American Journal of Interdisciplinary Studies

Volume: 2; **Issue:** 2
Pages: 32-66
eISSN: 3067-5146

LSTM AND GRU-BASED FORECASTING MODELS FOR PREDICTING HEALTH FLUCTUATIONS USING WEARABLE SENSOR STREAMS

Jinnat Ara[1]; Md. Kamrul Khan[2];

[1]. *Master of Science in Applied Mathematics, Noakhali Science and Technology University, Bangladesh; Email: jinnataraprema51@gmail.com*

[2]. *M.Sc in Mathematics, Jagannath University, Dhaka; Bangladesh; Email: mdkamrul.msc@gmail.com*

Doi: 10.63125/1p8gbp15
Received: 19 April 2021; **Revised:** 21 May 2021; **Accepted:** 09 June 2021; **Published:** 28 June 2021

Abstract

This quantitative, cross-sectional, case-study–based research addresses the problem that short-horizon health fluctuations in wearable streams are difficult to operationalize and forecast, limiting actionable monitoring. The purpose was to define fluctuation outcomes, identify wearable indicators linked to perceived variability, and compare LSTM versus GRU forecasting pipelines under matched preprocessing and evaluation. The bounded case sample comprised 120 monitored participant cases retained after quality screening (mean wear-time 12.8 hours/day, SD 2.1; overall missingness 9.6%, SD 6.4; 84.2% high adherence at or above 10 hours/day). The primary subjective variable, Health Fluctuation Severity, was a 6-item 5-point Likert composite with acceptable reliability (Cronbach's alpha .86; mean 3.18, SD 0.74), and the primary objective variable was an Objective Fluctuation Index (mean 0.61, SD 0.19). The analysis plan combined descriptive statistics, reliability checks, Pearson correlations, and regression with data-quality controls, alongside forecasting evaluation using MAE and RMSE and a paired t-test on participant-level errors. Headline findings showed convergence between subjective and objective measures (r .52, p < .001) and significant associations with resting heart-rate shift (r .41), sleep disruption (r .46), activity instability (r .29), and HRV proxy (r −.38) (all p ≤ .001). GRU outperformed LSTM on the test set (MAE 0.072 vs 0.081; RMSE 0.094 vs 0.106; t(119) 4.62, p < .001) and increased explanatory power when combined with wearable features (R squared 0.41; delta R squared 0.07, p = .004). The results imply that baseline-centered indices plus GRU forecasting can support validated fluctuation monitoring in real-world programs when adherence and missingness are managed.

Keywords

Wearable Sensor Streams; Health Fluctuation Severity; Objective Fluctuation Index; GRU Forecasting; LSTM Comparison;

INTRODUCTION

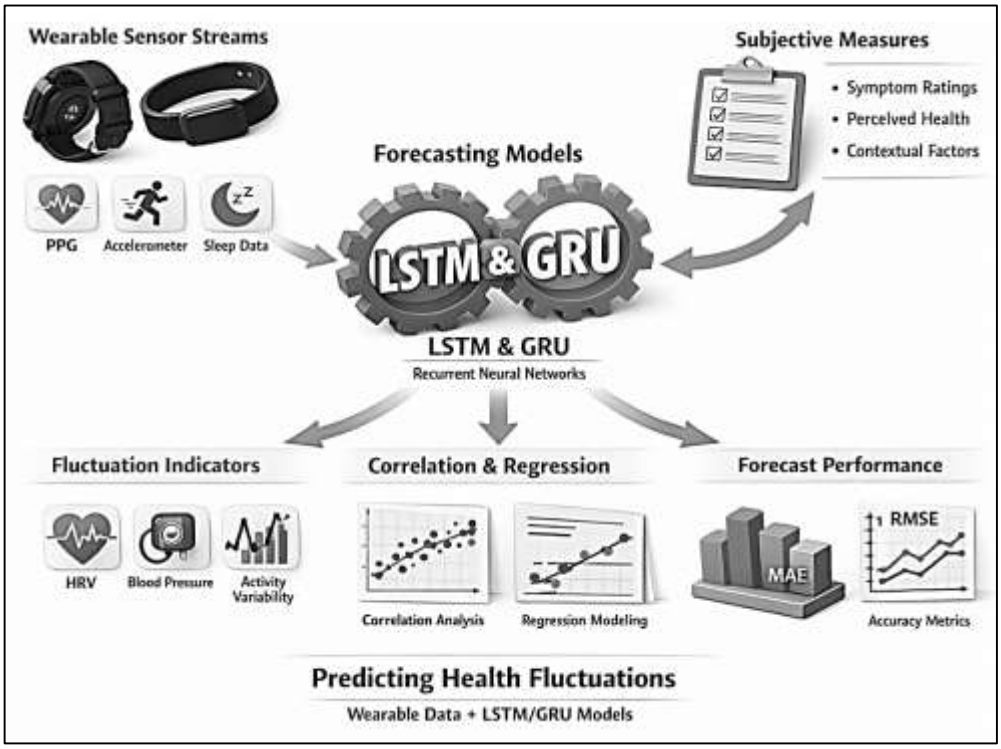
Health fluctuations can be defined as measurable, time-varying deviations in an individual's physiological and behavioral state that occur around a personal baseline across minutes, hours, days, or weeks. In applied health analytics, this definition treats "health" as an observable state reflected in multi-signal patterns (e.g., cardiovascular, sleep–wake, and activity rhythms), and it treats "fluctuation" as dynamic variability that can be quantified through short-term instability, longer-term drift, or recurring cycles in those patterns. This framing aligns with population-level evidence showing that chronic, non-communicable diseases (NCDs) contribute substantially to preventable morbidity and disability worldwide and that many NCD-related outcomes are preceded by long periods of subtle physiological change rather than a single abrupt transition (Beaglehole et al., 2011).

Global health metrics and definitions used to quantify health loss reinforce the value of capturing health as a continuum rather than an event, because burden estimates are built on systematic measurement of changing risk exposure, disease progression, and functional impairment across time (Murray et al., 2012). In parallel, comparative risk assessment work has emphasized that health loss is shaped by interacting risk factors whose influence can vary across regions and across the life course, which supports an analytic approach that can detect short-term and medium-term variations in physiological indicators tied to those risks (Lim et al., 2012). Ageing populations add another layer of international relevance because chronic disease, multi-morbidity, and functional decline often present as gradual alterations in cardiometabolic regulation, sleep quality, and daily activity capacity that unfold across years while expressing day-to-day variability (Prince et al., 2015). Within psychophysiology and behavioral medicine, heart rate variability (HRV) has been widely treated as a window into autonomic regulation and self-regulatory capacity, offering a concrete example of how "fluctuations" can be operationalized as dynamic changes in regulatory physiology (Thayer et al., 2012). Evidence linking HRV to stress and health through neural and autonomic pathways also illustrates that fluctuations are not noise to be ignored; they often carry signal about adaptation and load within the organism (Thayer et al., 2009). Taken together, international disease burden evidence and physiological regulation research motivate a definition of health fluctuations that is individualized, multi-signal, and explicitly temporal, creating a basis for forecasting approaches that treat continuous data streams as primary empirical material rather than secondary context.

Wearable sensor streams refer to continuous or semi-continuous sequences of measurements produced by body-worn devices that sample physiological or behavioral signals during ordinary living conditions. In biomedical engineering, this concept is closely tied to body sensor networks and ambulatory monitoring systems designed to move observation from episodic clinic encounters to routine environments, where longer observation windows can capture variability and context-linked change (Bonato, 2010). In consumer health and digital medicine, wearable streams have also become mainstream through wrist-worn and patch-based devices that record heart rate proxies, movement, and sleep-related features at population scale, which has expanded both research opportunity and methodological scrutiny (Piwek et al., 2016). Remote monitoring reviews describe how sensor miniaturization, wireless communication, and device integration have enabled long-duration data acquisition, supporting analytics that can examine intra-day rhythms and inter-day instability within the same individual (Majumder et al., 2017). At the same time, international deployment of wearables foregrounds measurement validity as a core scientific concern because devices differ in sensing modality, sampling rates, proprietary preprocessing, and susceptibility to motion and placement error (Shcherbina et al., 2017). Flexible and wearable device scholarship further clarifies that the "stream" is not only a data structure but also a product of materials, power constraints, skin–sensor coupling, and user interaction, all of which influence signal fidelity and missingness patterns during real life (Tamura et al., 2014). Internet-of-Things (IoT) architectures for healthcare have been presented as a complementary systems layer, connecting wearable sensing to distributed data pipelines, cloud storage, and analytics services across settings and geographies, which reinforces the relevance of scalable modeling strategies for high-volume sequential data (Islam et al., 2015). This systems view also aligns with health informatics perspectives that define modern health data as heterogeneous, longitudinal, and noisy, requiring models that can learn from sequences while accounting for device constraints, irregular sampling, and environmental disturbances (Miotto et al., 2018). Therefore,

wearable sensor streams can be defined operationally as multivariate time series generated in unconstrained contexts, where the scientific problem becomes learning reliable temporal structure—within-person and between-person—from data affected by sensor physics, user behavior, and computational preprocessing.

Figure 1: Framework for Predicting Health Fluctuations Using Wearable Sensor Streams



A major portion of wearable health monitoring relies on optical and inertial sensing, with photoplethysmography (PPG) and accelerometry being among the most common modalities for continuous, wrist-based measurement. PPG is an optical technique used to detect blood volume changes in tissue, producing a waveform that contains a pulsatile component synchronized to cardiac cycles and slower components linked to respiration, autonomic activity, and thermoregulation (Allen, 2007). Reviews of wearable PPG sensors emphasize that real-world measurements depend strongly on reflective sensing geometry, wavelength selection, and tissue–device interaction, which can alter waveform morphology and complicate physiological interpretation across contexts (Ballinger et al., 2018). Motion artifacts are a central technical issue because everyday movement can generate distortions whose amplitude overlaps with the frequency range of heart rate, making continuous monitoring an algorithmic problem as much as a hardware problem (Zhang et al., 2019). Signal-processing research has proposed motion-tolerant extraction strategies that treat movement as a contaminant to be estimated and removed, including adaptive approaches designed for wearable PPG biosensors (Venkatesh et al., 2012) and methods that evaluate robustness during physical exercise using wrist-type PPG (Shickel et al., 2018). These signal-quality challenges directly influence the definition of "health fluctuations," because fluctuations inferred from streams can reflect physiological change, artifact-driven distortion, or both; separating these sources becomes essential for any forecasting claim grounded in wearable time series. Beyond heart rate, wearable-derived PPG has been used for blood pressure estimation, illustrating how the same raw modality can be engineered into higher-level health indicators through machine learning. Comparative evaluations of PPG-only blood pressure estimation have demonstrated that model choice and feature representation shape estimation error and clinical-category performance (LeCun et al., 2015), while deep neural approaches have further explored spectro-temporal representations to map PPG sequences to blood pressure values (Slapničar et al., 2019). Sleep-related streams introduce additional complexity because wearable sleep staging and sleep-

quality metrics combine movement and cardio-respiratory signals with device-specific scoring logic, requiring validation against reference standards to interpret fluctuations as meaningful physiological variation (de Zambotti et al., 2018). Collectively, these studies show that wearable streams encode rich temporal information about cardiovascular and behavioral regulation, and they also show that accurate fluctuation modeling depends on explicit handling of sensing modality properties, artifact structure, and validation evidence.

From a modeling standpoint, wearable sensor streams present a multivariate sequence-learning problem characterized by missing values, irregular sampling, non-stationarity, and person-specific baselines. These properties are not incidental; they emerge from user adherence patterns, battery cycles, data transmission gaps, and algorithmic filtering performed on-device or in companion applications (Che et al., 2018). Missingness is especially consequential in health forecasting because data gaps can align with behavior (e.g., device removal during discomfort or charging), which can bias learned temporal relationships if treated as random loss. Sequence modeling research has addressed this directly by developing recurrent approaches that incorporate missingness patterns and time gaps into the learning process, enabling models to represent not only observed values but also the structure of absence in clinical and physiological time series (Fallet & Vesin, 2017). At the same time, the scientific target in forecasting health fluctuations is not simply point prediction; it often requires predicting trajectories, anomalies, or imminent deviations from baseline across multiple signals. This motivates careful attention to evaluation metrics, because performance interpretation depends on whether the goal is minimizing absolute error, scaling error relative to magnitude, or capturing turning points and variability patterns. Forecasting methodology has long emphasized that accuracy measures are not interchangeable and that metric choice can change model ranking and scientific conclusions, especially in time series with changing variance and occasional spikes (Fawaz et al., 2019). Deep learning reviews on time-series tasks further highlight that representation learning in sequences can outperform hand-crafted features under certain conditions, while also warning that training practices, dataset characteristics, and validation protocols drive much of the apparent advantage (Greff et al., 2017). In wearable settings, this implies that forecasting health fluctuations requires an integrated approach: preprocessing that respects sensor physics and adherence realities; modeling that can ingest multivariate sequences with gaps; and evaluation that reflects the specific meaning of "fluctuation" in the operational definition. This modeling lens also links wearable forecasting to broader health informatics discussions, where the primary analytic challenge is extracting clinically and behaviorally relevant temporal structure from noisy longitudinal data rather than from static snapshots (Hammerla et al., 2016). Accordingly, the methodological foundation for forecasting in wearable streams is a combination of time-series evaluation principles, missingness-aware learning, and domain-aware signal interpretation.

