Neural Networks, and Long Short-Term Memory networks across both single-modality and fused approaches. Our findings indicate that recurrent models offer the strongest performance, highlighting the necessity of capturing temporal dependencies in behavioural data. Furthermore, we find that cross-modal correlations between drift signals are weak, confirming that activities, sleep, and communication provide complementary, non-redundant insights into an individual's well-being. This work establishes the methodological basis for integrating multimodal sensing data to monitor mental health trajectories, providing a scalable path towards early intervention in digital health.

Keywords: behavioural drift; anomaly detection; unsupervised detection; digital phenotyping; multimodal



Academic Editor: Enzo Pasquale Scilingo

Received: 25 December 2025

Revised: 2 February 2026

Accepted: 16 February 2026

Published: 20 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The past decade has witnessed a rapid transformation in the study of human behaviour, driven by the ubiquity of smartphones and wearable devices. This paradigm, frequently referred to as digital phenotyping, allows for the continuous, unobtrusive collection of behavioural and physiological signals in real-world environments [1–3]. Unlike traditional clinical assessments, which provide only a static “snapshot” of health, passive sensing enables a high-resolution, longitudinal “video” of an individual's daily life. This includes fine-grained measurements of physical activity (e.g., step counts and movement variance, sleep behaviour (e.g., onset, duration, and nocturnal interruptions), and social connectivity (e.g., call and SMS frequency and contact diversity). The widespread adoption of smartphones and wearables,

particularly among young adults, has made such continuous behavioural monitoring increasingly feasible. As of 2023, 98% of UK residents aged 16–24 own a smartphone [4], while the use of wearables such as Fitbit and Apple Watch has transitioned from niche fitness tracking to mainstream health monitoring. For university populations, a demographic notoriously prone to fluctuating stress levels and burgeoning mental health challenges, this creates a unique opportunity for large-scale, real-time monitoring. However, the true value of these data streams lies not in their volume, but in our ability to detect meaningful departures from an individual's unique "status quo."

Despite the potential of digital sensing, much of the current computational landscape remains ill-suited to detect the gradual, cumulative changes that often characterise early psychological decline. In sensor-based behavioural analysis, human activity recognition (HAR) has long been the predominant approach. These systems typically segment data into short windows, for example, 10–60 s, to classify discrete activities such as walking or sitting [5,6]. While effective for identifying momentary states, HAR is limited for understanding sustained behavioural change; it assumes independence among segments and overlooks the long-term temporal continuity that defines human habits.

To bridge this gap, this paper introduces the concept of behavioural drift. While an anomaly might represent a single night of sleep deprivation or an isolated social event, drift refers to a sustained and gradual shift in the baseline itself. In a mental health context, drift may manifest as a progressive reduction in social entropy, a subtle but consistent delay in sleep onset, or a creeping decline in physical mobility, all of which are documented precursors to depressive or anxious episodes. The urgency of developing drift detection mechanisms is underscored by a worsening global mental health crisis. According to the World Health Organization [7], mental disorders affect more than one in four adults and represent 14% of the global disease burden. Among young adults, the figures are particularly stark—for example, in the United Kingdom, nearly 28% of individuals aged 16–29 reported moderate to severe depressive symptoms in late 2022 [8]. Traditional monitoring relying on episodic clinical visits or retrospective self-reports is often "too little, too late." By the time a student seeks help, they may have been experiencing behavioural drift for weeks. Detecting these deviations as they occur offers a path towards proactive, rather than reactive, intervention.

There is a significant need for computational approaches that can synthesise multi-modal data to detect these longitudinal shifts. This paper addresses this need by designing and evaluating an approach to behavioural drift detection using the NetHealth dataset, which contains over 500 university students' longitudinal data. The proposed approach models individual baselines and flags deviations across physical, social, and circadian domains to provide methodological insights for proactive health monitoring. In doing so, it contributes a principled approach for modelling sustained behavioural drift under real-world constraints of label scarcity, heterogeneity, and individual variability. More specifically, this paper makes three primary contributions:

- We propose a per-user modelling approach using sliding temporal windows to construct stable, intra-individual behavioural baselines.
- We provide a comparative analysis of three modelling strategies, Isolation Forest, Convolutional Neural Networks, and Long Short-Term Memory networks, to determine which architecture best captures the "memory" of human behaviour under real-world conditions.
- We introduce a late-fusion strategy that integrates drift signals from disparate modalities (activity, sleep, and communication) to produce interpretable alerts, demonstrating that while these signals are weakly correlated, they are highly complementary.

The remainder of this paper is structured as follows. Section 2 reviews related work on behavioural modelling and drift detection. Section 3 details the methodology, including the NetHealth dataset and our preprocessing pipelines. Section 4 presents our experimental findings, followed by a discussion of implications and future work in Section 5.

2. Related Work

Human activity recognition (HAR) provides a foundational framework for modelling sensor-based behaviour, focusing primarily on the detection and classification of physical actions via smartphones and wearables [5,9]. Early HAR systems utilised short-window architectures, typically segmenting high-frequency data from accelerometers and gyroscopes into 10–60 s intervals for independent analysis through statistical feature extraction and classical classifiers like Random Forest or Support Vector Machines [10,11]. While effective at identifying immediate activity states, these traditional approaches suffer from a significant structural limitation in that they treat activity segments as independent events. Consequently, the higher-order temporal continuity and routine context, such as circadian regularities and behavioural transitions, remain largely invisible to these models, making them ill-suited for detecting the gradual and cumulative behavioural shifts that often precede psychological decline [12,13].

Recent advances in deep learning have improved feature extraction and short-term temporal modelling through Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which excel at capturing local patterns and sequential dependencies in behavioural signals [14–16]. CNNs are particularly effective at extracting local temporal features from multi-modal data to identify structured activity states [17], while LSTMs extend these capabilities by retaining information over longer periods to model periodic behaviours [18]. More recently, transformer architectures and pretrained Large Language Models (LLMs) have been introduced to wearable sensing to leverage attention mechanisms for modelling global dependencies across entire sequences [19]. Despite these improvements, most deep learning systems remain restricted by their reliance on supervised learning and high-fidelity labels. In free-living mental health monitoring, ground-truth annotations of behavioural decline are scarce, forcing a shift towards semi-supervised or unsupervised approaches that can adapt to new users with minimal intervention [20].

Beyond deep learning, researchers have explored sequential models such as bag-of-words (BoW) and process mining to capture the higher-order dynamics of daily routines. BoW frameworks summarise activity streams as frequency histograms of discrete events within fixed intervals, successfully capturing inter-individual differences in routine structure [21,22]. However, BoW approaches discard the specific order and transitions of activities, which may encode critical signatures of behavioural change. Conversely, process mining aims to recover temporal dependencies and routine structures by analysing event logs to detect disruptions in sequential patterns [23]. While powerful for domains with clean, structured event logs, process mining remains computationally intensive and less resilient to the inherent noise of wearable and smartphone data, limiting its scalability in unconstrained longitudinal studies [24].

Current drift detection literature operationalises behavioural shifts as statistically atypical patterns relative to personalised baselines. Unlike sudden anomalies, behavioural drift refers to gradual, sustained changes in routine that persist beyond expected cyclical variations, such as weekend effects or academic holidays [25,26]. To address the limitations of population-level thresholds, recent studies utilise sliding window techniques to track deviations from individual-specific means and variances over multi-day or multi-week intervals [27,28]. As summarised in Table 1, the design axis for these systems centres on day-

scale windowing strategies that stabilise behavioural variance. Furthermore, the grouping of drift days into contiguous “streaks” or episodes has emerged as a vital mechanism for distinguishing meaningful disruption from stochastic noise, aligning with clinical perspectives on relapse where the duration and persistence of change are critical markers of health status [29]. Despite progress in single-modality drift detection, there remains a notable gap in testing whether these deviations co-occur across heterogeneous streams, motivating the need for the multimodal late-fusion approach proposed in this study.

Table 1. Approaches with day-scale windowing methods.

Approach	Window Size	Preprocessing	Reference
U-BEHAVED	3-day rolling baseline (configurable)	Aggregate steps/hour; compute rolling mean and IQR; flag consecutive outliers	[28]
DynAmo	Configurable length	Dynamic clustering to capture trends	[25]
PCAR	Multi-month windows	HAR aggregation into daily activity curves	[27]
PCAR	≤ 7 days	KL-divergence on day-level step summaries with smoothing to avoid zeros	[26]
eB ² feasibility study (GPS)	> 1 month	Clustering of daily profiles; change point detection on profile probabilities	[30]
Daily multivariate anomaly detection	Daily features over months; analysis focuses on two weeks pre-relapse	Daily mobility and sociability features; anomaly detection on multivariate daily vectors	[29]

Recent advances in modality-specific behavioural sensing further underscore the field’s transition from static summary statistics towards dynamic, context-aware modelling. In the domain of physical activity, one of the most frequently utilised inputs in digital phenotyping research has moved beyond simple daily step counts and aggregate movement intensities to emphasise temporal structure and within-person context. While early studies reduced physical behaviour to daily averages, thereby obscuring the nuance of sedentary bouts and movement patterns [31], more recent unsupervised approaches compare adjacent multi-day windows or model mixtures of recurring “day types” to characterise longitudinal change [26,29]. For example, U-BEHAVED [28] demonstrates that rolling personalised baselines with robust dispersion estimates substantially improve sensitivity to gradual habit shifts while filtering routine weekly variability. Sleep has likewise emerged as a critical modality for behavioural monitoring, with measures such as duration, onset, and efficiency serving as sensitive indicators of stress, fatigue, and emotional dysregulation. Recent unsupervised methods have moved away from aggregate sleep summaries towards modelling circadian structure and within-person phase shifts using techniques such as change-point detection, which are better suited to capturing sustained deviations that static averages often fail to detect [32,33]. More broadly, multimodal passive sensing has been widely explored for mental health inference using activity, sleep, mobility, and

communication features, often relying on supervised learning or population-level models. Recent systematic reviews of Digital Phenotyping of Mental Health synthesise these efforts and consistently highlight persistent challenges related to temporal modelling, label scarcity, and real-world deployment [34]. The present work directly addresses these gaps through an empirically validated, unsupervised approach for detecting sustained behavioural drift.