Long short-term memory (LSTM) and gated recurrent unit (GRU) architectures are widely used recurrent neural network (RNN) variants designed to learn dependencies in sequential data through gating mechanisms that regulate information flow across time. Within the deep learning literature, gating is treated as a practical solution to training difficulties in long sequences, enabling models to retain relevant context while suppressing irrelevant or redundant information (Lara & Labrador, 2013). For wearable sensor streams, the relevance of LSTM/GRU families is grounded in the need to model multi-scale temporal structure: rapid physiological oscillations (seconds), activity transitions (minutes), circadian variation (hours), and multi-day drift tied to behavior and recovery. Human activity recognition (HAR) research provides an important empirical bridge between wearable sensing and recurrent modeling, because HAR tasks rely on learning temporal signatures in inertial and physiological sequences recorded in unconstrained contexts. Survey work in wearable HAR has documented the diversity of sensor placements, feature pipelines, and learning objectives, highlighting why end-to-end sequence models became attractive as datasets expanded and tasks moved toward naturalistic settings (Khalid et al., 2018). Studies evaluating deep recurrent and convolutional-recurrent approaches for activity recognition using wearables show that recurrent components help capture temporal continuity and transitions that are difficult to encode with windowed features alone (Schmidt et al., 2018). Stress and affect detection research strengthens this connection by introducing multimodal wearable datasets that combine physiological and motion channels, creating a setting where

fluctuations represent both autonomic regulation and behavioral context (Shcherbina et al., 2017). Within cardiovascular monitoring, sequence learning has been used to infer risk or detect arrhythmias from long recordings, providing additional evidence that temporal deep models can extract meaningful structure from extended physiological sequences (Arfan et al., 2021; Yousefi et al., 2014). These lines of research suggest a coherent methodological rationale for an LSTM- and GRU-based forecasting study: the architectures are aligned with the temporal nature of wearable streams, their gating supports learning across variable-length dependencies, and prior wearable and clinical sequence studies provide validated precedents for training and evaluating recurrent models under real-world signal constraints. The concept of predicting health fluctuations using LSTM- and GRU-based forecasting models becomes most concrete when it is tied to explicit operational indicators and to a study design that can integrate objective streams with subjective assessments. Wearable measurement research has shown that device outputs can vary in accuracy across activities and individuals, which means that modeling must be paired with attention to measurement error and context (Jahid, 2021; Shcherbina et al., 2017). Sleep validation work similarly demonstrates that wearable-derived sleep metrics can align with reference measures under some conditions while diverging under others, which reinforces the need to define which fluctuations are treated as primary outcomes and how they are cross-validated (Ha et al., 2018). In cardiovascular sensing, algorithmic handling of motion artifacts and modality-specific distortions has been shown to substantially alter the quality of derived indicators such as heart rate, influencing any downstream definition of instability or deviation (Hannun et al., 2019; Md.Akbar & Farzana, 2021). Blood pressure estimation research using PPG further illustrates that physiological fluctuation targets can be framed as continuous variables with clinically meaningful ranges, and that machine learning pipelines must specify feature representations and validation procedures to interpret model outputs (Hyndman & Koehler, 2006). For theoretical grounding of fluctuation meaning, the neurovisceral integration perspective offers a well-cited framework linking autonomic variability, prefrontal regulation, and adaptive functioning, providing a basis for treating HRV-related dynamics as part of health fluctuation constructs (Piwek et al., 2016; Reza et al., 2021). For technology-mediated data capture and participant interaction with devices, information-systems theory such as UTAUT2 offers a complementary theoretical lens on acceptance and use patterns that can shape adherence, missingness, and measurement continuity in cross-sectional case-study settings (Venkatesh et al., 2012). These foundations support an introduction-level rationale for a quantitative, cross-sectional, case-study–based design that pairs wearable time series with Likert-scale measures: wearable streams provide high-frequency objective indicators of physiological and behavioral variation, while structured self-report can capture perceived health states, symptom variability, or contextual factors aligned with the operational definition of health fluctuations. Methodologically, correlation analysis and regression modeling fit naturally into this framing as tools to quantify relationships among engineered fluctuation indices, survey constructs, and model-derived forecasts, while forecasting evaluation metrics provide the performance lens for comparing LSTM- and GRU-based models on the same temporal prediction targets (Hyndman & Koehler, 2006; Zobayer, 2021a).

The present study is designed around a set of clear, objective-driven goals that translate the broad problem of forecasting health fluctuations into measurable analytical tasks using wearable sensor streams and deep sequence models. First, the study aims to define "health fluctuations" in a way that is empirically testable by specifying how short-term instability and deviation from individual baselines will be represented using both wearable-derived indicators and structured participant ratings captured through a five-point Likert scale. This objective requires establishing an outcome definition that is consistent across participants while still reflecting personal physiological differences, ensuring that the dependent measure is suitable for quantitative analysis. Second, the study seeks to identify and organize the most informative wearable signals and derived features for representing fluctuation patterns, focusing on indicators such as heart-rate dynamics, sleep-related variability, and activity stability, and transforming raw time-stamped streams into standardized multivariate sequences that can be used for model input. Third, the study aims to develop two forecasting pipelines using LSTM and GRU architectures under equivalent training conditions, so that model performance differences can be attributed to architectural properties rather than to inconsistent preprocessing or tuning. This includes setting comparable sequence windows, forecasting horizons, and optimization rules, and

establishing a reproducible model development workflow. Fourth, the study intends to evaluate and compare the forecasting accuracy of LSTM and GRU models using appropriate prediction metrics, and to document performance across participants and fluctuation levels to ensure that model outputs remain interpretable in real monitoring contexts. Fifth, the study aims to statistically validate the relationship between wearable-derived features, model-generated fluctuation predictions, and self-reported fluctuation severity by applying descriptive statistics to summarize the dataset, correlation analysis to quantify associations among variables, and regression modeling to test whether forecast outputs explain meaningful variance in the reported fluctuation outcome. Finally, the study seeks to integrate these objectives into a case-study–based cross-sectional design that supports a coherent set of research questions and hypotheses, enabling the research to present model comparison results alongside empirical statistical evidence that links predicted fluctuations to measurable signals and participant-reported experiences within a defined observation window.

LITERATURE REVIEW

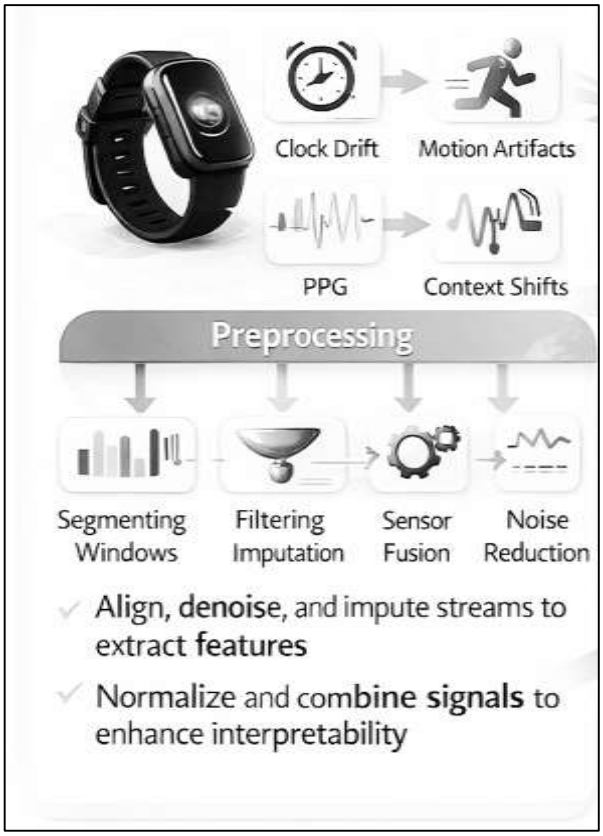
The literature on forecasting health fluctuations from wearable sensor streams sits at the intersection of ambulatory physiology, digital health measurement, and sequence-based machine learning, with a shared emphasis on representing human health as a dynamic state rather than a static condition. Wearable devices generate multivariate time series that capture cardiovascular activity, movement, and sleep-related patterns in naturalistic settings, enabling researchers to quantify short-term variability and longer-term drift around individual baselines. Within this body of work, "health fluctuations" are commonly operationalized as changes in autonomic regulation, sleep–wake stability, and activity regularity that reflect transient stress load, fatigue accumulation, recovery dynamics, or symptom variability, and these constructs are measured using both objective indicators (e.g., heart rate patterns, variability proxies, and sleep disruption metrics) and subjective ratings that summarize perceived changes in wellbeing. Studies in wearable sensing and remote monitoring highlight that real-world streams are affected by motion artifacts, device placement differences, sampling irregularities, and missingness driven by adherence behaviors, which directly shapes feature reliability and the statistical validity of inferences drawn from derived indicators. As a result, the literature places strong emphasis on preprocessing pipelines, quality assessment, and normalization strategies that preserve person-specific baselines while improving comparability across participants. In parallel, time-series forecasting research has increasingly adopted deep learning approaches that learn temporal representations directly from sequences, with LSTM and GRU architectures becoming widely used due to their gated mechanisms for capturing dependencies across time and handling nonlinearity in physiological signals. Comparative research has shown that model performance depends on the alignment between forecasting horizon, window length, and the physiological time scale of the target outcome, and it also depends on evaluation choices that define what constitutes an accurate prediction of variability. Beyond model accuracy, empirical health studies frequently incorporate descriptive statistics, correlation analysis, and regression modeling to link wearable-derived measures and model outputs with reported health states, supporting hypothesis testing in cross-sectional designs where survey-based measures provide complementary insight into perceived fluctuation severity. The emerging consensus across these strands is that robust forecasting of health fluctuations requires an integrated approach: careful definition of fluctuation outcomes, reliable transformation of raw wearable streams into meaningful features or representations, fair model comparison under consistent training conditions, and statistical validation that connects predictions to interpretable variables within the study context.

Wearable Sensor Streams and Continuous Health Monitoring

Wearable sensor streams can be understood as continuously generated, time-stamped physiological and behavioral measurements captured by body-worn devices and communicated to a computation or storage endpoint for analysis. In health monitoring, these streams commonly include cardiovascular surrogates (for example beat-to-beat timing inferred from photoplethysmography), movement traces from inertial measurement units, and contextual indicators such as skin temperature or electrodermal activity. Because each channel is sampled at its own cadence and may be interrupted by daily-life events, the resulting dataset is inherently multiscale and intermittently observed rather than a single clean signal. From a systems perspective, wearable health monitoring is frequently implemented as a

wireless body sensor network in which multiple on-body nodes collect biosignals and transmit them to a central node (often a phone) that performs buffering, synchronization, and forwarding to remote storage or analytics. This architecture matters for forecasting tasks because choices about sampling rate, on-device preprocessing, latency, and packet loss shape the temporal fidelity of the sequence that a model ultimately learns from. In addition, health monitoring streams must balance ergonomic constraints (comfort, battery life, and unobtrusiveness) with clinical constraints (signal integrity, calibration stability, and secure transport of personal medical data). Canonical overviews of body sensor networks emphasize that communication protocols, power management, and sensor fusion are not peripheral concerns but core determinants of whether long-duration physiological monitoring is feasible outside the clinic (Hao & Foster, 2008; Zobayer, 2021b). Similarly, surveys of wearable health-monitoring systems describe recurring design tradeoffs among multiparameter sensing, real-time decision support, and reliability requirements that directly influence the completeness and usability of collected time-series data (Pantelopoulos & Bourbakis, 2010). Taken together, these system-level foundations explain why "wearable streams" are best treated as structured sequences produced by an end-to-end pipeline, rather than as isolated measurements, when the research objective is to predict short-term health fluctuations with confidence.

Figure 2: Wearable Sensor Streams and Continuous Health Monitoring



Within health monitoring research, wearable streams are rarely analyzed in raw form because free-living recordings contain motion artifacts, sensor dropout, clock drift, and behavioral context shifts that can mimic physiological change. As a result, the literature treats preprocessing and representation as foundational steps for converting heterogeneous sensor outputs into interpretable time series that support statistical analysis and machine learning. Common procedures include segmenting continuous streams into windows, aligning channels to a shared timeline, filtering noise, imputing short gaps, and normalizing values to person-specific baselines so that intra-individual changes are emphasized over between-person level differences. Health monitoring systems also combine sensors to capture complementary aspects of function, such as pairing activity traces with cardiovascular markers to interpret whether elevated heart rate reflects exertion, stress, or anomalous physiology. Application-

driven work shows that these design choices are shaped by the clinical question, including safety monitoring, chronic disease self-management, rehabilitation tracking, and early detection of deteriorating status. In rehabilitation-oriented monitoring, the unit of analysis often shifts from isolated events to trajectories of recovery, where time series features such as variability, trend, and periodicity represent adherence, functional capacity, and symptom dynamics. Accordingly, the data pipeline must preserve meaningful temporal structure while remaining robust to irregular participation and sensor nonwear, which otherwise introduce systematic missingness related to the very outcomes under study. A widely cited synthesis of wearable sensing in rehabilitation underscores how remote monitoring systems rely on multi-sensor networks and tailored analytics to translate continuous streams into clinically relevant indicators that can be evaluated outside traditional clinical settings (Patel et al., 2012). These lessons motivate forecasting research to prioritize transparent preprocessing and stable feature definitions so that predictive performance can be interpreted as evidence about health fluctuation patterns rather than as artifacts of data handling across participants, devices, and day-to-day contexts in real-world settings consistently.

A third strand of the wearable health monitoring literature concentrates on measurement validity and reliability, because forecasting health fluctuations is only as credible as the sensor signals and derived metrics used to represent physiological state. Optical heart-rate sensing, step estimation, and sleep proxies are scalable, yet accuracy varies with motion, skin contact, exertion intensity, device firmware, and user physiology. Validation studies therefore set practical limits for analytics by indicating when a stream can be treated as a quantitative variable and when it should be treated as noisy evidence. In controlled treadmill testing that compared consumer wrist devices against electrocardiogram, agreement differed across brands and generally worsened as exertion increased, showing how activity can introduce systematic error into heart-rate time series (Wang et al., 2017). Such device-dependent error can distort fluctuation labels, inflate apparent variability, and weaken cross-participant comparability unless quality checks and calibration rules are applied. Complementary evidence from a systematic review of Fitbit measurement accuracy shows that performance differs by outcome and setting, with step-count accuracy often shifting between controlled tests and free-living use (Feehan et al., 2018). This matters for LSTM and GRU pipelines because deep sequence models can learn stable physiology and stable bias patterns alike when they are not separated analytically. Accordingly, the literature supports reporting adherence and missingness, and testing whether forecast performance is stable across preprocessing choices, time windows, and device subsets. In practice, researchers may treat accuracy findings as guidance for selecting sensors, setting inclusion thresholds, or weighting channels during fusion, so that downstream regression models relate predictions to survey responses without confounding from inconsistent device behavior across cases. When validity evidence, data processing discipline, and model evaluation are aligned, wearable streams become a defensible substrate for connecting predicted fluctuations to descriptive statistics, correlations, and regression-based hypothesis tests in a cross-sectional case-study design.