Social behaviour and communication represent another vital dimension, where social withdrawal is a primary hallmark of psychological distress. Passive sensing of communication logs allows for the inference of social rhythm stability, with studies showing that reduced phone interaction and social entropy correlate strongly with mood instability [35,36]. Shannon entropy, originally derived from information theory [37], has been successfully adapted to characterise ego-network diversity, providing a proxy for the stability of social ties that extends beyond raw event volume [38]. A decline in contact entropy often signals social withdrawal and isolation risk, even when total interaction counts remain stable, yet many existing implementations still prioritise mobility metrics over these complex communication structures [30,39].

The inherent multimodality of human behaviour involving movement, sleep, and social interaction suggests that relying on a single data stream risks overlooking critical patterns of strain. While reduced physical activity might stem from various causes, its co-occurrence with disrupted sleep and social withdrawal provides a far more robust indicator of psychological decline. However, the integration of these heterogeneous streams poses significant technical challenges, particularly regarding scale heterogeneity and data missingness [40]. Early fusion strategies, which concatenate streams before modelling, are often vulnerable to these issues, whereas recent work suggests that decision-level late fusion is more effective. This strategy allows modality-specific models to be trained independently, preserving the distinct statistical properties of each domain while enabling flexible integration through neural network architectures or weighted ensembles [41].

Despite these advances, several critical gaps remain in the current state of behavioural drift research. A primary challenge is the absence of standardised continuous ground-truth labels for evaluating behavioural change, which undermines the comparability and generalisability of proposed systems [42,43]. Furthermore, most existing approaches continue to rely on pooled models or cross-user normalisation, implicitly assuming that deviations can be assessed against population averages. Evidence from human activity recognition suggests that such models often fail to generalise, reinforcing the need for truly individualised historical baselines [42,44]. There is also a persistent lack of temporal sensitivity; while deep sequence models like recurrent neural networks have been applied in adjacent fields like affect recognition, their application in unsupervised drift detection remains limited. Finally, most models still treat behavioural domains in isolation, overlooking the joint evolution and temporal alignment of activity, sleep, and social patterns. This lack of integration prevents the identification of cross-domain signatures of drift and limits the interpretability of automated alerts in clinical settings.

These limitations motivate the need for a personalised, temporally sensitive, and multimodal approach to behavioural drift detection. In response, we propose a methodological approach that models behavioural change relative to individual historical baselines. Temporal persistence is explicitly considered to separate sustained drift from short-term fluctuations. Activity, sleep, and communication signals are integrated to capture cross-domain behavioural dynamics. The resulting approach provides the foundation for the methodology described in Section 3.

3. Proposed Methodology

The proposed behavioural drift detection approach is illustrated in Figure 1. Building on the motivation outlined in Section 1, the methodology is designed to support personalised detection of behavioural change utilising longitudinal sensing data by defining adaptive individual baselines, preserving the temporal structure of heterogeneous behavioural modalities, and integrating modality-specific signals through decision-level fusion.

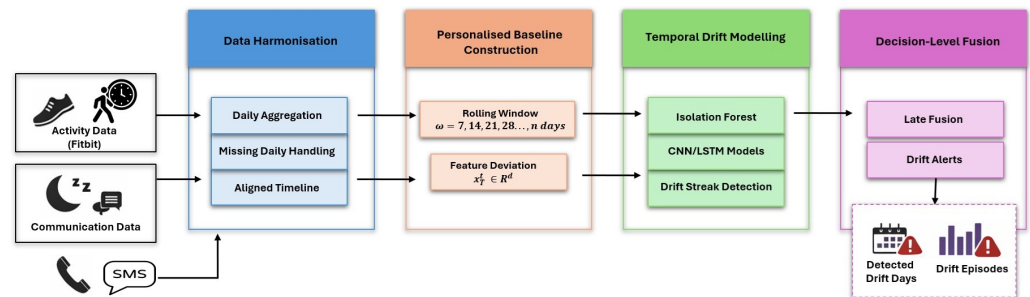


Figure 1. Schematic representation of the proposed multimodal drift detection approach, illustrating the pipeline from raw sensor ingestion and daily data aggregation to personalised baseline modelling and late-fusion alert generation.

As shown in Figure 1, heterogeneous sensing streams, Fitbit-derived physical activity and sleep records and smartphone-based communication events are first harmonised via aggregation to a common daily resolution. This alignment produces a unified, per-user longitudinal timeline that preserves contextual richness and enables cross-modal comparison of behavioural dynamics. For each user u and day t , multimodal daily behaviour is represented by a feature vector $x_t^u \in \mathbb{R}^d$. Personalised behavioural baselines are then constructed using rolling statistical windows $\omega \in \{7, 14, 21, 28\}$ days, which are continuously updated to remain sensitive to both transient anomalies and gradual long-term behavioural shifts. Deviations from these individual-specific baselines are quantified over time and aggregated into contiguous drift streaks, allowing sustained behavioural change to be distinguished from isolated short-term fluctuations. Drift detection is performed independently within each modality in order to preserve their distinct temporal and statistical properties. The resulting modality-specific drift signals are subsequently integrated using a decision-level late fusion strategy, generating interpretable behavioural drift indicators and fusion alerts at the individual level.

3.1. Problem Formulation

To formalise the detection of behavioural drift, we represent the longitudinal data as a set of time-ordered observations for each individual. Let $U = \{u_1, u_2, \dots, u_m\}$ denote the set of participants. For each user $u \in U$, we define a multivariate time series:

$$X^u = x_1^u, x_2^u, \dots, x_T^u \quad (1)$$

where $x_t^u \in \mathbb{R}^d$ is a d -dimensional feature vector representing daily metrics across the physical activity, sleep, and communication domains at time t . Unlike traditional point-anomaly detection, which seeks to identify isolated outliers, our objective is to identify sustained departures from a rolling baseline. We define a “drift day” as a day where the observed behaviour significantly deviates from the established intra-individual distribution. To capture the cumulative nature of these shifts, we introduce the drift-streak label, s_t^u . This label is defined as

$$s_t^u = \begin{cases} k, & \text{if day } t \text{ is the } k\text{-th day of a contiguous drift episode,} \\ 0, & \text{if behaviour at day } t \text{ is within baseline expectations,} \end{cases} \quad (2)$$

We define $s_t^u > 0$ as the length of the current streak; we encode the temporal persistence of the deviation directly into the target variable. This allows the model to differentiate between a “one-off” disruption, for example, a single night of poor sleep, and a developing trend, such as a week-long decay in sleep efficiency. The primary computational task is to learn a mapping function $\mathcal{F}$ that predicts the state of behavioural drift based on recent history. Given a look-back window of size ω , the goal is to learn:

$$\mathcal{F} : (x_{t-\omega+1}^u, \dots, x_t^u) \rightarrow s_t^u \quad (3)$$

where $\mathcal{F}_0$ ingests the sequence of feature vectors over the preceding ω days and outputs the drift-streak magnitude s_t^u for the current day t . This formulation enables the evaluation of different architectural strategies, ranging from a memoryless Isolation Forest to sequence-aware models such as LSTM, to assess their efficacy in mapping raw behavioural patterns to sustained longitudinal drift. Modality-specific drift probabilities are subsequently combined at the decision level using a late-fusion rule, which is formally defined and empirically evaluated in Section 4.3.

3.2. Dataset Description

For experimental purposes, we utilise the NetHealth dataset [45], a longitudinal repository comprising Fitbit and smartphone-derived sensing data collected from a large cohort of undergraduate students between Fall 2015 and Spring 2019 at the University of Notre Dame. Recruitment was conducted in a three-tier strategy, as detailed in Table 2, starting with an initial cohort of 698 participants in the first year. While 686 participants remained active by the second year, the cohort experienced natural attrition due to non-compliance and graduation, reducing to approximately 320 by the fourth year [45]. The study was conducted under strict institutional IRB protocols, ensuring informed consent and the de-identification of all participants via case codes. To provide environmental context, the dataset includes an academic calendar containing semester dates, holidays, and exam periods, which allow the approach to distinguish between academic-induced stress and underlying behavioural drift.

Table 2. NetHealth dataset cohort dynamics: recruitment phases and longitudinal participant retention over the four-year study period.

Longitudinal Retention	Timeline	Participants (<i>n</i>)
Year 1	2015–2016	698
Year 2	2016–2017	686
Year 3	2017–2018	~435
Year 4	2018–2019	~320

The approach uses three primary behavioural modalities to quantify drift. Physical activity data were captured via wrist-worn Fitbit devices with Charge HR and Charge 2, providing daily step counts and intensity minutes. These are used to derive Moderate-to-Vigorous Physical Activity (MVPA) and sedentary duration, with data quality controlled by a wear-compliance metric based on valid heart rate minutes. Sleep hygiene is similarly derived from Fitbit logs, where individual sleep segments separated by less than 30 min are merged into a primary daily episode. This provides granular metrics, including sleep onset, waking time, efficiency, and Wake After Sleep Onset (WASO), serving as proxies for circadian stability. Social and communication dynamics were monitored through a back-

ground smartphone application. This application recorded over 60 million timestamped events, including incoming and outgoing calls and SMS messages. These logs enable the construction of daily interaction counts and social entropy, a measure of contact diversity and social balance based on Shannon entropy [46].

Although the NetHealth dataset includes self-reported mental health assessments, these labels are not used in this study, as our objective is to evaluate behavioural drift detection using passive sensing data alone.

3.3. Data Processing Pipelines

To transform raw sensing logs into the structured feature vector x_t^u defined in Section 4, each modality underwent a rigorous preprocessing pipeline. All streams were first validated against the official NetHealth codebooks for temporal consistency and identifier integrity. To ensure data fidelity, we excluded days with a compliance of device wear-time below 80%, a threshold standard for avoiding spurious drift detections caused by poor compliance [47].

3.3.1. Physical Activity Pipeline

The activity pipeline processes daily summaries of step counts and intensity. We applied plausibility thresholds to remove sensor artefacts based on established Fitbit validation studies [48,49]. Missing data were not imputed but were explicitly flagged. In longitudinal monitoring, the absence of data can serve as a meaningful indicator of behavioural disengagement rather than random sensor failure [50]. A participant-day is considered valid for the activity modality if wearable device wear-time exceeds 80% of daily minutes (≥ 1152 min). Daily activity features are computed only on valid days. Days that do not meet the compliance threshold are excluded from downstream analysis. After filtering, activity features exhibit low feature-level missingness ($<1\%$), and participant-days failing the compliance criterion are excluded from downstream analysis. Outliers were identified using the Interquartile Range (IQR) rule— $[Q1 - 1.5IQR, Q3 + 1.5IQR]$. To reduce dimensionality, we calculated MVPA as a health-relevant composite of fairly active minutes and very active minutes [51].

3.3.2. Sleep Hygiene Pipeline

The sleep pipeline extracts nightly metrics (duration, onset, efficiency, and WASO). To maintain temporal alignment, all sleep records were keyed to the “wake date.” We applied standard sleep hygiene formulas to verify metric consistency:

$$\text{Sleep Efficiency} = \frac{\text{Total Sleep Time}}{\text{Total Minutes In Bed}} = \frac{\|\text{Sleep Period}\| - \text{WASO}}{\|\text{Sleep Period}\| + \text{Latency}} \quad (4)$$

Records indicating a “Time in Bed” of less than 5 min or exceeding 24 h were removed as erroneous.

3.3.3. Communication and Social Pipeline

Smartphone logs were aggregated at the daily per-user level, and days with no communication activity were retained as valid observations. To quantify social diversity, we calculated social entropy (H_t) based on the distribution of unique communication partners:

$$H_t = - \sum_{i=1}^n p_i \log p_i \quad (5)$$

where p_i represents the proportion of events with contact i on day t . This metric distinguishes between concentrated interactions with a single primary contact (low entropy) and broad social engagement across a diverse network (high entropy) [39]. To distinguish

between well-being-related drift and academic-induced stress, each participant-day was merged with an academic calendar. We integrated binary flags for weekends and holidays, alongside categorical variables for semester phases for example orientation, midterms and finals. This allows the subsequent models to adjust for population-wide shifts in behaviour during exam periods [52].

A summary of modality-level missingness and the number of excluded participant-days for each modality is provided in Table 3.

Table 3. Modality-level valid-day criteria and missingness statistics.

Modality	Valid-Day Criterion	Feature	Missing Rate (%)
Activity	Wear time $\geq$ 80%	Steps	0.30
		Floors	0.25
		Activity minutes	0.55
Sleep	Complete sleep episode	Bed exit time	0.0006
		Core sleep metrics	0.00
Communication	Daily aggregation	Call duration	97.7
		Message length	10.9
		Media bytes	96.5

3.4. Machine Learning Architectures and Fusion Strategy

To address unsupervised behavioural drift detection across heterogeneous sensing streams, we evaluate a set of complementary modelling approaches that serve distinct analytical purposes rather than forming a single sequential pipeline. Specifically, we employ a non-parametric statistical model, Isolation Forest, as a memoryless baseline, and compare it against deep sequence architectures—1D-CNN and LSTM—that explicitly model temporal dependencies using personalised sliding windows. These models are trained and evaluated independently under a common data formulation to assess the contribution of temporal memory. Only at the final stage are modality-specific LSTM outputs combined via decision-level late fusion to produce an integrated drift signal.

3.4.1. Baseline Modelling: Isolation Forest

Following the rolling window feature transformation, we implemented an Isolation Forest as a baseline for daily drift detection. Models were trained independently for each participant to ensure that anomalies were defined relative to a person’s own historical behaviour. The Isolation Forest was selected for its computational efficiency and its ability to identify outliers in high-dimensional feature spaces without requiring labelled ground truth, a critical requirement for digital phenotyping where labelled outcomes are often sparse [53,54]. Although Isolation Forest is an unsupervised anomaly detector, we report classification-style metrics by thresholding anomaly scores and comparing them against proxy drift labels for retrospective evaluation only; no labels are used during model training. As shown in Table 4, performance varies substantially across modalities.

Table 4. Test set performance of the unsupervised Isolation Forest across modalities. Classification-style metrics are computed by thresholding anomaly scores and comparing them against proxy drift labels for retrospective evaluation only.

Modality	Accuracy	Precision	Recall	ROC-AUC	PR-AUC
Activity	0.829	0.614	0.476	0.789	0.561
Sleep	0.932	0.878	0.783	0.981	0.926
Communication	0.828	0.514	0.515	0.770	0.495

The model assigns anomaly scores based on the path length required to isolate a given observation, effectively flagging behavioural days that lie at the periphery of an individual's typical feature distribution. Importantly, the Isolation Forest operates in a memoryless manner, treating each day independently and providing a statistical reference point for drift detection. These per-day anomaly scores are interpreted as proxy indicators of behavioural deviation and are used solely for comparative evaluation against sequence-aware models, rather than as clinical or ground-truth labels.

3.4.2. Deep Sequence Architectures

While the Isolation Forest treats days as independent events, behavioural drift often manifests as a sustained decay of routine. To capture these temporal motifs, we implemented convolutional and recurrent neural networks.

The CNN architecture comprised two 1D convolutional layers with varied kernel sizes $\{3, 5, 7\}$ to capture multi-scale local motifs, such as sudden drops in communication volume or bursts of sedentary activity. Each layer was followed by ReLU activation and dropout for regularisation. A global average pooling layer was utilised to summarise the temporal context before a dense sigmoid output layer. CNNs are particularly effective for identifying stationary temporal patterns in behavioural time series while remaining computationally more straightforward to train than fully recurrent architectures [13].

To model sequential dependencies within each sliding window, we implemented a stacked LSTM network. The architecture comprised two LSTM layers with 64 and 32 hidden units, respectively. The first layer returned the full sequence of hidden states (`return_sequences=True`), while the second returned the final hidden state (`return_sequences=False`). Batch normalisation and dropout ($p = 0.3$) were applied after each recurrent layer. The recurrent stack was followed by a fully connected layer with 64 ReLU-activated units (dropout $p = 0.3$) and a final sigmoid-activated output layer for binary drift prediction.

The model was trained using the Adam optimiser with binary cross-entropy loss and evaluated using accuracy, precision, and recall. Class imbalance was addressed through class-weighted loss. Training was regularised using early stopping (patience of five epochs), learning rate reduction on plateau (factor 0.5, patience of three epochs), and model checkpointing based on validation loss. Models were trained for a maximum of 50 epochs with a batch size of 32. LSTMs are well-suited for capturing gradual behavioural change due to their ability to retain information over extended temporal horizons [12].

3.4.3. Multimodal Fusion Strategy

A central challenge in multimodal learning is the integration of information from disparate sources. In this paper, we adopted decision-level late fusion. This choice was motivated by three technical considerations. First, it prevents the statistical noise of one modality from distorting the signal of a stable modality. Second, it provides a robust mechanism for handling missing data; if a participant forgets to wear their Fitbit, a drift alert can still be reliably inferred from their smartphone entropy. Finally, late fusion ensures interpretability, allowing clinicians to audit which specific behavioural domain activity, sleep, or communication triggered a high-confidence alert. For each user-day, the modality-specific models generated a drift probability. To synthesise these into a global alert, we applied a Max-Fusion operator. This approach prioritises sensitivity by raising an alert if any single modality strongly signals drift. This ensures that domain-specific behavioural decay is not “suppressed” or averaged out by stability in other domains, a strategy aligned with the requirements of early warning systems in clinical health.

4. Experiments and Results

This section details the empirical evaluation of the behavioural drift detection approach across three distinct experiments. All computational tasks were performed in a Python 3.13 environment using an Anaconda distribution. While primary activity and sleep processing occurred on a standard workstation (16 GB of RAM), the high-velocity smartphone communication logs exceeding 60 million records necessitated a dedicated secondary machine to handle the significant I/O and processing load required for feature engineering.

To ensure a unified analysis, all raw sensing streams were temporally aligned into a consistent daily representation. Fitbit-derived metrics and smartphone event logs were summarised into daily indices of volume, diversity (Shannon entropy), and quality. Quality control was enforced by excluding days with wearable compliance below 80% to prevent the detection of spurious anomalies. Missing data were not discarded but were explicitly flagged, under the assumption that a lack of sensor activity may represent behavioural disengagement rather than a technical failure.

We conducted a series of experiments based on the analysis of behavioural constructs (Experiment I), followed by an assessment of temporal look-back windows on baseline stability (Experiment II). The final evaluation (Experiment III) benchmarks statistical and deep learning models to determine the efficacy of multimodal fusion in identifying sustained behavioural drift.

4.1. Experiment I: Exploratory Data Analysis and Feature Selection

We conducted an Exploratory Data Analysis to establish the latent structure of the multimodal dataset and identify redundant behavioural constructs. This step is critical to ensure that subsequent deep learning models are not biased by multicollinearity or high-dimensional noise. The analysis of physical activity and sleep metrics revealed significant feature overlap. As shown in the activity correlation matrix in Figure 2, several feature pairs exhibited high collinearity, particularly between intensity categories and caloric expenditure. Similarly, the sleep domain analysis, as depicted in Figure 3a, demonstrated that Time in Bed and Total Sleep Time were nearly identical in their variance ($r = 0.96$), while Wake After Sleep Onset (WASO) and Sleep Efficiency showed a strong inverse relationship ($r = -0.89$).