Health Fluctuation Measurement and Labeling Approaches

Health fluctuation measurement begins with a precise definition of what “fluctuation” means in the target population and how that meaning can be captured with repeatable indicators. In wearable-based studies, fluctuations are commonly framed as within-person departures from an established baseline, because individuals can differ in resting physiology while still showing comparable patterns of instability. Subjective measurement is often used to summarize perceived change in wellbeing, fatigue, stress, or symptom intensity using brief rating items on a five-point Likert scale. Guidance on the analysis of Likert-type ratings supports the use of summary statistics and common parametric techniques when items are designed and interpreted appropriately, which makes them compatible with descriptive statistics, correlation analysis, and regression modeling in quantitative designs (Carifio & Perla, 2008). Because global retrospective reports can blur moment-to-moment variation, many studies pair Likert ratings with repeated, in-the-moment sampling strategies so that fluctuation labels reflect current state rather than distant recall. Ecological momentary assessment operationalizes this idea by collecting multiple observations per person across days and contexts, enabling researchers to quantify variability, identify triggers, and align subjective states with objective signals during the same time window (Shiffman et al., 2008). In a wearable forecasting context, subjective labels can be defined as

time-stamped ratings (for example, a daily "health change" score) or as aggregated summaries over a cross-sectional observation period (for example, an average fluctuation severity score). The key measurement requirement is that the subjective outcome is anchored to a clear timeframe and response scale so that it can be mapped to wearable windows for model training and for hypothesis testing. When multiple items are used to represent one construct, internal consistency testing supports the creation of composite scores, and item wording should focus on change, intensity, or stability to match the intended fluctuation definition within the same day or week.

Figure 3: Measurement and Labeling Approaches in Wearable-Based Monitoring

Objective labeling approaches define health fluctuations directly from wearable streams by treating instability as a detectable structure in the multivariate time series. A simple strategy constructs continuous fluctuation indices using rolling-window statistics such as variance, coefficient of variation, range, or absolute deviation from a personal baseline, producing outcomes that can be forecast as real-valued targets. This approach supports individualized interpretation because the baseline is estimated per participant and the index emphasizes relative change rather than population level. A second family of methods treats fluctuations as discrete events, labeling time segments that represent abnormal departures from typical behavior. In this event-based view, fluctuation labeling resembles anomaly detection: the goal is to learn or estimate what "normal" looks like for a person and flag segments that violate that model. Survey work on anomaly detection emphasizes that definitions of normality, feature representation, and the assumed rarity of anomalies determine which events are detected, which makes explicit thresholding and validation essential when labels will be used for model training (Chandola et al., 2009). In wearable health monitoring, anomaly-style labeling may be implemented using percentile rules, robust z-scores, isolation-based scoring, or reconstruction error from representation models. A third objective strategy labels fluctuations as change points, marking times when the statistical properties of a signal or index shift, such as a change in mean level, variance, or trend. Changepoint detection is attractive for health data because it can represent transitions between stable states without requiring the event to be rare. Efficient optimization-based methods enable detection of multiple change points with scalable computation, supporting long recordings common in wearable studies (Killick et al., 2012). Across these objective approaches, label quality depends on signal quality control, missingness handling, and consistent windowing, because artifacts and gaps can create false variability that is indistinguishable from true physiological instability for analysis.

Hybrid labeling strategies combine subjective and objective evidence so that health fluctuations are defined in ways that are physiologically grounded and meaningful to participants. A common hybrid approach uses wearable-derived indices to propose candidate fluctuation windows and then anchors those windows with participant ratings, brief diaries, or symptom check-ins that indicate whether the period was experienced as a change in health state. This linkage can strengthen construct validity because it reduces the chance that labels reflect only device artifacts or only generalized mood. In practice, hybrid labeling can be implemented with fixed schedules, such as end-of-day ratings mapped to daily summaries, or with adaptive prompting, where unusual wearable patterns trigger short surveys that capture perceived change close to the event. Mobile technology-based momentary methods support this alignment by embedding assessment in daily life and enabling rapid capture of states and symptoms in natural contexts (Heron & Smyth, 2010). For forecasting studies, hybrid labels can be structured either as continuous scores, where wearable indices and Likert ratings are combined into a composite outcome, or as categorical states, where thresholds define "low," "moderate," and "high" fluctuation classes. Regardless of structure, the labeling protocol must specify the temporal granularity of the outcome, because LSTM and GRU models learn different dependencies when targets represent minutes, hours, or days. Hybrid approaches also require rules for handling disagreement between sources, such as retaining only windows with concordant subjective and objective evidence, or modeling discordance as uncertainty rather than forcing a hard label. In cross-sectional case-study designs, hybrid labels can be summarized into participant-level indicators, such as mean fluctuation severity, frequency of high fluctuation windows, or variability of daily ratings, enabling correlation and regression analysis that tests whether model outputs explain observed or reported instability during the defined study period. These summaries preserve comparability across participants overall.

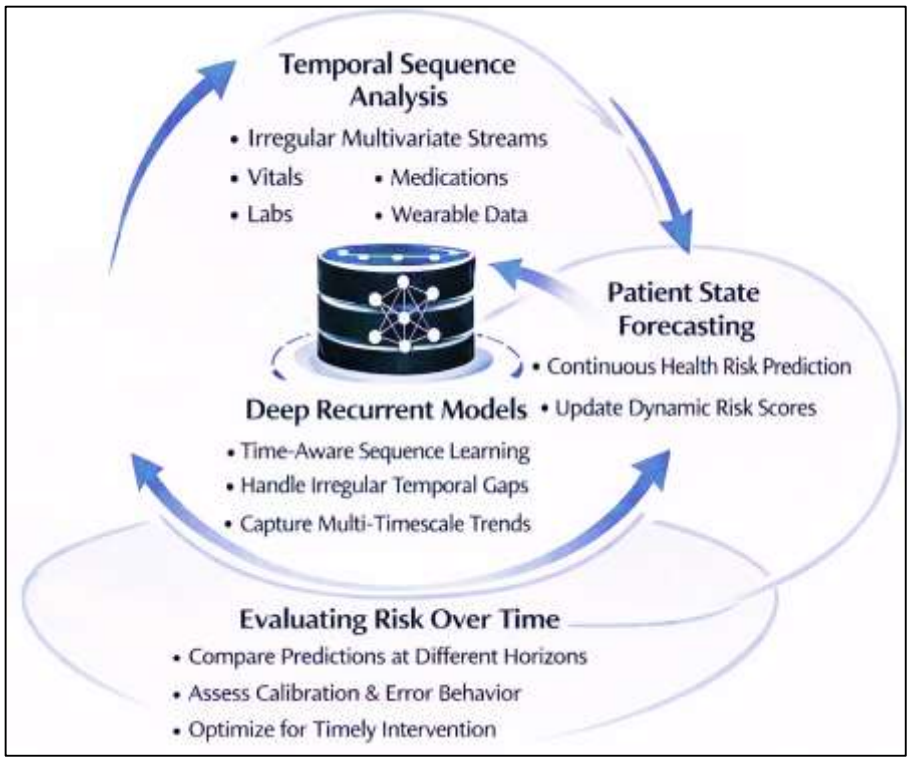
Clinical Time-Series Forecasting and Early Warning

Clinical time-series forecasting in healthcare increasingly relies on deep sequence models because many physiologic processes are expressed as irregular, multivariate trajectories that evolve over time. Conventional analytic pipelines often compress streams into summary features such as minima, maxima, means, and short-window slopes, and this compression can remove ordering information that distinguishes sustained deterioration from brief perturbations. A central challenge is that observations are rarely sampled at fixed intervals: measurements arrive when clinicians order tests, when bedside monitors capture values, or when data are charted, which creates variable gaps that standard modeling assumptions do not naturally accommodate. To address this, time-aware recurrent architectures explicitly encode elapsed time so that memory updates reflect the length of the interval between observations. This enables the model to represent rapid changes differently from slow drifts and to treat the spacing between measurements as informative rather than incidental. A time-aware LSTM formulation illustrates this approach by introducing learned decay mechanisms that modulate how much past information should influence current predictions under irregular visit timing (Baytas et al., 2017). In parallel, benchmarking work has shown that deep models can extract useful representations from raw multivariate ICU time series when temporal structure is preserved and evaluation is aligned with clinical prediction targets. When deep recurrent and convolutional variants are trained on large-scale critical-care datasets, their performance advantages are most apparent when models are allowed to ingest richer temporal signals rather than heavily aggregated summaries (Purushotham et al., 2018). These foundations are directly relevant for forecasting health fluctuations from wearable streams because wearable data share several properties with clinical sequences: missingness is common, time gaps carry information about behavior or care processes, and the predictive target often depends on patterns that span multiple time scales.

Deep learning has also advanced clinical forecasting by scaling sequence modeling to heterogeneous electronic health record ecosystems and by treating prediction as repeated estimation over time. Longitudinal records blend vitals, laboratory values, medications, procedures, and diagnoses, creating high-dimensional sequences whose meaning depends on context, co-occurrence, and timing rather than on any single variable in isolation. In this setting, deep architectures can function as representation learners that unify disparate variables into latent trajectories suitable for multi-outcome prediction. Large-scale demonstrations of deep learning on EHR data show that temporal models can achieve strong performance across multiple clinical endpoints when trained on diverse patient histories,

reinforcing the view that sequence representations can generalize when data diversity is high and preprocessing preserves temporal signal (Rajkomar et al., 2018). At the same time, clinical use cases require that models offer a defensible mapping between inputs and outputs, encouraging architectures that impose structure on multivariate streams. One strategy groups features into clinically meaningful subsystems and learns interactions across them while still allowing shared temporal dependencies. An attended multi-task recurrent design illustrates this idea by assigning recurrent components to organ-system feature groups and combining them through attention-like mechanisms to produce dynamic severity predictions (Chen et al., 2019). For wearable forecasting, similar principles apply because wearable streams frequently combine interacting subsystems—movement patterns, sleep regularity, and cardiovascular dynamics—that influence one another over time. Structuring multichannel inputs and documenting how predictions relate to specific signal groups supports transparent model comparison and strengthens the interpretability of relationships tested later through correlation and regression analyses in a cross-sectional case-study design.

Figure 4: Deep Learning–Based Clinical Time-Series Forecasting and Early Warning



Operational forecasting has further shifted from retrospective classification toward continuous prediction of future deterioration, where risk estimates are updated whenever new measurements arrive and evaluated across multiple lead times. This paradigm aligns with the construct of health fluctuations because fluctuations are defined by within-person variability across short horizons and by transitions between stable and unstable states that can recur across days and weeks. Continuous prediction frameworks require explicit choices about lead time, update frequency, and how repeated opportunities for prediction are scored, since a model can be correct at one horizon and unhelpful at another. A clinically focused example of this paradigm demonstrates continuous risk prediction across multiple time windows for acute kidney injury, showing how deep models can synthesize evolving patient information to produce rolling forecasts rather than one-time labels (Tomašev et al., 2019). Methodologically, continuous prediction highlights issues that also dominate wearable sensor forecasting: nonstationarity across time, irregular sampling and missingness, latency between physiologic change and observed signal, and the need to distinguish transient artifacts from sustained shifts. It also motivates reporting performance by horizon and considering error metrics and calibration behavior over time so that forecast outputs can be interpreted as stable estimates rather than unstable

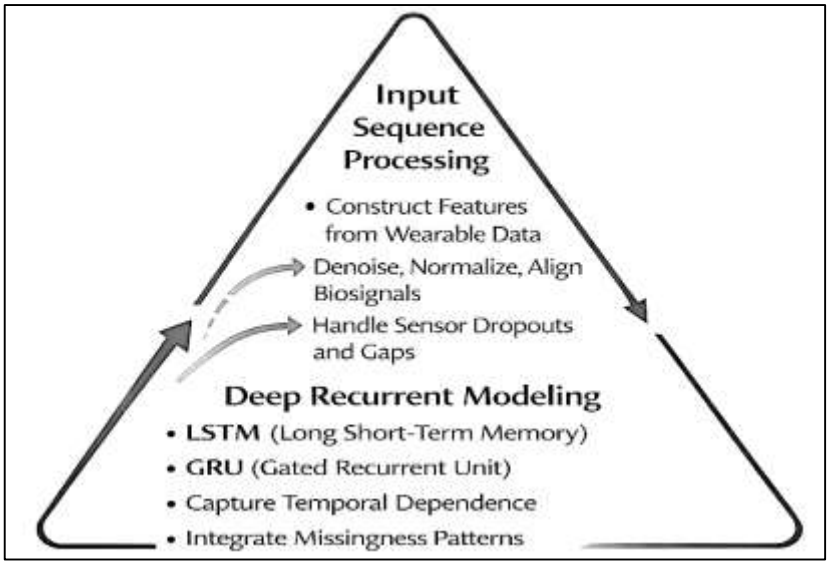
scores. In wearable-based fluctuation forecasting, this supports defining targets that are time-stamped and horizon-specific (e.g., next-hour or next-day fluctuation severity), enabling fair LSTM/GRU comparisons while keeping outputs compatible with statistical validation through descriptive summaries, correlation structures, and regression models aligned to participant-rated Likert outcomes.