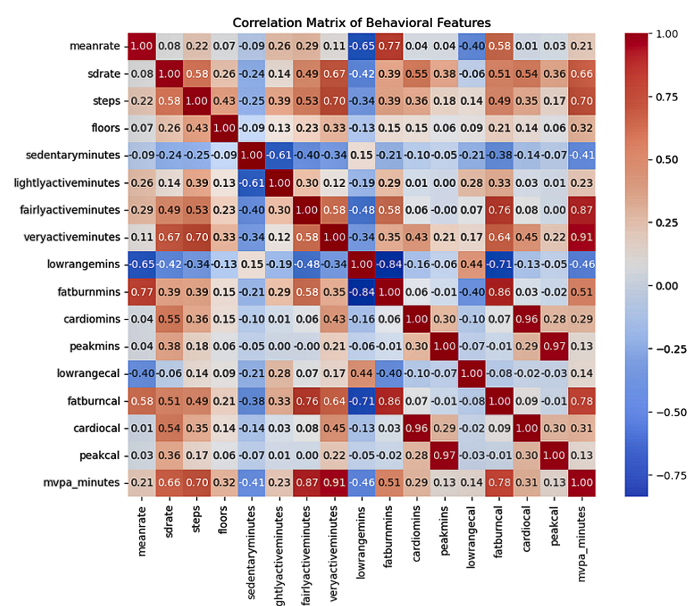


Figure 2. Correlation matrix of physical activity features, identifying overlapping constructs in movement intensity and energy expenditure.

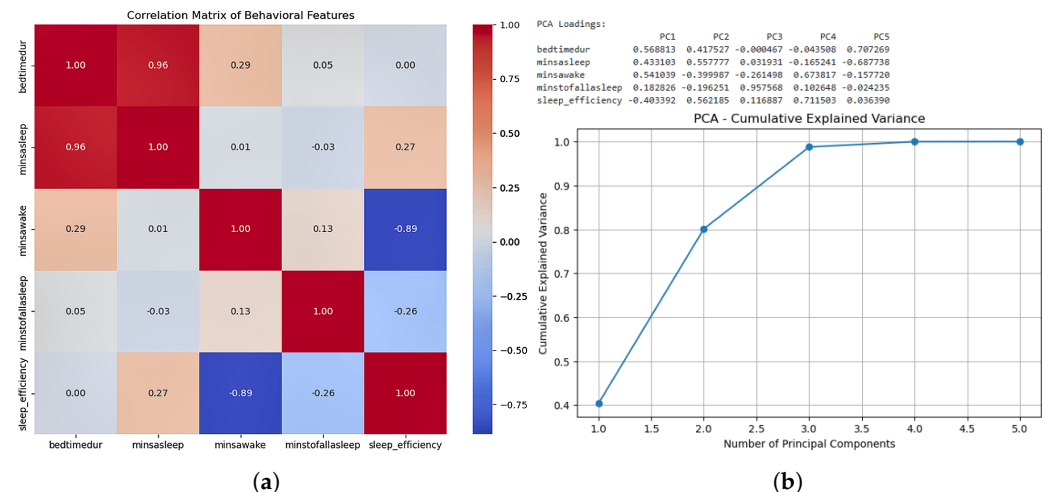


Figure 3. Pairwise correlation and PCA for sleep data to identify redundant features: (a) Correlation Matrix (b) PCA Variance Plot.

To quantify this redundancy, Principal Component Analysis (PCA) was applied. The results for sleep, visualised in Figure 3b, confirmed that just two components account for 80% of the domain's variance, while the first four activity components explain 75% of total variance.

Following the identification of redundant features, we utilised ANOVA F-scores to rank the discriminative power of the remaining behavioural predictors across all modalities. As illustrated in Figure 4, this univariate analysis highlighted the features most sensitive to state changes, such as sedentary minutes in activity, sleep efficiency, and contact entropy in communication. Based on the combined insights from the correlation matrices, PCA, and F-score rankings, we executed a feature reduction step, removing highly collinear variables like *veryactiveminutes* and *peakcal*. This reduction ensures a parsimonious input for the LSTM and CNN architectures, focusing the models on independent dimensions like circadian stability and general activity load.

The social interaction logs revealed a diverse range of social phenotypes within the student cohort. The distribution of Daily Contact Entropy (Figure 5) indicates a broad peak between 1.0 and 3.5, suggesting moderate contact diversity daily. However, a significant density of user-days is clustered at low entropy levels, reflecting days dominated by communication with only one or two primary contacts. The relationship between social volume and diversity is non-linear; as illustrated in Figure 6, contact entropy increases with communication events until a saturation point (~1000 events), suggesting a “social ceiling” regarding the active ties students can maintain daily.

Furthermore, temporal trend analysis (Figure 7) reveals distinct social behaviours: User 51,495 maintains consistently high entropy and event volume, indicating broad and regular social engagement. In contrast, User 90,905 exhibits long periods of social withdrawal (low entropy) punctuated by sudden, transient spikes in interaction. These distinct profiles underscore the necessity of the personalised baseline approach defined in our methodology, as a “normal” social day for one user may represent a significant “drift” for another.

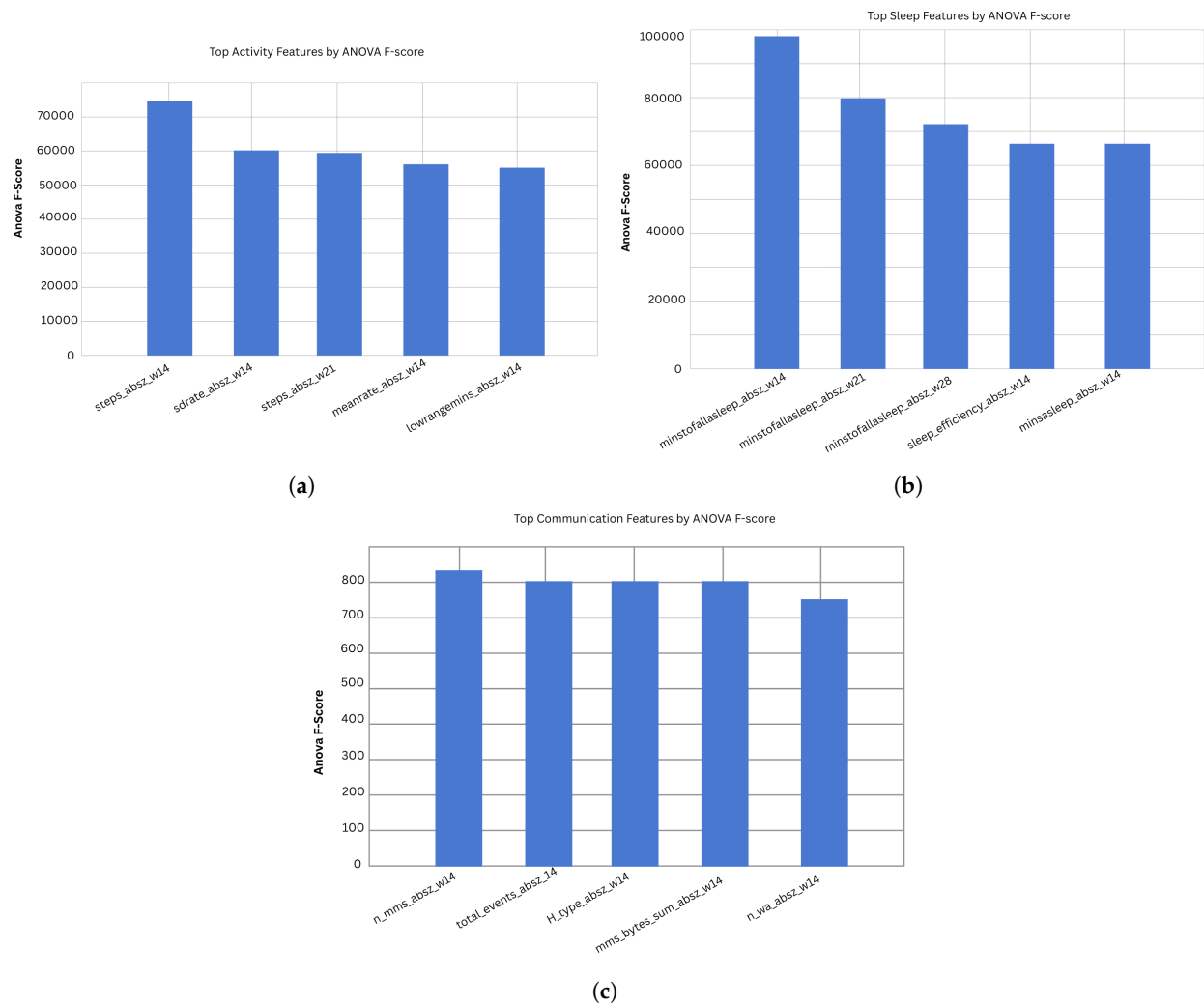


Figure 4. ANOVA F-score top-ranked predictors for each modality: (a) Activity data; (b) sleep data; (c) communication data.