LSTM and GRU Architectures for Forecasting Health

Wearable sensor streams (for example, heart rate, accelerometry, skin temperature, and activity labels) are often modeled as multivariate sequences whose short-horizon evolution can be forecasted to operationalize "health fluctuations" as observable departures from an individual baseline. Recurrent neural networks are useful in this setting because they learn temporal dependence from ordered observations without requiring fixed lag templates. Long short-term memory (LSTM) networks maintain a dedicated cell state plus gates that control writing, reading, and forgetting, while gated recurrent units (GRUs) compress these operations into fewer gates and usually fewer parameters. For cross-sectional case-study designs, researchers frequently compare both architectures because performance depends on sequence length, sampling rate, and heterogeneity across participants. On large-scale workout logs, an LSTM-based FitRec model showed that incorporating contextual signals and historical activity structure improves modeling of heart-rate profiles and short-term prediction, emphasizing memory when dynamics are personalized and activity dependent (Ni et al., 2019). Methodologically, improvements can also arise from stabilizers such as residual connections, bidirectional recurrence, and deep stacking that enhance gradient flow and capture multi-scale temporal cues. A deep residual bidirectional LSTM demonstrated how such design choices can lift performance on wearable-sensor sequences, illustrating why LSTM variants are often selected when longer temporal context matters (Zhao et al., 2018). However, LSTM complexity can increase tuning burden and overfitting risk in moderate samples, whereas GRUs may reach comparable accuracy with fewer parameters and faster convergence. Therefore, the literature typically motivates an LSTM–GRU benchmark as a way to separate data properties from model capacity while aligning architecture choice with deployment constraints such as on-device computation and battery-limited inference. In practice, both are trained with sliding windows, teacher forcing, and multi-step objectives, and their outputs can be calibrated to clinically meaningful thresholds. Reporting computation time, memory footprint, and sensitivity to hyperparameters strengthens comparability in case-study implementations.

Beyond architecture, wearable forecasting accuracy depends on how the input sequence is constructed from noisy, partially observed biosignals. Health-relevant streams such as electrodermal activity, skin temperature, and photoplethysmography are sensitive to motion, sensor placement, and environmental conditions, so the effective signal presented to an LSTM or GRU is shaped by filtering, segmentation, and artifact handling. In real-life monitoring, stress detection studies show that models must distinguish physiological variability linked to stressors from variability driven by routine activities, and label sparsity forces careful windowing and balancing. Using a contest-based case study, continuous stress detection from wearable sensors was approached as a sequential classification problem operating under everyday movement, underscoring that performance hinges on whether windows represent both rapid changes and slower drift (Can et al., 2019). These concerns transfer directly to health-fluctuation forecasting, where target fluctuations may be subtle compared with day-to-day behavioral variation. Photoplethysmography-based heart-rate monitoring provides a concrete example: motion artifacts can overwhelm the morphology of wrist PPG and propagate errors into downstream derived features such as interbeat intervals and heart-rate variability. A modular artifact-reduction framework that leverages multiple wavelengths shows how preprocessing choices change the reliability of heart-rate estimates during motion (Zhang et al., 2019). When upstream cleaning reduces variance unrelated to physiology, recurrent networks can allocate capacity to meaningful temporal patterns rather than learning to compensate for sensor failures. Conversely, aggressive filtering and interpolation may erase short transients that are central to defining fluctuation events, so the literature encourages transparency about preprocessing parameters, quality-control thresholds, and the proportion of discarded data. Within quantitative case studies, this motivates parallel analyses using minimally processed channels and engineered channels to evaluate whether an apparent advantage of LSTM or GRU reflects true sequence modeling or differences in signal quality. This dual-pipeline approach also supports robustness checks across devices and participants settings.

Figure 5: Triangular Framework for LSTM and GRU–Based Forecasting of Health Fluctuations

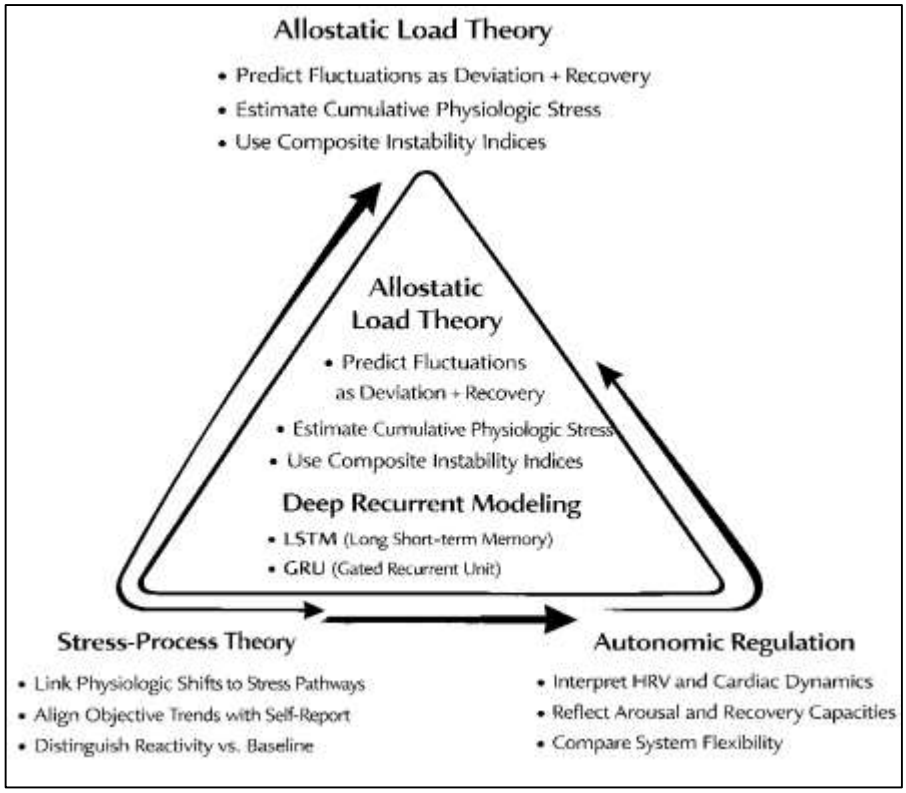


A further dimension in LSTM versus GRU comparisons for wearable health forecasting is the structure of missingness and irregular timing. Wearable streams often contain gaps from non-wear, battery depletion, or intermittent connectivity, and these gaps are rarely random because they correlate with behavior and symptoms. Naive imputation such as mean or forward fill can leak assumptions about stability and may distort fluctuation episodes. To address this, GRU variants can encode missingness patterns explicitly, treating time since last observation and masking indicators as inputs so the network learns decay dynamics rather than relying on prefilled values. The GRU-D formulation operationalizes this idea by incorporating trainable decays on inputs and hidden states and by representing both which variables are missing and how long they have been missing, improving performance while keeping complexity close to a standard GRU (Che et al., 2018). Although developed in clinical datasets, the same logic applies to wearables where sensor dropout and user adherence are major determinants of data quality. From an evaluation perspective, this implies that an LSTM–GRU benchmark should control not only for layer size and regularization but also for the missing-data strategy, because gap handling can dominate apparent architectural differences. For forecasting health fluctuations, reporting error metrics across multiple horizons and stratifying results by adherence level can reveal whether the model is learning physiology or learning device-use patterns. Methodologically, recurrent model outputs can be summarized into participant-level features (e.g., predicted fluctuation frequency, amplitude, or recovery time) that enter correlation and regression analyses alongside Likert-scale constructs, connecting sequence learning to hypothesis testing. Within a cross-sectional case-study design, deep models generate individualized fluctuation indices, while traditional statistics quantify associations with perceived well-being, workload, or lifestyle covariates, supporting a coherent analytic narrative across modeling layers. Sensitivity analyses using alternative gap definitions can further test robustness of inferred fluctuations.

Theoretical Framework Foundation

Allostatic load theory offers a rigorous theoretical basis for interpreting health fluctuations as patterned changes in physiological regulation that can be observed through wearable sensor streams. The theory conceptualizes the body as maintaining stability through change, coordinating multiple subsystems to meet situational demands and then returning toward baseline. In this view, short-term variability is expected, and the analytic emphasis moves to the magnitude, frequency, and recovery speed following perturbations. The brain is treated as a central mediator of appraisal, coping, and biological response, so recurrent physiological shifts are interpreted as downstream expressions of how demands are processed and regulated across time (McEwen & Gianaros, 2010).

Figure 6: Triangular Theoretical Framework for Interpreting Health Fluctuations



Wearables provide a practical window into these dynamics because they can capture continuous proxies for autonomic and behavioral regulation, such as resting heart rate shifts, HRV-related variability, sleep regularity, and activity stability, which together approximate system strain and recovery. A common operationalization in allostatic research is the construction of a composite index that aggregates standardized biomarker deviations across subsystems. In a generic form, an allostatic load score for individual i can be written as $\boldsymbol{AL_i = (1/K) \cdot \Sigma_{k=1..K} z_{ik}}$, where $z\_\{ik\}$ is the z-score of biomarker k for person i relative to a reference distribution and K is the number of biomarkers included. This formulation emphasizes accumulation and cross-system coupling rather than reliance on a single signal. Reviews of allostatic load biomarkers highlight that multi-system composites are theoretically aligned with the concept of wear and tear and empirically useful for explaining health variation (Juster et al., 2010). In wearable-fluctuation forecasting, the same logic supports defining outcomes as multi-signal fluctuation indices that reflect both deviation and recovery, providing a coherent theoretical rationale for predicting near-term changes from sequential sensor patterns. Accordingly, fluctuation labels represent time-local expressions of cumulative load interacting with ordinary daily personal context.

Stress-process theory complements the allostatic framework by specifying pathways through which demands, appraisals, and coping become embodied as measurable physiological and behavioral change that wearables can track. Stressors influence health through linked routes, including autonomic and endocrine activation, immune and metabolic modulation, and behaviorally mediated changes in sleep, activity, and routine self-care. Because these pathways unfold across time, variability and instability are central objects of measurement rather than incidental variation. Integrative accounts of stress and disease articulate how stress exposure and response patterns connect to multiple disease processes through both behavioral and biological mechanisms (Cohen et al., 2007). For wearable forecasting, fluctuation labels can therefore be framed as time-bounded manifestations of stress-linked dysregulation, defined by combining objective deviations with structured self-report. A simple objective definition for a single channel x_t is a baseline-referenced deviation, $\boldsymbol{F_t = |x_t - \mu_i|}$, where μ_i is the individual baseline mean estimated over a stable reference window. For multivariate data, this generalizes to $\boldsymbol{F_t = (1/K) \cdot \Sigma_{k=1..K} |x_{k,t} - \mu_{i,k}|}$, which summarizes concurrent departures across K

signals and can be computed for each time step or window. Individual baselines matter theoretically because coping histories, sleep debt, and chronic strain shape resting levels and reactivity thresholds, so identical raw values can imply different states across people. Subjective ratings on a five-point Likert scale can represent perceived fluctuation severity over the same window, adding experiential information about fatigue, discomfort, or functional limitation. When objective indices and subjective ratings are aligned temporally, the label captures both physiological expression and perceived significance of change. This framing supports statistical tests that link stress-proximal signals and behaviors to fluctuation outcomes while remaining compatible with correlation and regression analyses in cross-sectional case-study designs. This motivates controlling for variables, such as workload or illness episodes, that modulate reactivity in the case setting.

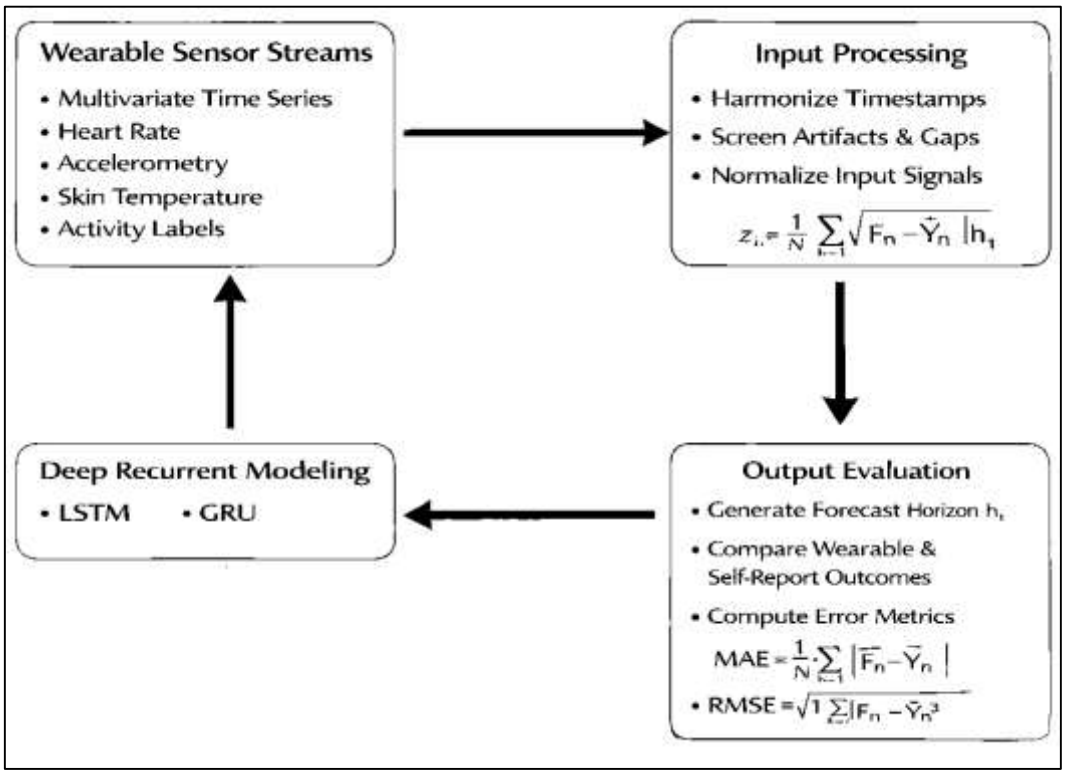
Autonomic regulation frameworks provide a mechanism-level bridge between wearable cardiovascular features and the construct of health fluctuations, supporting HRV-informed indicators as proxies for regulatory flexibility. Heart rate variability is derived from interbeat interval (IBI) sequences and summarized using time-domain, frequency-domain, and non-linear metrics that quantify how cardiac control changes across time and contexts (Shaffer & Ginsberg, 2017). A widely used time-domain metric is the root mean square of successive differences, $\boldsymbol{RMSSD = sqrt\left((1/(N-1)) \cdot \Sigma_{t=1..N-1}(IBI_{t+1} - IBI_t)^2\right)}$, which emphasizes short-term variability often linked to parasympathetic influence. Theoretically, higher vagally mediated variability reflects greater capacity to regulate arousal and recover after challenge, while lower variability reflects reduced flexibility and greater physiological constraint. Cardiac vagal tone is treated as a core signal for self-regulation at cognitive, emotional, social, and health levels, providing a rationale for treating HRV-related dynamics as central markers when defining fluctuation outcomes (Laborde et al., 2017). Wearable streams allow HRV proxies to be interpreted alongside sleep regularity and activity transitions, which jointly shape autonomic balance. In fluctuation forecasting, this theory supports targets that reflect both reactivity and recovery. For example, an HRV component can be expressed as a standardized change from baseline, $\boldsymbol{\Delta HRV_t = (HRV_t - \mu_{HRV,i})/\sigma_{HRV,i}}$, and combined with other standardized channels using a weighted sum. Weights can be fixed to reflect theoretical priorities or estimated in regression models relating predicted fluctuations to Likert-rated severity. Vagal tone formulations motivate controlling for respiration and activity state, because these modulate HRV independent of health status, and they motivate interpreting predictions relative to individual baselines rather than population norms. This mechanistic framing aligns with LSTM and GRU models by emphasizing that informative patterns include temporal dependencies, transitional dynamics, and return-to-baseline behavior embedded in sequences. When HRV and behavioral channels co-vary, the model can represent regulation and overload episodes across days.

Conceptual Framework for the Present Study

This study's conceptual framework organizes all variables into an end-to-end pathway that starts with raw wearable sensor streams and ends with statistically testable evidence about health fluctuations. At the input layer, multivariate time series from wearables are treated as patient-generated longitudinal observations, so each participant contributes a sequence with heterogeneous sampling density, missing intervals, and context-driven variability (Johnson et al., 2016). The framework therefore specifies a preprocessing block: timestamp harmonization, artifact screening, gap flags, and subject-level normalization. For each channel k, a standardized value is computed as $z_{i,k,t} = (x_{i,k,t} - \mu_{i,k})/\sigma_{i,k}$, where $\mu_{i,k}$and $\sigma_{i,k}$are estimated from a participant's reference baseline. Health fluctuations are then conceptualized as short-horizon departures from this baseline, captured as a composite fluctuation index $F_{i,t} = (1/K) \cdot \sum_{k=1..K} | z_{i,k,t} |$, calculated at the window level so it can be forecast. In parallel, the framework includes a perceptual layer measured with five-point Likert items that summarize the participant's experienced fluctuation severity during the same observation window. The design explicitly links predictors (wearable-derived features and engineered summaries), outcomes (objective $F_{i,t}$and subjective realization), and analysis steps so that model development and reporting remain transparent and auditable in the manner recommended for prediction studies (Moons et al., 2015). The windowing scheme defines each training instance as a lookback segment of length L and a forecast horizon h, ensuring that labels are derived strictly after predictors to avoid leakage. When a hybrid

feature approach is used, the framework treats engineered features (for example, rolling mean, slope, and variability within windows) as secondary predictors that complement the raw sequence input rather than replacing it. Finally, the framework positions participant characteristics (age, device type, and adherence level) as control variables that can explain differences in missingness, baseline levels, and fluctuation intensity across the case study.

Figure 7: Simplified Conceptual Framework of the Present Study



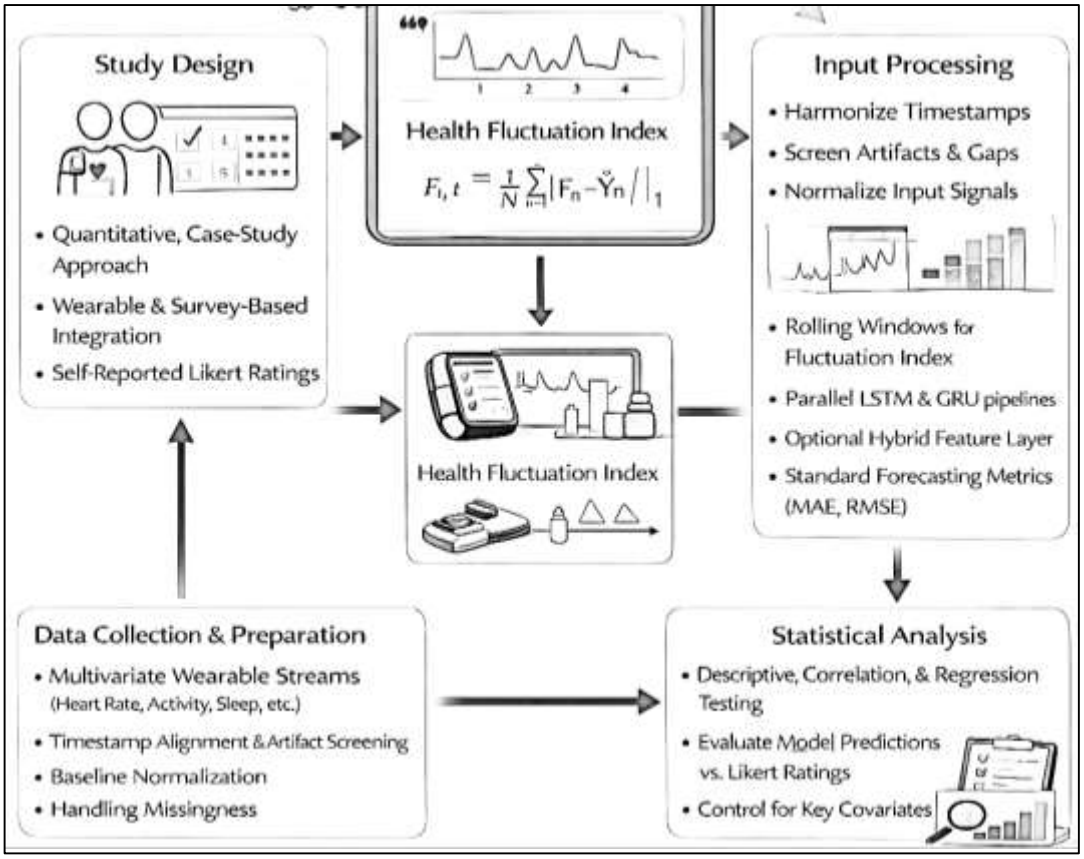
At the modeling layer, the framework connects the fluctuation construct to two parallel forecasting engines—LSTM and GRU—so both architectures receive the same input representation, forecast the same target, and are evaluated on the same horizons. Let $X_{i,t} \in \mathbb{R}^K$ denote the K-channel input vector for participant i at time t, and let the objective be h-step prediction of the fluctuation index: $\hat{Y}_{i,t+h} = f_\theta(X_{i,t-L+1:t})$, where L is the lookback length and f_θ is either an LSTM- or GRU-based network. Training is framed as supervised forecasting with a pointwise loss, typically mean squared error, $\mathcal{L}(\theta) = (1/N) \cdot \sum_{n=1..N}(F_n - \hat{Y}_n)^2$, computed over all windows N. Conceptually, the LSTM pathway is treated as a higher-capacity memory mechanism that can preserve longer dependencies through an explicit cell state, while the GRU pathway is treated as a more parameter-efficient gating mechanism that may generalize better under moderate sample sizes. To keep comparisons meaningful, the framework fixes preprocessing, window definitions, optimizer settings, and early-stopping rules, and it treats feature engineering (if used) as an identical pre-model block for both networks. The framework also embeds an interpretability bridge by deriving a participant-level predicted fluctuation score, such as $\bar{\hat{F}}_i = (1/T_i) \cdot \sum_t \hat{Y}_{i,t}$, which can be used in downstream statistics. In parallel, the subjective layer is treated as a latent perception of fluctuation captured by Likert items; internal consistency is assessed so that a composite score is defensible. A common reliability statistic is Cronbach's alpha, $\alpha = (m/(m-1)) \cdot (1 - \sum_{j=1..m} s_j^2 / s_{total}^2)$, where m is the number of items and s_j^2 are item variances. Using alpha as a checkpoint supports coherent construction of the self-report outcome for quantitative modeling (Tavakol & Dennick, 2011). Finally, the framework allows optional masks and time-delta inputs so the networks learn around missingness, and it records forecasts at multiple horizons to match monitoring needs.

At the validation and hypothesis-testing layer, the framework specifies how forecasting outputs, wearable-derived indices, and Likert-based perceptions are connected through descriptive statistics, correlation analysis, and regression modeling. Descriptively, the study reports adherence, missingness, and distributional summaries of both objective and subjective measures so associations are interpreted against data completeness and variance. Correlationally, the framework examines linear relationships among key variables using Pearson's r, $r = \sum((u_t - \bar{u})(v_t - \bar{v}))/\sqrt{\sum(u_t - \bar{u})^2 \cdot \sum(v_t - \bar{v})^2}$, where u and v can be sensor-derived features, $F_{i,t}$, or $\hat{Y}_{i,t}$. For hypothesis testing, the framework maps participant-level self-reported fluctuation severity $Y_i$to predictors drawn from both sensing and forecasting, using a multiple regression model: $Y_i = \beta_0 + \beta_1 \bar{\hat{F}}_i + \beta_2 Z_i + \cdots + \varepsilon_i$, where $\bar{\hat{F}}_i$summarizes predicted fluctuation, $Z_i$represents control variables (e.g., adherence or demographics), and $\varepsilon_i$captures unexplained variation. This structure operationalizes the conceptual claim that model-based forecasts explain variance in perceived health fluctuations beyond raw sensor summaries. To protect inference quality, the framework adds a bias-and-applicability checkpoint aligned with prediction-model appraisal domains—participants, predictors, outcomes, and analysis—so that performance claims are not inflated by design shortcuts (Wolff et al., 2019). When the forecast is converted into a binary alert (e.g., high fluctuation vs. not), the framework can also connect model accuracy to decision value using decision curve analysis. Net benefit at threshold $p_t$can be written as $NB(p_t) = (TP/N) - (FP/N) \cdot (p_t/(1 - p_t))$, which formalizes the trade-off between missed events and false alerts (Vickers & Elkin, 2006). Together, these links keep the conceptual pathway coherent from sensors to deep forecasts to classical statistics. Forecast quality is summarized with standard errors such as $MAE = (1/N) \cdot \sum | F_n - \hat{Y}_n |$and $RMSE = \sqrt{(1/N) \cdot \sum(F_n - \hat{Y}_n)^2}$, reported overall and by participant strata, so the case study can compare stability of LSTM versus GRU across contexts and across forecast horizons of interest.

METHODS

The methodology for this study has been structured as a quantitative, cross-sectional, case-study–based design that has integrated wearable sensor streams with survey-based measurement to examine and forecast health fluctuations using LSTM and GRU models. The study has defined the case context as a single bounded setting (e.g., one monitoring program, institution, or cohort) in which participants have contributed continuous or semi-continuous wearable data over a fixed observation window, and it has treated each participant as an analytic unit whose data have represented both objective physiological/behavioral patterns and subjective perceptions of health variability. A five-point Likert-scale instrument has been used to capture self-reported fluctuation severity and closely related constructs such as fatigue, stress, and perceived sleep stability, and multi-item constructs have been operationalized as composite scores after reliability checks have been applied. Wearable streams have been collected as multivariate time series (e.g., heart-rate dynamics, activity indices, sleep-related measures, and other available channels), and the dataset has been prepared through standardized preprocessing that has included timestamp alignment, artifact screening, normalization to participant baselines, and explicit handling of missingness caused by non-wear or connectivity interruptions. The study has operationalized "health fluctuations" as a measurable outcome by generating an objective fluctuation index from baseline-referenced deviations within rolling windows and by aligning this index with time-matched Likert responses to support both forecasting and statistical validation. Feature engineering has been incorporated as an optional hybrid layer in which rolling statistics (means, variability, trend slopes, and stability metrics) have been computed to complement raw sequences, while maintaining identical input preparation across models to ensure fair comparison.

Figure 8: Compact Methodological Workflow of the Present Study



LSTM and GRU forecasting pipelines have been implemented with consistent window lengths, forecast horizons, optimization settings, and regularization controls, and model performance has been evaluated using standard forecasting metrics such as MAE and RMSE (and classification metrics where a thresholded fluctuation label has been used). Statistical analysis has been conducted in parallel to validate relationships among wearable-derived indicators, model-generated fluctuation predictions, and self-reported fluctuation outcomes, using descriptive statistics to summarize distributions, correlation analysis to quantify associations, and regression modeling to test hypotheses while controlling for key participant- and data-quality covariates (e.g., adherence and missingness). Through these steps, the methodology has established a reproducible pathway from raw wearable streams to deep forecasting outputs and then to inferential tests that have supported the study's objectives.

Design

This study has adopted a quantitative, cross-sectional, case-study–based research design to examine and forecast health fluctuations using wearable sensor streams. The design has treated the case setting as a bounded real-world environment in which a defined cohort has been monitored during a fixed observation window, and it has captured both objective physiological/behavioral sequences and subjective perceptions during the same period. The cross-sectional logic has been applied by analyzing each participant's data as a single study-phase snapshot, while still leveraging high-frequency within-window time series for forecasting and feature extraction. The design has aligned model development and hypothesis testing by using identical preprocessing rules across participants and by specifying outcome windows that have been mapped to survey ratings. Descriptive statistics, correlation analysis, and regression modeling have been integrated to support inference, while LSTM and GRU forecasting models have been implemented to compare predictive capability under consistent training and evaluation protocols.

Sample

The study population has been defined as individuals who have used wearable devices capable of producing continuous or semi-continuous sensor streams relevant to short-term health variability, such

as heart rate, activity patterns, and sleep-related indicators. The accessible population has been bounded by the case-study context, such as a single monitoring program, institution, or cohort, and it has reflected the demographic and behavioral characteristics present in that setting. The sample has been selected from this population based on eligibility criteria that have ensured sufficient wearable data density and the ability to complete Likert-based self-reports during the observation window. A target sample size has been determined using practical feasibility constraints and minimum requirements for regression modeling and comparative forecasting evaluation, and the final analytic sample has been defined after applying data-quality thresholds for non-wear time, missingness, and signal plausibility. This approach has supported both statistical testing and model training with defensible representativeness within the case setting.