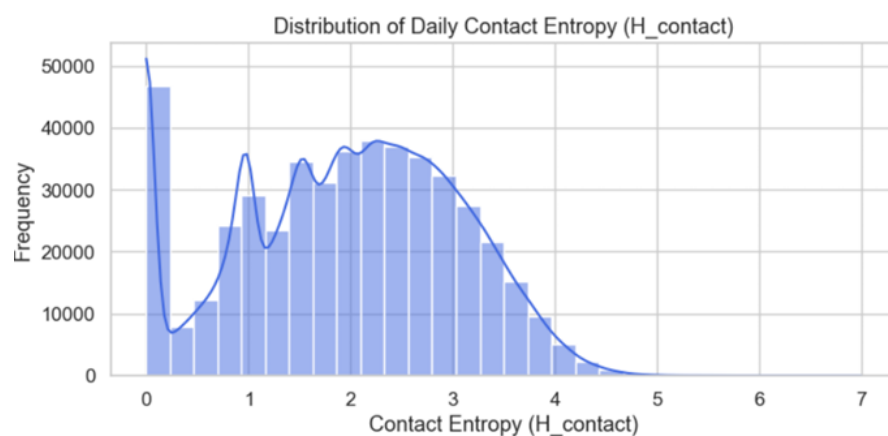


Figure 5. Distribution of Daily Contact Entropy across the student cohort.

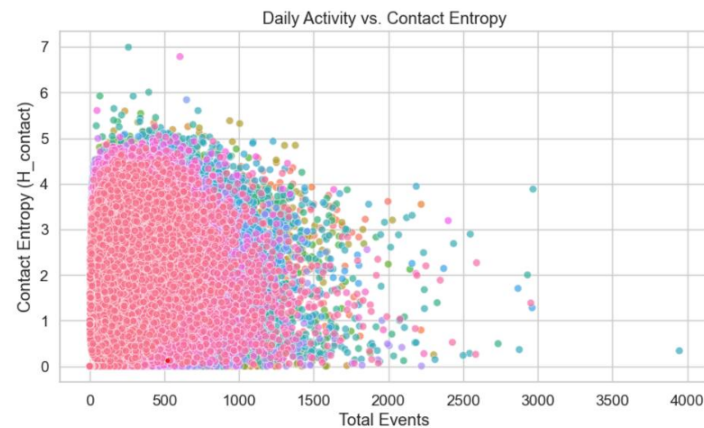


Figure 6. Relationship between daily communication volume and contact entropy.

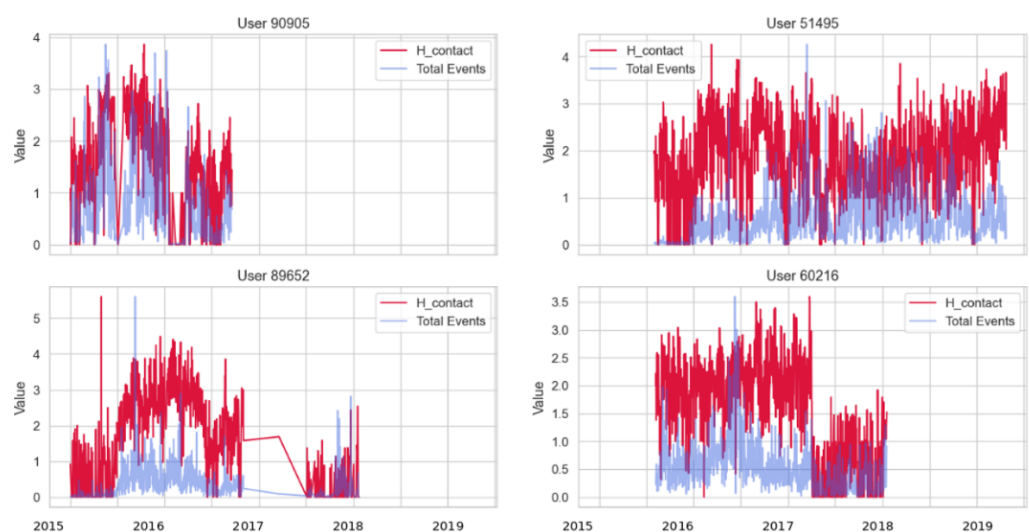


Figure 7. Temporal social trends for two sample users, illustrating high-diversity versus withdrawal-spike phenotypes.

4.2. Experiment II: Temporal Sensitivity Analysis

This experiment evaluates the impact of the look-back window parameter $\omega \in \{7, 14, 21, 28\}$ on the stability and discriminative power of the drift detection approach. The choice of ω determines the temporal resolution of the individual baseline. To maintain a scale-invariant comparison across heterogeneous features, for example, steps vs. minutes of sleep, we transformed raw observations into rolling z-scores.

4.2.1. Dynamic Standardisation and Composite Scoring

For each user u , day t , and feature f , the observation $x_{u,t,f}$ was standardised relative to its local historical distribution:

$$z_{u,t,f} = \frac{x_{u,t,f} - \mu_{u,t,f}(\omega)}{\sigma_{u,t,f}(\omega)} \quad (6)$$

where $\mu_{u,t,f}(\omega)$ and $\sigma_{u,t,f}(\omega)$ represent the rolling mean and standard deviation, respectively, over the preceding ω days. To capture the magnitude of deviation irrespective of its direction (e.g., both hypersomnia and insomnia), we utilised the absolute z-score. To synthesise these signals into a single indicator per modality, we calculated a daily composite drift score $S_{u,t}$:

$$S_{u,t} = \sum_{f \in F} |z_{u,t,f}| \quad (7)$$

4.2.2. Drift-Streak Identification and Window Impact

The approach identifies a “drift day” when the composite score $S_{u,t}$ exceeds a participant-specific quantile threshold. However, to distinguish sustained behavioural shifts from stochastic noise, we introduced a streak detection mechanism. A drift episode is defined as a contiguous sequence of k drift days:

$$E = d_t, d_{t+1}, \dots, d_{t+k} \mid S_{u,i} > \tau_u, \forall i \in [t, t+k] \quad (8)$$

where τ_u is the individual threshold. As shown in Table 5, the choice of the rolling window size ω significantly influences the morphology and frequency of these episodes, revealing a critical sensitivity-stability trade-off. Short windows of $\omega = 7$ yielded the highest frequency of episodes ($n = 367$ for communication) but were susceptible to “flickering,” a phenomenon where a single day of baseline recovery prematurely terminates a valid drift streak. Conversely, longer windows of $\omega = 28$ produced the most stable episodes but introduced a “lag effect.” In this scenario, the moving average $\mu(\omega)$ slowly absorbs the shifted behaviour into the baseline, eventually causing the z-score to return to zero even if the behaviour remains fundamentally altered. Consequently, a 14-day window was identified as optimal for this population, as it spans two full weekend-weekday cycles to filter out cyclic routine variance while remaining responsive to sustained behavioural decay. This threshold was selected to reflect clinically informed notions of persistence rather than short-term behavioural variability. In clinical psychiatry, the DSM-5 diagnostic criteria for Major Depressive Disorder require symptoms to persist for a minimum duration of two weeks to be considered clinically significant. Although the proposed approach is not intended as a diagnostic tool, adopting a 14-day persistence threshold provides an interpretable and principled basis for distinguishing sustained behavioural drift from transient anomalies.

We also examined alternative values of k (7, 21 and 28 days) and observed qualitatively similar trends in Table 5.

Table 5. Drift streak episodes across rolling window sizes and modalities.

Modality	Window (Days)	Episodes Detected	Drift Days Flagged/Valid Days
Activity	7	225	54,936/274,679
	14	234	54,936/274,679
	21	221	54,936/274,679
	28	215	54,936/274,679
Sleep	7	232	59,472/297,357
	14	222	59,472/297,357
	21	226	59,472/297,357
	28	241	59,472/297,357
Communication	7	367	92,480/471,110
	14	363	92,832/471,110
	21	364	93,000/471,110
	28	357	93,116/471,110

The 14-day window was identified as the optimal configuration for student populations, as it spans two full weekend-weekday cycles. This provides a sufficiently robust reference to filter out weekly routine variances while remaining responsive to the be-

havioural “decay” patterns visualised in the sample user trajectory in Figure 8. The streak logic ensures that the approach ignores isolated spikes, such as a single late-night party, but flags the third or fourth consecutive day of social withdrawal as a high-confidence drift episode.

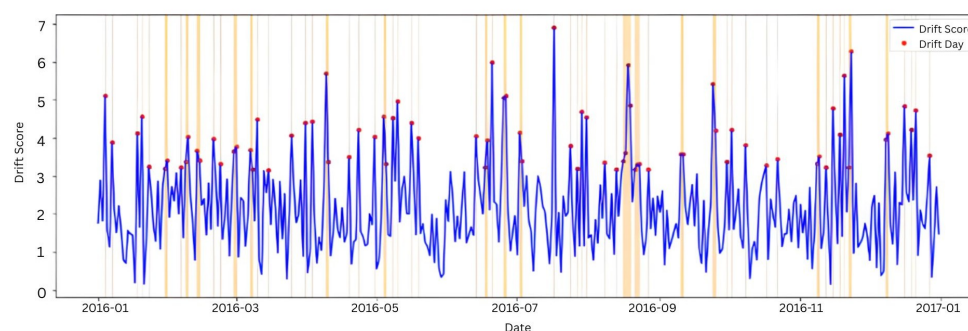


Figure 8. Sample drift streak score for a sample user for sleep data.

4.3. Experiment III: Multimodal Fusion and Model Performance Analysis

Experiment III evaluates the predictive performance and cross-modal complementarity of the proposed approach. By comparing a non-parametric baseline with deep sequence architectures, we quantify the value of temporal memory in drift detection. This analysis culminates in a late-fusion strategy that synthesises heterogeneous behavioural streams into a unified individual-level alert.

As part of this, we performed an initial benchmarking using the Isolation Forest algorithm, which provides a non-parametric assessment of daily drift. By modelling users independently with a 70/15/15 split for training, validation, and testing, the algorithm identifies anomalies based on their isolation in the feature space. To ensure comparability across diverse behavioural profiles, we applied per-user calibration mapping raw anomaly scores derived via the `score_samples` function to probability-like scores, where the median represents baseline behaviour, and the 95th percentile anchors extreme deviations. As detailed in Table 4, the Isolation Forest proved most effective in the sleep domain, achieving an accuracy of 0.932 and an ROC AUC of 0.981. However, its performance on activity and communication was more conservative, with PR-AUC scores of 0.561 and 0.495, respectively. These results highlight a fundamental limitation of static anomaly detectors: they treat each day as an independent event, frequently failing to capture the cumulative temporal decay characteristic of behavioural drift.