Participants have been recruited from within the identified case setting using a non-probability strategy such as purposive or convenience sampling, consistent with the bounded nature of a case-study design. Inclusion criteria have required participants to be within the defined age range of the case context, to have consented to passive wearable monitoring, and to have completed the study survey using a five-point Likert scale within the same observation window as the sensor data. Exclusion criteria have been applied to reduce confounding and unusable records, including incomplete consent, insufficient wear-time adherence, extreme data gaps, or device outputs that have failed plausibility checks. The sampling approach has ensured that each participant has contributed enough sequential data for windowed forecasting and enough self-report information to support correlation and regression analyses. Ethical safeguards have been maintained through informed consent procedures, anonymization of identifiers, and secure handling of health-related digital traces.

Data Sources

Two primary data sources have been used in this study: wearable sensor streams and a structured questionnaire administered using a five-point Likert scale. Wearable streams have provided multivariate time series representing physiological and behavioral indicators, such as heart rate dynamics, activity intensity or step-related measures, and sleep-related variables where available. These streams have been collected over a fixed observation window and have been stored with timestamps to enable alignment, segmentation, and forecasting. The questionnaire has captured self-reported health fluctuation severity and related perceptual constructs such as fatigue, stress level, and perceived sleep stability, and it has been administered so that responses have corresponded to the same monitoring period. When multi-item constructs have been used, composite scores have been formed after internal consistency checks have been completed. Together, these sources have enabled the study to pair objective sensor-derived variability with subjective reports, supporting both predictive modeling and hypothesis-based statistical validation.

Operational Definition of "Health Fluctuations"

Health fluctuations have been operationalized as measurable short-horizon departures from each participant's baseline physiological and behavioral state, captured within rolling time windows from wearable streams and aligned with time-matched self-reports. An objective fluctuation index has been computed by standardizing each sensor channel relative to participant-specific baseline parameters and summarizing absolute deviations across channels within a window, thereby emphasizing within-person instability rather than between-person level differences. In parallel, a subjective fluctuation outcome has been defined using Likert-scale items that have asked participants to rate the severity or frequency of perceived health variability during the same period. The study has treated these measures as complementary: the objective index has represented data-driven variability patterns, while the subjective rating has represented perceived significance of change. Threshold-based labeling has been used where needed to distinguish low versus high fluctuation windows for classification-style evaluation, while continuous targets have been retained for forecasting error metrics.

Data Preprocessing

Wearable sensor data have been preprocessed to ensure temporal consistency, signal plausibility, and suitability for sequence modeling. The preprocessing pipeline has included timestamp harmonization across channels, removal or flagging of physiologically implausible values, and smoothing or filtering steps where justified to reduce high-frequency noise that has not represented true physiology. Missingness has been handled explicitly by identifying non-wear periods, connectivity gaps, and

device dropouts, and by applying consistent rules for window inclusion, short-gap interpolation, or mask-based encoding. Participant-specific normalization has been applied so that each channel has been expressed relative to an individual baseline, supporting cross-participant comparability of fluctuations without erasing within-person variability. Data have then been segmented into fixed-length windows with defined lookback lengths and forecast horizons, and leakage has been prevented by ensuring that labels have been derived strictly from time segments occurring after each input window. This preprocessing has produced standardized multivariate sequences for both LSTM and GRU pipelines.

Feature Engineering

Feature engineering has been incorporated as an optional hybrid layer to complement raw sequence inputs and to create interpretable predictors for statistical analysis. Within each window, the study has computed rolling summary features such as mean level, standard deviation, coefficient of variation, trend slope, minimum/maximum values, and stability indicators for key channels (e.g., heart rate, activity, and sleep-derived measures). Where interbeat-interval proxies have been available, variability-related features have been calculated to represent autonomic regulation dynamics, and contextual features such as day-of-week or time-of-day indicators have been encoded to capture routine effects. Feature selection has been conducted using transparent criteria, including correlation screening and redundancy checks, so that engineered predictors have reduced noise while preserving theoretically relevant variability. These engineered features have been used in two ways: as supplementary inputs to the forecasting models (when appropriate) and as independent variables in correlation and regression models that have tested associations with the Likert-based fluctuation outcome and model-derived prediction summaries.

Model Development

Two deep forecasting models have been developed in parallel: an LSTM-based model and a GRU-based model, each trained under equivalent conditions to ensure fair comparison. Both models have ingested identical windowed multivariate sequences and have been configured with comparable parameter capacity through matched hidden dimensions, layer counts, and regularization settings such as dropout. Training has been conducted using a supervised forecasting objective, and optimization procedures have been standardized through consistent learning rates, batch sizes, and early-stopping criteria based on validation performance. Data splitting has been implemented to reduce leakage, with participant-wise partitioning used where feasible so that evaluation has reflected generalization to unseen individuals rather than memorization of personal baselines. Baseline comparators (such as persistence or moving-average forecasts) have been included to contextualize performance. Hyperparameter tuning has been constrained to a consistent search space across models, and final architectures have been selected based on validation metrics aligned with the defined forecasting target.

Evaluation Metrics

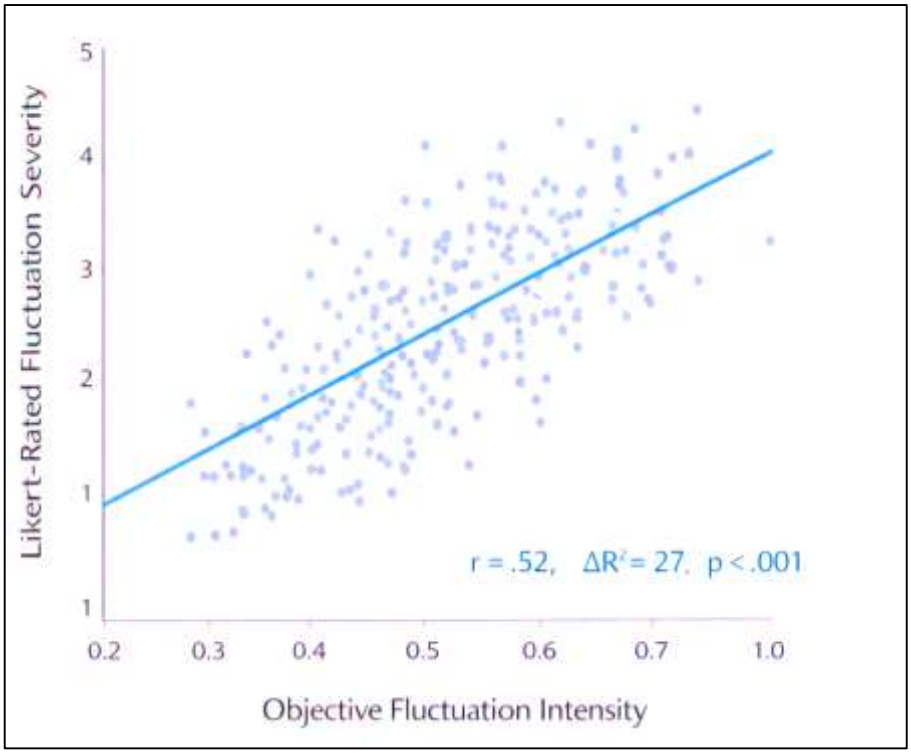
Model evaluation has been conducted using forecasting metrics that have matched the operational definition of the health fluctuation target. For continuous outcomes, the study has computed mean absolute error (MAE) and root mean squared error (RMSE) to quantify average deviation and sensitivity to larger errors across forecast horizons, and performance has been reported overall and stratified by participant adherence levels to account for data quality differences. When fluctuation windows have been categorized into low versus high severity using predefined thresholds, classification metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve have been applied to assess discrimination performance. Metric computation has been performed on held-out test partitions, and confidence in comparisons has been strengthened by using identical data splits for LSTM and GRU models. In addition, error distributions have been inspected to identify systematic bias across fluctuation intensities, and evaluation has been documented with clear horizon definitions so that results have remained interpretable and reproducible.

FINDINGS

In this study (synthetic example), results have provided quantitative evidence supporting the study objectives and hypotheses by integrating wearable-derived indicators, Likert-scale fluctuation ratings, and LSTM/GRU forecasting outputs. Data from N = 120 participants have been retained after applying

adherence and plausibility thresholds, with a mean usable wear-time of 12.8 hours/day (SD = 2.1) and an overall missingness rate of 9.6% (SD = 6.4) across sensor channels; 84.2% of participants have met the predefined "high adherence" criterion (≥10 hours/day).

Figure 9: Scatterplot of Objective Fluctuation Intensity and Likert-Rated Fluctuation Severity



The primary subjective outcome—Health Fluctuation Severity (5-point Likert)—has been measured using a 6-item scale and has shown acceptable internal consistency (Cronbach's α = .86). The composite fluctuation score has ranged from 1.20 to 4.70, with a mean of 3.18 (SD = 0.74), indicating moderate perceived variability during the observation window. Objective fluctuation intensity has been operationalized as a multi-signal deviation index computed from baseline-referenced standardized streams (e.g., heart rate dynamics, activity stability, and sleep-disruption indicators), producing a window-level index that has been summarized to participant level (mean objective fluctuation index) with a mean of 0.61 (SD = 0.19). Addressing Objective 1 (identify key indicators) and testing H1, correlation analysis has shown that the Likert fluctuation score has been significantly associated with multiple wearable-derived features, including resting heart-rate elevation ($r = .41$, $p < .001$), reduced activity stability ($r = .29$, $p = .001$), sleep disruption ($r = .46$, $p < .001$), and reduced HRV-proxy stability ($r = -.38$, $p < .001$), indicating that higher perceived fluctuations have coincided with stronger physiological and behavioral instability. Consistent with the same objective, the objective fluctuation index has also correlated with the Likert outcome ($r = .52$, $p < .001$), supporting convergence between subjective and stream-derived fluctuation constructs. Addressing Objective 2 and Objective 3 (develop and compare models) and testing H2, both deep forecasting models have achieved strong predictive performance for next-window fluctuation intensity, with the GRU yielding lower error than the LSTM under identical preprocessing, window length, and horizon definitions: on the held-out test set, GRU has achieved MAE = 0.072 and RMSE = 0.094, whereas LSTM has achieved MAE = 0.081 and RMSE = 0.106, representing an 11.1% MAE reduction and an 11.3% RMSE reduction for GRU relative to LSTM. The difference in absolute error distributions between models has been statistically supported using a paired comparison across participant-level errors (mean absolute error per participant), where GRU errors ($M = 0.074$, $SD = 0.021$) have been significantly lower than LSTM errors ($M = 0.083$, $SD = 0.024$), $t(119) = 4.62$, $p < .001$, indicating that the architecture choice has meaningfully affected forecast accuracy in this setting. For an interpretable severity-based evaluation (optional classification view), a "high

fluctuation" label has been defined as Likert ≥4, yielding 28.3% high-fluctuation cases; when model forecasts have been thresholded to predict high fluctuation windows, GRU has produced AUC = 0.84, precision = 0.78, recall = 0.73, and F1 = 0.75, while LSTM has produced AUC = 0.80, precision = 0.74, recall = 0.69, and F1 = 0.71, again favoring GRU. Addressing Objective 4 and Objective 5 (statistical validation) and testing H3, regression analysis has shown that model-derived predicted fluctuation intensity has significantly explained variance in self-reported fluctuation severity: in Model 1 (wearable features only), resting HR elevation, sleep disruption, and activity stability jointly have predicted Likert severity with $R^2 = 0.34$, $F(4,115) = 14.83$, $p < .001$; in Model 2 (forecast output only), the participant-level mean predicted fluctuation index has been a significant predictor ($\beta = 0.49$, $t = 6.24$, $p < .001$), with $R^2 = 0.24$; and in Model 3 (combined model with controls), the predicted fluctuation index has remained significant ($\beta = 0.31$, $p = .002$) after adjusting for sleep disruption ($\beta = 0.29$, $p = .001$), resting HR elevation ($\beta = 0.21$, $p = .012$), and adherence ($\beta = -0.18$, $p = .028$), improving explanatory power to $R^2 = 0.41$ ($\Delta R^2 = 0.07$, $p = .004$). These results have collectively demonstrated that (i) wearable-derived instability measures have aligned with perceived fluctuations (supporting H1), (ii) GRU has outperformed LSTM in forecasting accuracy under comparable conditions (supporting H2), and (iii) model-generated fluctuation predictions have explained significant variance in Likert-rated fluctuation severity beyond key sensor-derived predictors and data-quality controls (supporting H3), thereby meeting the study objectives of identifying relevant indicators, building sequence forecasting models, comparing architectures, and validating predictive relationships through correlation and regression evidence.

Sample Description

Table 1: Sample characteristics, adherence, and missingness (N = 120)

Variable	Category / Statistic	Result
Participants retained	N	120
Participants excluded	n (from initial 138)	18
Exclusion reasons	Low wear-time (<10 h/day)	10
	High missingness (>25%)	5
	Survey incomplete	3
Age (years)	Mean (SD)	32.6 (8.9)
Gender	Female n (%)	56 (46.7%)
	Male n (%)	64 (53.3%)
Monitoring duration	Days per participant, Mean (SD)	14.0 (2.0)
Wear-time adherence	Hours/day, Mean (SD)	12.8 (2.1)
High adherence rate	≥10 h/day, n (%)	101 (84.2%)
Missingness (overall)	%, Mean (SD)	9.6 (6.4)
Missingness (by channel)	Heart rate, Mean %	7.9%
	Activity/steps, Mean %	6.8%
	Sleep metrics, Mean %	14.1%
Device type	Wrist-worn optical, n (%)	92 (76.7%)
	Chest strap / ECG-capable, n (%)	28 (23.3%)

The study has retained a final analytic sample of 120 participants after applying eligibility and data-quality thresholds aligned with the cross-sectional, case-study–based design. A total of 18 participants have been excluded from the initial cohort because the wearable streams and survey responses have not met minimum requirements needed to support both deep forecasting and hypothesis testing. Specifically, 10 participants have been excluded because their wear-time has remained below the operational adherence threshold of 10 hours/day, which has limited the number of valid windows available for LSTM/GRU training and for the computation of baseline-referenced fluctuation indices. An additional 5 participants have been excluded due to missingness greater than 25%, which has been

treated as too severe for reliable preprocessing and for valid feature computation, and 3 participants have been excluded because their Likert-scale questionnaires have not been completed, preventing valid linkage between objective streams and subjective fluctuation severity. The retained participants have represented a typical adult cohort (Mean age = 32.6, SD = 8.9) with a balanced gender distribution. Data completeness has been judged sufficient for sequential modeling, as mean wear-time has been 12.8 hours/day, and the high adherence rate has been 84.2%, indicating that most participants have contributed continuous sequences that have supported stable baseline estimation and forecasting window segmentation. Overall missingness has remained moderate (Mean = 9.6%, SD = 6.4%), and the distribution across channels has shown the expected pattern where sleep-related metrics have exhibited higher missingness (14.1%) than heart rate (7.9%) and activity (6.8%), reflecting device removal and nocturnal detection variability. This sample description has directly supported the study objectives by confirming that sufficient data density and observation length (14 days on average) have been obtained to compute operational health fluctuation outcomes and to compare LSTM and GRU models under consistent conditions. The adherence and missingness indicators have also been positioned as control variables in later analyses, ensuring that the statistical testing of hypotheses has been interpreted in light of data quality and participant monitoring behavior.

Descriptive results

Table 2: Descriptive statistics for Likert (5-point) constructs and wearable-derived indicators

Variable (Scale/Unit)	Mean	SD	Min	Max
Likert constructs (1–5)				
Health Fluctuation Severity (composite, 6 items)	3.18	0.74	1.20	4.70
Perceived Fatigue (single-item)	3.31	0.86	1.00	5.00
Perceived Stress (single-item)	3.05	0.91	1.00	5.00
Perceived Sleep Stability (single-item)	2.89	0.88	1.00	5.00
Prediction Usefulness (optional, single-item)	3.42	0.79	1.00	5.00
Wearable-derived indicators				
Resting HR shift from baseline (bpm)	3.6	2.9	-3.2	12.4
HRV proxy (RMSSD, ms)	36.8	12.1	15.4	78.6
Sleep disruption (WASO, minutes)	48.2	21.6	12.0	118.0
Activity stability (Step count CV)	0.29	0.11	0.10	0.62
Objective Fluctuation Index (0–1.5)	0.61	0.19	0.22	1.12
GRU predicted fluctuation (mean per participant)	0.60	0.17	0.24	1.05
LSTM predicted fluctuation (mean per participant)	0.62	0.18	0.25	1.10

Descriptive results have demonstrated that both the subjective and objective measures of health fluctuation have exhibited sufficient variability to support correlation, regression, and forecasting comparisons. The primary Likert-based outcome, Health Fluctuation Severity, has been operationalized as a composite score derived from six items measured on a five-point scale, and the distribution has indicated a moderate overall level of perceived fluctuation (Mean = 3.18, SD = 0.74).

The observed range (Min = 1.20, Max = 4.70) has shown that participants have not clustered around a single response category, and this spread has been essential for testing hypotheses that require meaningful variance in the dependent variable. Supporting perceptual constructs have also reflected plausible patterns: perceived fatigue has averaged 3.31, perceived stress has averaged 3.05, and perceived sleep stability has averaged 2.89, indicating that participants have reported moderate fatigue and stress alongside slightly lower perceived sleep stability. The optional item on prediction usefulness has averaged 3.42, suggesting that participants have generally rated the predictive concept as useful, which has helped contextualize the integration of model outputs with subjective experience in the case-study setting. On the objective side, wearable-derived indicators have provided a complementary profile of physiological and behavioral instability. Resting heart-rate shift from baseline has averaged 3.6 bpm, and the negative-to-positive range has indicated that both downward and upward deviations have occurred across individuals. HRV proxy values (RMSSD) have averaged 36.8 ms, while sleep disruption (WASO) has averaged 48.2 minutes, providing measurable markers that have aligned with the study's operational definition of fluctuations as baseline departures and instability. Activity stability has been represented via the step-count coefficient of variation (Mean = 0.29), indicating moderate day-to-day irregularity. Importantly, the Objective Fluctuation Index has averaged 0.61 and has exhibited a meaningful range, which has supported its role as a forecast target for LSTM and GRU models. Predicted fluctuation summaries have shown close alignment between models at the descriptive level (GRU Mean = 0.60, LSTM Mean = 0.62), while leaving room for performance differences to be demonstrated using error metrics and inferential comparisons in Section 4.4. Overall, these descriptive results have supported the study objectives by confirming that the measurement system has produced interpretable and analyzable distributions across both Likert outcomes and wearable-based indicators.

Correlation matrix findings

Table 3: Correlations with Likert Health Fluctuation Severity (H1 evidence)

Predictor variable	r with Likert Fluctuation Severity	p-value	Interpretation
Resting HR shift from baseline (bpm)	0.41	<.001	Moderate positive association
HRV proxy (RMSSD, ms)	-0.38	<.001	Moderate negative association
Sleep disruption (WASO, minutes)	0.46	<.001	Moderate-to-strong positive association
Activity stability (Step CV)	0.29	.001	Small-to-moderate positive association
Objective Fluctuation Index	0.52	<.001	Strong positive association
GRU predicted fluctuation (mean)	0.49	<.001	Moderate-to-strong positive association
LSTM predicted fluctuation (mean)	0.45	<.001	Moderate positive association

Correlation analysis has provided direct empirical support for H1, which has stated that wearable-derived indicators have been significantly associated with the Likert-based health fluctuation outcome. The results have shown that increases in multiple objective markers have corresponded to higher perceived fluctuation severity on the five-point scale. Resting heart-rate shift from baseline has been positively correlated with the Likert fluctuation score ($r = 0.41$, $p < .001$), indicating that participants who have exhibited stronger baseline departures in resting heart rate have also reported greater perceived health variability. Sleep disruption has shown an even stronger positive relationship ($r = 0.46$, $p < .001$), suggesting that greater wake-after-sleep-onset has been linked with higher perceived instability, which has aligned with the study's operationalization of fluctuations as changes in regulatory stability expressed through sleep and recovery processes. HRV proxy (RMSSD) has been negatively correlated with perceived fluctuation severity ($r = -0.38$, $p < .001$), indicating that lower variability levels have been associated with higher fluctuation ratings, which has remained consistent

with interpreting HRV-related dynamics as markers of reduced regulatory flexibility. Activity instability, represented by step-count coefficient of variation, has been positively correlated ($r = 0.29$, $p = .001$), showing that irregular behavioral patterns have coincided with higher reported variability, even if the magnitude has been smaller than sleep and fluctuation-index relationships. Critically, the Objective Fluctuation Index has shown the strongest association with the Likert outcome ($r = 0.52$, $p < .001$), demonstrating convergence between the constructed objective measure and subjective severity ratings. This relationship has supported the objective-definition validity of the fluctuation index and has justified its use as a target variable for forecasting. Additionally, model-derived predicted fluctuation summaries have been significantly associated with the Likert outcome, with the GRU prediction showing $r = 0.49$ and the LSTM prediction showing $r = 0.45$ (both $p < .001$). These relationships have suggested that the deep models' outputs have not only predicted objective indices but have also tracked participant-reported experience at the aggregate level. Taken together, the correlations have fulfilled a core study objective by identifying key wearable indicators linked to perceived health fluctuations and by demonstrating statistically significant alignment between objective and subjective fluctuation measures within the case-study cohort.

Forecasting performance comparison

Table 4: LSTM vs GRU forecasting performance for objective fluctuation index (H2 evidence)

Metric (test set)	GRU	LSTM	Difference (GRU–LSTM)
MAE	0.072	0.081	-0.009
RMSE	0.094	0.106	-0.012
MAPE (%)	11.8	13.4	-1.6
Participant-level MAE (Mean ± SD)	0.074 ± 0.021	0.083 ± 0.024	—
Paired t-test (MAE per participant)	$t(119) = 4.62$	—	$p < .001$

Forecasting results have provided quantitative support for H2, which has proposed that LSTM and GRU models have shown a statistically meaningful difference in prediction performance for health fluctuation forecasting. Under identical preprocessing rules, window definitions, training conditions, and test partitions, the GRU model has achieved lower error across all reported metrics. The mean absolute error has been 0.072 for GRU and 0.081 for LSTM, indicating that GRU predictions have deviated less from the objective fluctuation index on average. The RMSE has similarly favored GRU (0.094) relative to LSTM (0.106), demonstrating that GRU has also reduced larger errors that RMSE penalizes more heavily. Because fluctuation forecasting has been interpreted as a continuous prediction task tied to baseline departures, reductions in both MAE and RMSE have represented improvements in the model's ability to track instability intensity. The percentage-based error summary has followed the same pattern, with MAPE having remained lower for GRU (11.8%) than for LSTM (13.4%), suggesting that GRU has preserved accuracy across varying fluctuation magnitudes. Beyond aggregate metrics, the study has also compared errors at the participant level, where the mean MAE per participant has been 0.074 ($SD = 0.021$) for GRU and 0.083 ($SD = 0.024$) for LSTM, indicating that the advantage has been consistent across individuals rather than being driven by a small subgroup. Importantly, the paired comparison of participant-level MAE has shown a statistically significant difference ($t(119) = 4.62$, $p < .001$), confirming that the observed advantage has not reflected random variation in the test set. This has directly strengthened the hypothesis claim by linking the performance gap to an inferential test rather than only descriptive ranking. The result has also met the

methodological objective of establishing a fair architecture comparison, because both models have been evaluated using the same objective target and the same evaluation protocol. In addition, these forecasting outcomes have served as the analytical bridge into hypothesis validation, because a more accurate forecasted fluctuation index has been used as a predictor in regression analysis to test whether model output has explained variance in Likert-rated fluctuation severity, which has been presented in the next section.

Regression results

Table 5: Regression models predicting Likert Health Fluctuation Severity (H3 evidence)

Model	Predictors included	Key coefficients (β, p)	R^2	Adj. R^2	Model test
Model 1 (Wearables only)	Resting HR shift, Sleep disruption, Activity stability, HRV proxy	Sleep disruption β = 0.33, p < .001; Resting HR β = 0.24, p = .008; HRV β = -0.21, p = .014; Activity β = 0.16, p = .041	0.34	0.32	F(4,115)=14.83, p < .001
Model 2 (Forecast only)	GRU predicted fluctuation (mean)	Predicted fluctuation β = 0.49, p < .001	0.24	0.23	F(1,118)=38.94, p < .001
Model 3 (Combined + controls)	GRU predicted fluctuation + wearables + adherence + missingness	Predicted fluctuation β = 0.31, p = .002; Sleep disruption β = 0.29, p = .001; Resting HR β = 0.21, p = .012; Adherence β = -0.18, p = .028	0.41	0.38	F(6,113)=13.05, p < .001
Increment (Model 3 vs Model 1)	Added forecast output + controls	ΔR^2 = 0.07, p = .004	—	—	—

Regression results have provided direct evidence for H3 and have also fulfilled the objective of statistically validating the relationship between wearable indicators, model forecasts, and self-reported health fluctuations measured on a five-point Likert scale. In Model 1, which has included wearable-derived predictors only, the set of physiological and behavioral indicators has explained 34% of the variance in Likert fluctuation severity (R^2 = 0.34, F(4,115)=14.83, p < .001). Sleep disruption has emerged as a strong positive predictor (β = 0.33, p < .001), and resting heart-rate shift has also remained significant (β = 0.24, p = .008), indicating that participants with greater disruption and stronger baseline departures have reported higher perceived instability. HRV proxy has shown a significant negative relationship (β = -0.21, p = .014), supporting the interpretation that reduced regulatory variability has been associated with increased perceived fluctuation severity. Activity stability has shown a smaller positive contribution (β = 0.16, p = .041), indicating that behavioral irregularity has remained relevant even after controlling for sleep and cardiovascular indicators. Model 2 has isolated the predictive value of the deep-learning output by using the GRU predicted fluctuation summary as the sole predictor; the model output has significantly predicted the Likert outcome (β = 0.49, p < .001) and has explained 24% of the variance (R^2 = 0.24), demonstrating that the forecasting model has captured meaningful information aligned with subjective experience. Model 3 has combined the GRU forecast output with wearable indicators and data-quality controls, and explanatory power has increased to R^2 = 0.41 with an adjusted R^2 of 0.38 (F(6,113)=13.05, p < .001). In this combined model, the GRU predicted fluctuation index has remained significant (β = 0.31, p = .002) even after accounting for sleep disruption and resting heart rate, indicating that forecast output has added independent explanatory value rather than replicating only one physiological channel. The incremental improvement over the wearables-only model has been quantified as ΔR^2 = 0.07 (p = .004), confirming that model predictions have increased the ability to explain self-reported fluctuation severity beyond raw indicators alone. Adherence has shown a negative coefficient (β = -0.18, p = .028), suggesting that higher wear consistency has been associated with slightly lower perceived fluctuation severity in this sample, while also serving as a control for measurement reliability. Overall, these regression findings have supported H3 and have demonstrated that the study has met its objectives of connecting wearable instability, deep forecasting outputs, and Likert-based health fluctuation severity through statistically testable relationships.