To address the temporal limitations of the statistical baseline, we implemented 1D-CNN and LSTM architectures. These models transform daily records into fixed-length windows, using an 80/20 stratified split to preserve class balance and prevent target leakage. We adopt this window-level stratified split intentionally to evaluate within-participant behavioural drift detection. While sliding windows may result in temporally adjacent samples from the same participant appearing in both training and test sets, this design reflects the intended deployment scenario, in which historical data from an individual is used to detect subsequent deviations in their own behaviour. Accordingly, the evaluation focuses on learning personalised temporal dynamics rather than cross-participant generalisation. Class weights were dynamically computed to handle the inherent sparsity of drift labels. The comparative results in Table 6 demonstrate that sequence modelling significantly improves detection sensitivity across all modalities.

In the activity domain, the LSTM achieved the highest ranking performance, with a PR-AUC of 0.703, a substantial improvement over the Isolation Forest’s 0.561. Similarly, the LSTM’s ability to model long-range dependencies provided the most balanced metrics in the communication domain, yielding an ROC-AUC of 0.844. Across all modalities, while

the CNN was adept at identifying short-term motifs, the LSTM was consistently superior in detecting gradual, longitudinal shifts. Consequently, the LSTM was selected as the primary input for the multimodal fusion layer.

Table 6. Test-set performance comparison of CNN and LSTM models across activity, sleep, and communication modalities.

Modality	Model	Accuracy	Precision	Recall	ROC-AUC	PR-AUC
Activity	CNN	0.814	0.535	0.726	0.857	0.690
	LSTM	0.806	0.520	0.759	0.867	0.703
Sleep	CNN	0.852	0.610	0.772	0.896	0.797
	LSTM	0.868	0.655	0.753	0.915	0.813
Communication	CNN	0.755	0.435	0.753	0.831	0.578
	LSTM	0.771	0.457	0.747	0.844	0.606

Multimodal Late Fusion and Complementarity

The decision-level late fusion was executed by aligning the modality-specific LSTM probabilities for each participant-day. Prior to fusion, we assessed the Pearson correlation between sensing streams. The results revealed near-zero correlations between activity, sleep, and communication ($r \approx 0.01$), statistically confirming that drift in one domain rarely overlaps immediately with another. This cross-modal independence validates our use of a Max-Fusion rule:

$$P_{\text{fused}}(t) = \max\{P_{\text{activity}}(t), P_{\text{sleep}}(t), P_{\text{communication}}(t)\}, \quad (9)$$

By adopting this operator, the system ensures that a significant drift signal in any single modality is preserved rather than averaged out. As demonstrated in the Confusion Matrix (Figure 9), this strategy optimises the detector for sensitivity, which is vital for clinical early warning systems. Different fusion rules correspond to different operating points along the precision–recall trade-off, and the choice of Max-Fusion reflects a deliberate bias towards sensitivity appropriate for early-warning settings rather than a claim of universal optimality. The fused model achieved a high recall of 0.871 and a PR-AUC of 0.836, indicating a robust ability to flag true drift days across diverse student behaviours.

The qualitative advantage of this fusion is visualised in Figure 10, where the aggregate drift score provides a smoothed, interpretable trajectory. The fused score consistently rises during periods of high disruption, even when individual modalities provide fluctuating signals. This provides a robust individual-level “wellbeing trajectory” that accounts for the weak association between sensing streams while leveraging their collective predictive power.

The global effectiveness of the fused approach is summarised in Table 7, which aggregates the longitudinal performance across the entire test cohort. The system achieved a recall of 0.871, a critical result for behavioural health applications where the cost of a “false negative” (failing to detect a burgeoning crisis) significantly outweighs the cost of a false alarm. This high sensitivity ensures that the “Max-Fusion” logic successfully captures drift episodes, even when they are manifest in only a single behavioural domain. While the accuracy (0.700) and precision (0.630) reflect the inherent noise and “class imbalance” of real-world sensing data, the ranking metrics provide a more robust picture of the model’s reliability. The ROC-AUC of 0.831 and PR-AUC of 0.836 indicate that the model maintains a strong discriminatory boundary between stable and drifting states. Specifically, the high PR-AUC confirms that the model’s “Top-K” drift alerts are highly likely to be true positives, making it a viable tool for prioritising which students might require a wellness check-in from university support services.

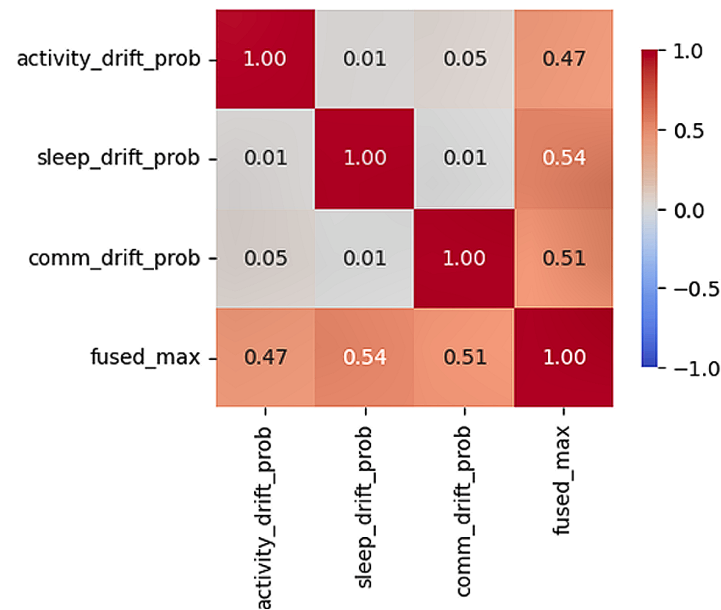


Figure 9. Late Fusion Confusion Matrix, demonstrating high sensitivity to behavioural drift events.

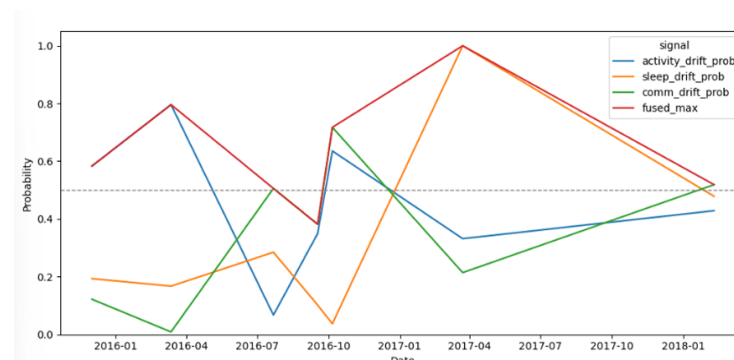


Figure 10. Longitudinal trajectory of the multimodal fused drift score for a representative participant. The fused signal (red line) integrates independent evidence from activity, sleep, and communication streams to identify sustained drift episodes relative to the 0.5 decision threshold.

Table 7. Global performance metrics for the decision level Late Fusion.

Metric	Accuracy	Precision	Recall	ROC-AUC	PR-AUC
Value	0.700	0.630	0.871	0.831	0.836

Ultimately, these metrics demonstrate that by leveraging the weakly correlated nature of activity, sleep, and social data ($r \approx 0.01$), the approach achieves a level of robustness that single-modality systems cannot. The high recall suggests that the “drift-streak” logic, combined with LSTM temporal memory, effectively filters out the stochastic “noise” of student life while remaining highly attuned to the sustained behavioural decay that characterises shifts in mental wellbeing.

5. Discussion

Although this study focuses on a student population, the proposed approach does not rely on population-level norms or demographic-specific features. Instead, it models behavioural drift relative to personalised, rolling baselines constructed on a per-user basis. As such, the same approach can be applied to other age groups or clinical cohorts, provided sufficient longitudinal passive sensing data are available. Because behavioural

drift is assessed relative to individual historical patterns rather than fixed schedules, it is expected to accommodate diverse life rhythms; however, empirical validation beyond student populations remains essential. The results demonstrate that adaptive, user-specific baselines, when integrated with deep temporal modelling and decision-level fusion, can robustly identify both transient anomalies and sustained behavioural shifts without the need for labelled outcomes. These findings align with and extend the current consensus in digital phenotyping, which emphasises that individualised reference frames and temporal continuity are paramount for meaningful behavioural monitoring in ecological settings.

5.1. Personalised Baselines and Temporal Sensitivity

A central pillar of the proposed approach is the hypothesis that behavioural change must be evaluated relative to an individual's historical patterns rather than population-level norms. By operationalising this through rolling statistical windows, we allowed behavioural baselines to update dynamically. The stability of drift detection across multiple window lengths ($\omega \in \{7, 14, 21, 28\}$) suggests a robust methodology that is not overly sensitive to specific temporal configurations. This consistency supports prior research indicating that medium-term baselines are sufficient to capture meaningful deviations in student populations without the overhead of exhaustive labelling [26,28]. Furthermore, the identification of contiguous “drift streaks” represents a significant methodological shift from standard anomaly detection. Visual inspection of these trajectories, for example, in Figure 8, indicates that the approach successfully separates isolated stochastic noise common in a university environment from prolonged episodes of behavioural decay. This distinction is critical, as sustained clusters of anomalies have been shown to be far more predictive of clinical deterioration than single-day spikes [29].

5.2. The Value of Sequential Memory in Deep Modelling

Our comparison between static and sequence-based models confirmed the hypothesis that the latter are superior for capturing the gradual, cumulative nature of behavioural change. While the Isolation Forest provided a conservative and stable statistical baseline, the 1D-CNN and LSTM architectures consistently improved recall and ranking performance (PR-AUC) across all modalities. The superior performance of LSTMs specifically supports the observation that behavioural “drift” rarely manifests as an abrupt event; instead, it unfolds as a multi-day decay of routine. This finding is particularly relevant for preventative mental health monitoring, where the ability to detect subtle, long-range temporal dependencies in activity or sleep may allow for intervention before a behavioural shift reaches a crisis threshold. The results reinforce the growing consensus in the field that temporal continuity is not merely a feature but a requirement for high-fidelity behaviour change detection [12,13].

5.3. Modality Independence and the Rationale for Max-Fusion

The investigation into multimodal integration revealed a striking lack of correlation between drift signals across activity, sleep, and communication ($r \approx 0.01$). This statistical independence is one of the most compelling findings of the study, indicating that each modality serves as a distinct, non-redundant sensor for specific behavioural phenotypes. Such modality independence has been suggested in prior literature, in which different physiological and social streams respond to stressors in complementary, rather than synchronous, ways [31,52]. Given this lack of cross-modal association, our choice of decision-level Max-Fusion proved technically and clinically sound. By prioritising the strongest evidence of drift from any single modality, we achieved a high recall (0.871), ensuring that significant disengagement in one domain is not suppressed by stability in others. This approach provides a clear path to interpretability; the fused trajectory (Figure 10)

allows practitioners to trace a global “wellbeing alert” back to its source modality, be it a collapse in social entropy or a shift in physical MVPA, enabling targeted rather than generalised interventions.

5.4. Contextualising Performance in Real-World Deployment

While the proposed work demonstrates a strong ranking performance of PR-AUC of 0.836, the precision (0.630) acknowledges the inherent difficulty of differentiating between pathological drift and benign lifestyle shifts. Inspection of modality-specific errors provides insight into these performance differences. Sleep achieves the strongest performance due to its circadian regularity and relatively low day-to-day variance; false positives in this modality often correspond to short-term schedule disruptions such as examination periods, travel, or late-night deadlines that do not persist as sustained behavioural drift, while false negatives tend to occur when gradual shifts in sleep timing remain within adaptive baseline thresholds. In contrast, communication behaviour is inherently sparse and episodic, leading to greater noise. False positives frequently arise from brief bursts of social activity, whereas false negatives occur when disengagement unfolds slowly or is masked by an already low baseline of interaction. Activity-related errors typically lie between these extremes, with transient workload or exercise changes occasionally producing false alarms. In practical deployments, alert thresholds can be adaptively calibrated to individual risk tolerance and contextual constraints, enabling precision to be increased once candidate drift periods have been prioritised. However, in the context of university counselling and early warning systems, the cost of a false negative far outweighs that of a false positive. By providing a consistent and reliable ordering of days by drift likelihood, this approach serves as a viable decision-support tool, helping wellness services prioritise students whose behavioural signatures deviate most significantly from their established norms.

5.5. Limitations

This study presented a methodological exploration of personalised, multimodal drift detection using multimodal wearable and smartphone data. However, several limitations must be acknowledged. A key limitation of this study is the absence of clinical or psychological ground-truth labels, such as validated mental health assessments. Consequently, detected behavioural drift cannot be interpreted as evidence of mental health decline. Accordingly, reported performance metrics should be interpreted as consistent with behavioural deviation heuristics rather than validation against clinical outcomes. The proposed approach is not intended as a diagnostic tool but rather as an early warning mechanism that identifies sustained statistical deviations in passively sensed behaviour that may warrant further attention. The evaluation relied on proxy drift labels generated from rolling-window baselines rather than externally validated annotations. This is a common constraint in unsupervised behavioural sensing [26,28], and it allowed a systematic comparison of modelling approaches. However, it also introduces a degree of circularity, as performance metrics primarily reflect consistency with statistical heuristics rather than true behavioural or clinical outcomes. Second, the dataset was drawn from a single university student population, whose daily routines are structured by the academic calendar. As a result, it limits the generalisability of findings to other populations, such as older adults, working professionals, or patients in clinical care. Third, for reasons of comparability, all modalities were aggregated to the day level. This simplification enabled alignment across activity, sleep, and communication but removed finer-grained dynamics that may precede behavioural change. Finally, the models do not yet explain why change occurs. Moreover, the analysis was retrospective, meaning that questions of real-time scalability, robustness to missing data, and ethical deployment remain open.

6. Conclusions

This study has successfully demonstrated that a personalised, multimodal approach can effectively identify behavioural drift within a university student cohort using passive sensing data from wearables and smartphones. By shifting the focus from population-level anomalies to individual-level “drift-streaks,” we have shown that a temporal context modelled through user-specific rolling baselines and deep sequential architectures is essential for capturing the gradual decay of behavioural routines. Our experimental results, which compared static Isolation Forest baselines with deep 1D-CNN and LSTM architectures, highlight that sequence-based models are significantly more capable of modelling the cumulative nature of behavioural change. Specifically, the high sensitivity achieved by the late-fusion LSTM model, which reached a recall of 0.871, underscores the potential of this approach to act as a robust decision-support tool. This high performance is driven by the approach’s ability to leverage the striking lack of correlation between drift signals in activity, sleep, and communication, proving that behavioural disengagement is a sparse, cross-domain phenomenon that requires multimodal integration to be fully captured.

Ultimately, this research provides a scalable, unsupervised methodology for longitudinal monitoring in environments where labelled clinical outcomes are scarce. By adopting a “Max-Fusion” strategy, the system remains interpretable and actionable, allowing practitioners to trace global alerts back to specific behavioural domains. This provides a transparent “wellbeing trajectory” that respects individual baseline variations while maintaining the sensitivity required for early warning applications. As digital phenotyping moves towards real-time clinical integration, the “drift-streak” concept offers a promising bridge between raw sensor data and the identification of meaningful psychological or lifestyle shifts. While the proposed approach demonstrates scalable behavioural drift detection under realistic data constraints, its translation into real-time clinical workflows will require prospective feasibility studies and close collaboration with mental health services.

Future work will incorporate self-report measures and clinically validated assessments to enable quantitative validation of detected behavioural drift against meaningful behavioural or health-related outcomes. Although the proposed approach is designed to be population-agnostic, its generalisability should be evaluated using more diverse demographic groups and clinical cohorts with comparable passive sensing data.

There is also significant potential in adopting routine-aware modelling through process mining and conformance checking, which would allow for a more granular understanding of how daily event logs deviate from established structures. Furthermore, the development of multi-scale temporal models using finer-grained, sub-daily features could enhance the approach’s ability to detect micro-patterns that anticipate broader drift episodes. Technical advancements such as GRU-D architectures or attention-based fusion mechanisms should be explored to improve robustness against missing data and irregular sampling. Finally, ensuring the ethical deployment of these systems through federated learning and improved model interpretability will be essential for fostering trust among clinicians and end-users in sensitive health contexts.

Author Contributions: Conceptualisation, E.T.; methodology, E.T. and I.K.I.; software, E.T. and I.K.I.; validation, E.T., P.M., S.W.Y., A.L. and I.K.I.; formal analysis, E.T., P.M., S.W.Y., A.L. and I.K.I.; investigation, E.T., P.M., S.W.Y., A.L. and I.K.I.; writing—original draft preparation, E.T. and I.K.I.; writing, review and editing, E.T., P.M., S.W.Y., A.L. and I.K.I.; visualisation, E.T., P.M., S.W.Y., A.L. and I.K.I.; supervision, I.K.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: All the participating authors have signed and given informed consent statements.

Data Availability Statement: The dataset for this study is publicly available.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mendes, J.P.; Moura, I.R.; Van de Ven, P.; Viana, D.; Silva, F.J.; Coutinho, L.R.; Teixeira, S.; Rodrigues, J.J.; Teles, A.S. Sensing apps and public data sets for digital phenotyping of mental health: Systematic review. *J. Med. Internet Res.* **2022**, *24*, e28735. [CrossRef] [PubMed]
2. Montag, C.; Baumeister, H. Digital phenotyping and mobile sensing. In *Digital Phenotyping and Mobile Sensing*, 2nd ed.; Springer International Publishing: Cham, Switzerland, 2023.
3. Insel, T.R. Digital phenotyping: Technology for a new science of behavior. *JAMA* **2017**, *318*, 1215–1216. [CrossRef]
4. Statista. Smartphone Penetration Rate in the United Kingdom (UK). 2023. Available online: <https://www.statista.com> (accessed on 10 January 2025).
5. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognit. Lett.* **2019**, *119*, 3–11. [CrossRef]
6. Zeng, M.; Nguyen, L.T.; Yu, B.; Mengshoel, O.J.; Zhu, J.; Wu, P.; Zhang, J. Convolutional neural networks for human activity recognition using mobile sensors. In *Proceedings of the 6th International Conference on Mobile Computing, Applications and Services*; IEEE: Austin, TX, USA, 2014; pp. 197–205.
7. World Health Organization. *Depression and Other Common Mental Disorders: Global Health Estimates*; WHO: Geneva, Switzerland, 2023.
8. Office for National Statistics. Cost of Living and Depression in Adults, Great Britain. 2022. Available online: <https://www.ons.gov.uk> (accessed on 10 January 2025).
9. Lara, O.D.; Labrador, M.A. A survey on human activity recognition using wearable sensors. *IEEE Commun. Surv. Tutor.* **2012**, *15*, 1192–1209. [CrossRef]
10. Kwapisz, J.R.; Weiss, G.M.; Moore, S.A. Activity recognition using cell phone accelerometers. *ACM Sigkdd Explor. Newsl.* **2011**, *12*, 74–82. [CrossRef]
11. Reiss, A.; Stricker, D. Introducing a new benchmarked dataset for activity monitoring. In *Proceedings of the IEEE International Symposium on Wearable Computers*; IEEE: Newcastle, UK, 2012; pp. 108–109.
12. Mohr, D.C.; Zhang, M.; Schueller, S.M. Personal sensing: Understanding mental health using ubiquitous sensors. *Annu. Rev. Clin. Psychol.* **2017**, *13*, 23–47. [CrossRef]
13. Hammerla, N.Y.; Halloran, S.; Plötz, T. Deep, convolutional, and recurrent models for human activity recognition. *arXiv* **2016**, arXiv:1604.08880. [CrossRef]
14. Murad, A.; Pyun, J.Y. Deep recurrent neural networks for human activity recognition. *Sensors* **2017**, *17*, 2556. [CrossRef]
15. Guan, Y.; Plötz, T. Ensembles of deep LSTM learners for activity recognition using wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2017**, *1*, 1–28. [CrossRef]
16. Ihianle, I.K.; Nwajana, A.O.; Ebeunuwa, S.H.; Otuka, R.I.; Owa, K.; Orisatoki, M.O. A deep learning approach for human activities recognition from multimodal sensing devices. *IEEE Access* **2020**, *8*, 179028–179038. [CrossRef]
17. Ignatov, A. Real-time human activity recognition from accelerometer data using convolutional neural networks. *Appl. Soft Comput.* **2018**, *62*, 915–922. [CrossRef]
18. Greff, K.; Srivastava, R.K.; Koutník, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2222–2232. [CrossRef] [PubMed]
19. Ferrara, E. Large language models for wearable sensor-based human activity recognition and health monitoring. *Sensors* **2024**, *24*, 5045. [CrossRef] [PubMed]
20. Deng, S.; Hua, W.; Wang, B.; Wang, G.; Zhou, X. Few-shot human activity recognition on noisy wearable sensor data. In *Proceedings of the International Conference on Database Systems for Advanced Applications*; Springer: Cham, Switzerland, 2020; pp. 54–72.
21. Kheirkhanian, M.; Mehta, S.; Nath, M.; Wanigatunga, A.A.; Corbett, D.B.; Manini, T.M.; Ranka, S. A bag-of-words approach for assessing activities of daily living using wrist accelerometer data. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; IEEE: Kansas City, MO, USA, 2017; pp. 678–685.
22. Ihianle, I.K.; Naeem, U.; Tawil, A.R. Recognition of activities of daily living from topic model. *Procedia Comput. Sci.* **2016**, *98*, 24–31. [CrossRef]
23. Zheng, J.; Papapanagiotou, P.; Fleuriot, J.D.; Hillston, J. Process-aware human activity recognition. *arXiv* **2024**, arXiv:2411.08814. [CrossRef]

24. Winter, M.; Langguth, B.; Schlee, W.; Pryss, R. Process mining in mHealth data analysis. *NPJ Digit. Med.* **2024**, *7*, 299. [CrossRef]
25. Prenkaj, B.; Velardi, P. Unsupervised detection of behavioural drifts with dynamic clustering and trajectory analysis. *IEEE Trans. Knowl. Data Eng.* **2023**, *36*, 2257–2270. [CrossRef]
26. Sprint, G.; Cook, D.J.; Schmitter-Edgecombe, M. Unsupervised detection of changes in everyday physical activity. *J. Biomed. Inform.* **2016**, *63*, 54–65. [CrossRef]
27. Dawadi, P.N.; Cook, D.J.; Schmitter-Edgecombe, M. Modeling patterns of activities using activity curves. *Pervasive Mob. Comput.* **2016**, *28*, 51–68. [CrossRef]
28. Diaz, C.; Caillaud, C.; Yacef, K. Unsupervised early detection of physical activity behaviour changes. *Sensors* **2022**, *22*, 8255. [CrossRef]
29. Barnett, I.; Torous, J.; Staples, P.; Sandoval, L.; Keshavan, M.; Onnela, J.-P. Relapse prediction in schizophrenia through digital phenotyping. *Neuropsychopharmacology* **2018**, *43*, 1660–1666. [CrossRef] [PubMed]
30. Berrouiguet, S.; Perez-Rodriguez, M.M.; Larsen, M.; Baca-García, E.; Courtet, P.; Oquendo, M. From eHealth to iHealth: Transition to participatory and personalized medicine in mental health. *J. Med. Internet Res.* **2018**, *20*, e2. [CrossRef] [PubMed]
31. Saeb, S.; Zhang, M.; Karr, C.J.; Schueller, S.M.; Corden, M.E.; Kording, K.P.; Mohr, D.C. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *J. Med. Internet Res.* **2015**, *17*, e4273. [CrossRef] [PubMed]
32. Chen, S.; Sun, X. Validating CircaCP: A generic sleep–wake cycle detection algorithm for unlabelled actigraphy data. *R. Soc. Open Sci.* **2024**, *11*, 231468. [CrossRef]
33. Minaeva, O.; Riese, H.; Booi, S.H.; Lamers, F.; Giltay, E.J.; Scheer, F.A.J.L.; Hu, K. Fractal motor activity during wakefulness and sleep: A window into depression recency and symptom recurrence. *Psychol. Med.* **2024**, *54*, 4429–4437. [CrossRef]
34. Moura, I.; Teles, A.; Viana, D.; Marques, J.; Coutinho, L.; Silva, F. Digital phenotyping of mental health using multimodal sensing of multiple situations of interest: A systematic literature review. *J. Biomed. Inform.* **2023**, *138*, 104278. [CrossRef]
35. Abdullah, S.; Matthews, M.; Frank, E.; Doherty, G.; Gay, G.; Choudhury, T. Automatic detection of social rhythms in bipolar disorder. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 538–543. [CrossRef]
36. Harari, G.M.; Lane, N.D.; Wang, R.; Crosier, B.S.; Campbell, A.T.; Gosling, S.D. Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspect. Psychol. Sci.* **2016**, *11*, 838–854. [CrossRef]
37. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
38. Eagle, N.; Pentland, A.; Lazer, D. Inferring friendship network structure by using mobile phone data. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 15274–15278. [CrossRef]
39. Blair, D.S.; Miller, R.L.; Calhoun, V.D. A dynamic entropy approach reveals reduced functional network connectivity trajectory complexity in schizophrenia. *Entropy* **2024**, *26*, 545. [CrossRef]
40. Vaizman, Y.; Weibel, N.; Lanckriet, G. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *1*, 122. [CrossRef]
41. Khoo, L.S.; Lim, M.K.; Chong, C.Y.; McNaney, R. Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches. *Sensors* **2024**, *24*, 348. [CrossRef] [PubMed]
42. Vos, G.; Trinh, K.; Sarnyai, Z.; Azghadi, M.R. Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review. *Int. J. Med. Inform.* **2023**, *173*, 105026. [CrossRef] [PubMed]
43. Shen, S.; Qi, W.; Zeng, J.; Li, S.; Liu, X.; Zhu, X.; Dong, C.; Wang, B.; Shi, Y.; Yao, J.; et al. Passive sensing for mental health monitoring using machine learning with wearables and smartphones: A scoping review. *J. Med. Internet Res.* **2025**, *27*, e77066. [CrossRef]
44. Yu, L.; Ge, Y.; Ansari, S.; Imran, M.; Ahmad, W. Multimodal sensing-enabled large language models for automated emotional regulation: A review of current technologies, opportunities, and challenges. *Sensors* **2025**, *25*, 4763. [CrossRef]
45. University of Notre Dame. NetHealth Study Dataset and Documentation, 2015–2019. Available online: <https://sites.nd.edu/nethealth/> (accessed on 10 January 2025).
46. Faust, L.; Purta, R.; Hachen, D.; Striegel, A.; Poellabauer, C.; Lizardo, O.; Chawla, N.V. Exploring Compliance: Observations from a Large-Scale Fitbit Study. In *SocialSen'17: Proceedings of the 2nd International Workshop on Social Sensing, Pittsburgh, PA, USA, 18–21 April 2017*; ACM: New York, NY, USA, 2017; pp. 55–60.
47. Migueles, J.H.; Cadenas-Sanchez, C.; Ekelund, U.; Delisle Nyström, C.; Mora-Gonzalez, J.; Löf, M.; Labayen, I.; Ruiz, J.R.; Ortega, F.B. Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. *Sports Med.* **2017**, *47*, 1821–1845. [CrossRef]
48. Case, M.A.; Burwick, H.A.; Volpp, K.G.; Patel, M.S. Accuracy of smartphone applications and wearable devices for tracking physical activity data. *JAMA* **2015**, *313*, 625–626. [CrossRef]
49. Wang, C.; Lizardo, O.; Hachen, D.S. Using Fitbit data to examine factors that affect daily activity levels of college students. *PLoS ONE* **2021**, *16*, e0244747. [CrossRef]

50. Currey, S.; Doherty, K.; Doherty, G. Missing data in digital phenotyping: A review of reporting and handling methods. *J. Am. Med. Inform. Assoc.* **2023**, *30*, 1486–1496.
51. Alzahrani, H. Dose–response association between physical activity and health-related quality of life in the general population. *Healthcare* **2022**, *10*, 1460. [[CrossRef](#)]
52. Wang, R.; Chen, F.; Chen, Z.; Li, T.; Harari, G.; Tignor, S.; Zhou, X.; Ben-Zeev, D.; Campbell, A.T. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*; ACM: Seattle, WA, USA, 2014; pp. 3–14.
53. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation forest. In *Proceedings of the IEEE International Conference on Data Mining*; IEEE: Pisa, Italy, 2008; pp. 413–422.
54. Tabassum, M.; Mahmood, S.; Bukhari, A.; Alshemaimri, B.; Daud, A.; Khalique, F. Anomaly-based threat detection in smart health using machine learning. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 347. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.