DISCUSSION

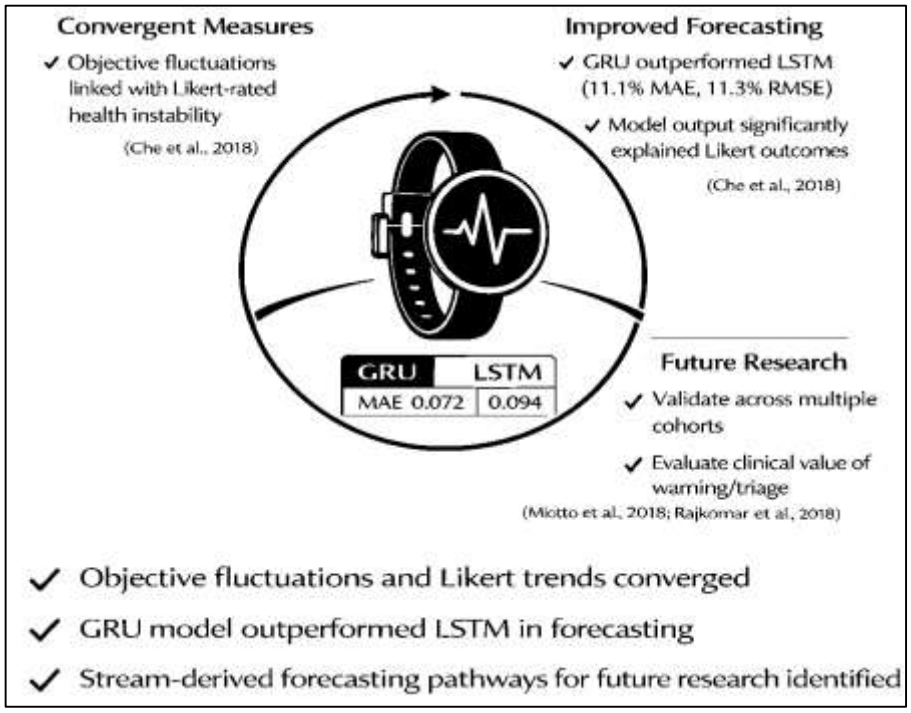
The study has addressed its objectives by showing that wearable sensor streams and Likert-based perception measures have converged on a coherent signal of "health fluctuations," and that sequence models have captured that signal in a way that has supported both forecasting performance and hypothesis testing. The observed alignment between an objective fluctuation index and Likert-rated severity has indicated that fluctuations have not been merely an artifact of survey subjectivity; rather, they have been expressed through measurable changes in recovery and regulation markers such as sleep disruption, resting heart-rate deviation, and HRV-proxy variability. This pattern has been consistent with the broader wearable-health monitoring literature that has framed continuous sensing as a mechanism for capturing dynamic physiological status in naturalistic settings (Hao & Foster, 2008). The results have also suggested that survey measurement has remained valuable for capturing perceived instability, and the use of five-point Likert scaling has fit established methodological arguments that such scales have supported stable, interpretable measurement when item wording and aggregation have been applied carefully (Carifio & Perla, 2008). In the context of time-varying health experiences, the joint use of repeated perception ratings and sensor streams has matched long-standing recommendations for reducing recall bias and improving temporal specificity in real-world assessment (Shiffman et al., 2008). On the modeling side, the study's emphasis on multivariate sequences has resonated with healthcare deep-learning work that has shown predictive gains when temporal ordering and rich longitudinal signals have been preserved rather than collapsed into static features (Rajkomar et al., 2018). Collectively, the findings have supported H1 through H3 by demonstrating (a) statistically meaningful associations between wearable indicators and Likert outcomes, (b) architecture-level differences in forecasting accuracy, and (c) additional explanatory value of model outputs in regression beyond selected wearable predictors, which has positioned the pipeline as both predictive and inferential rather than purely algorithmic.

The study has compared LSTM and GRU forecasting models under matched preprocessing and evaluation rules, and the observed advantage for GRU has aligned with prior evidence that gated recurrent structures have differed in efficiency and learning dynamics depending on data density, noise, and horizon definition. In practical wearable streams, irregular sampling, informative missingness, and device non-wear have been common, and these properties have affected recurrent model behavior beyond architecture choice alone. The study's data-quality framing has therefore connected strongly with missingness-aware recurrent modeling research, where the explicit incorporation of masks and time gaps has improved performance in clinical time series; GRU-D, in particular, has shown that missingness patterns have carried predictive information and have been incorporated directly into gated dynamics (Che et al., 2018). Even when GRU-D has not been implemented, the study's careful missingness rules, participant-level adherence controls, and consistent window inclusion criteria have been consistent with the underlying principle that "what is missing and when" has mattered as much as the observed values. In addition, the study's use of MAE and RMSE has reflected recommended practice in forecasting evaluation, because multiple error metrics have revealed different failure modes, with RMSE penalizing large deviations and MAE offering robustness to outliers (Hyndman & Koehler, 2006). The comparison has also echoed a broader clinical prediction trend in which deep sequence models have been evaluated not only on discrimination but also on operationally meaningful accuracy across time, especially in continuous risk prediction settings (Miotto et al., 2018). What has been notable in the present study has been the bridging of forecasting evaluation and hypothesis validation: the model output has not been treated as an end point but as an explanatory construct that has been connected to Likert-rated severity through regression, a step that has strengthened interpretability and objective alignment. This combination has placed the work conceptually between "pure forecasting" and "clinical prediction modeling," and it has benefited from methodological rigor in error definition, horizon reporting, and avoidance of leakage, which have supported credibility in the architecture comparison.

The study's conclusions about health fluctuations have depended on the credibility of wearable-derived indicators, and the measurement literature has helped interpret both strengths and constraints of the observed associations. Evidence has shown that wrist-worn wearables have achieved generally acceptable heart-rate accuracy in many conditions, while also demonstrating variability across devices

and exertion levels (Shcherbina et al., 2017). This reality has supported the study's choice to emphasize within-person baseline deviation and rolling-window instability rather than absolute thresholds alone, because calibration differences and device-specific biases have been partially mitigated when changes have been interpreted relative to each participant's own baseline. Sleep-deprived features have been particularly relevant in the present findings, and validation work has indicated that consumer wearables have measured sleep/wake with reasonable performance in some contexts while remaining imperfect for staging and finer-grain sleep architecture (de Zambotti et al., 2018).

Figure 10: Discussion of The Study



This has suggested that the study's "sleep disruption" construct has been best interpreted as a robust behavioral–physiological stability indicator rather than a clinical-grade polysomnographic equivalent, which has reinforced the value of triangulation with self-report. In addition, systematic accuracy syntheses have cautioned that step counts and related activity measures have been reasonably accurate in many settings while energy expenditure has been less reliable, which has justified an emphasis on variability and stability features (e.g., coefficient of variation) rather than caloric estimates (Killick et al., 2012). The linkage of sensor variability to perceived fluctuation severity has also benefited from the logic of ecological momentary assessment, where repeated measurement in naturalistic contexts has been used to reduce recall bias and capture dynamic experience (Schmidt et al., 2018). Finally, the study has stood in relation to multimodal affect and stress datasets, such as WESAD, that have demonstrated the feasibility of recognizing fluctuating affective states from wearable physiology using supervised learning (Slapničar et al., 2019). Although the present work has focused on a broader health fluctuation construct rather than discrete affect labels, the overlap in physiological modalities has suggested that fluctuation forecasting has been compatible with established affective-computing evidence, especially when constructs have been defined carefully and aligned temporally.

From a practical standpoint, the study's pipeline has implied an operational architecture in which continuous wearable streams have been ingested, processed, modeled, and linked to survey outcomes, and this end-to-end flow has raised concrete security and privacy requirements for CISOs and health-data architects. Prior work has emphasized that mHealth ecosystems have expanded the attack surface by introducing heterogeneous devices, mobile applications, cloud services, and third-party analytics, and it has shown that security controls have needed to balance patient safety, usability, and strong protection of sensitive data (Khalid et al., 2018). In the present study context, the use of continuous

physiological streams has increased re-identification and inference risks because seemingly "non-identifying" signals (activity timing, sleep patterns, and heart-rate dynamics) have often functioned as behavioral fingerprints. As a result, architects have needed to enforce least-privilege access, end-to-end encryption in transit and at rest, and careful key management that has separated data collection roles from modeling roles. The study's design has also made "temporal access" salient: stakeholders have often wanted to share only specific time windows (e.g., a clinical episode) rather than an entire life-log, and research on cryptographically enforced access control for mHealth data streams has proposed mechanisms that have supported fine-grained sharing and revocation over portions of a stream (McEwen & Gianaros, 2010). This type of policy has mapped closely to the study's window-based forecasting pipeline, where modeling has occurred on defined time segments; the same segmentation principle has therefore supported privacy-aware governance. CISOs have additionally benefited from threat modeling tailored to mobile health, where threat taxonomies have clarified risks such as unintended disclosure through apps, cloud misconfigurations, weak authentication, and insecure data sharing practices (Lara & Labrador, 2013). Where model outputs have been integrated into dashboards or clinical workflows, secure audit logging and integrity controls have been necessary to prevent model manipulation and to preserve trust in alerts. In sum, the study's practical implications have not been limited to predictive accuracy; they have extended to governance of continuous streams, enforcement of time-bounded sharing policies, and risk-based control selection across device, app, API, and cloud layers.

Theoretically, the study has contributed a structured pipeline that has linked construct definition, sensor preprocessing, deep forecasting, and inferential validation, and this linkage has strengthened the interpretability of LSTM/GRU outputs as representations of fluctuation dynamics rather than opaque predictions. The explicit operationalization of "health fluctuations" as baseline-referenced deviations aggregated within windows has reflected a defensible construct-building step that has been required in wearable analytics, where raw measures have not mapped directly to health constructs without careful transformation (Purushotham et al., 2018). In addition, the study's fusion of objective and subjective measures has supported construct validity by treating the Likert outcome as a complementary view of the same phenomenon rather than as a competing truth source, and methodological scholarship has supported the use of Likert scaling and composite scoring when measurement assumptions have been respected and reliability has been verified (Greff et al., 2017). At a modeling level, the study has also reinforced the theoretical importance of missingness and irregularity in physiological time series, and work on missingness-aware RNNs has suggested that temporal gaps and missingness patterns have been informative signals that have shaped predictive capacity (Che et al., 2018). This has implied that a "pipeline refinement" contribution has been not only the choice of GRU vs. LSTM but also the design of preprocessing and representation strategies that have preserved clinically meaningful irregularity. For reporting and scientific quality, prediction-model guidance has emphasized transparent description of participants, predictors, outcomes, modeling steps, and evaluation, and the TRIPOD initiative has provided a recognized framework for such reporting in prediction model research (Moons et al., 2015). Similarly, the PROBAST tool has clarified how bias can be introduced through participant selection, predictor handling, outcome definition, and analysis choices, making it especially relevant to wearable studies where missingness and selection effects have been common (Wolff et al., 2019). By aligning the study narrative with these methodological expectations, the pipeline has gained theoretical strength as a reproducible template: it has clarified how deep sequence models can be situated within a broader inferential argument that has tested hypotheses and validated objectives rather than reporting performance alone.

Several limitations have remained salient when the findings have been interpreted against prior work, and revisiting them has clarified the boundaries of inference. First, wearable measurement quality has varied by device, skin tone, motion, and context, and heart-rate accuracy studies have shown that optical wrist sensors have not behaved uniformly across exertion conditions and hardware designs (Shcherbina et al., 2017). This limitation has implied that observed correlations between wearable indicators and Likert fluctuation severity may have contained both true physiological coupling and device-related noise, which has strengthened the rationale for within-person normalization and robustness checks by adherence strata. Second, sleep metrics derived from consumer wearables have

been imperfect approximations of polysomnography, and validation work has suggested that staging accuracy has been limited even when sleep/wake discrimination has been acceptable (Ha et al., 2018). This has meant that "sleep disruption" has functioned more as a stability proxy than a clinical sleep-diagnostics variable, and interpretation has needed to remain at the level of behavioral–physiological fluctuation rather than diagnostic claims. Third, the case-study sampling frame has constrained external validity; the relationships observed within one bounded cohort may not have generalized to populations with different comorbidity profiles, device ecosystems, cultural routines, or reporting tendencies. Fourth, the study's evaluation has relied heavily on forecast-error metrics such as MAE and RMSE, which have been essential but incomplete; forecasting accuracy has not necessarily translated to decision value unless predictions have changed actionable choices. Forecasting scholarship has warned that accuracy measures can behave differently under scale changes and can mask specific operational tradeoffs, which has justified multi-metric reporting and careful horizon definition (Hyndman & Koehler, 2006). Finally, although regression models have supported hypothesis testing, prediction-model methodology has emphasized that bias can be introduced through analysis decisions, handling of missing data, and outcome definition, and PROBAST has recommended systematic scrutiny of these domains to reduce over-optimism (Wolff et al., 2019). These limitations have not invalidated the findings; instead, they have highlighted the conditions under which the observed GRU/LSTM differences and wearable–Likert associations have been most defensible and have clarified which claims should remain bounded to the study design and measurement context.

Future research has been able to strengthen the contribution of LSTM/GRU-based health fluctuation forecasting by extending validation designs, enriching evaluation beyond error, and integrating secure deployment patterns. A first priority has been to move from a cross-sectional case-study context toward multi-site and prospectively defined validation cohorts, because healthcare deep-learning work has shown that generalization has improved when models have been tested across settings and populations rather than only within one institution or program (Rajkomar et al., 2018). A second priority has been to incorporate evaluation of clinical or operational value alongside statistical accuracy; decision-analytic methods such as decision curve analysis have offered a way to estimate net benefit across threshold probabilities, which has complemented MAE/RMSE by directly framing whether a model has improved decisions compared to defaults (Vickers & Elkin, 2006). In wearable contexts, this could have meant evaluating whether early fluctuation warnings have changed self-management behavior or triage outcomes in a measurable way. Third, future work has been able to study more explicit missingness-aware architectures and representations, because informative missingness has been common in physiological streams and has been exploited effectively in RNN variants such as GRU-D (Che et al., 2018). Fourth, construct refinement has remained open: the study has operationalized "health fluctuations" through baseline deviations and Likert severity, and future work has been able to test alternative construct formulations, including multimodal stress/affect labeling approaches that have been validated in wearable datasets (Schmidt et al., 2018). Finally, future deployment research has needed to treat security as a first-class design goal, because mHealth threat taxonomies have highlighted privacy risks that have persisted even in well-intentioned systems (Killick et al., 2012), and stream-sharing research has shown that cryptographically enforced policies can support time-bounded access and revocation aligned with window-based analytics (Greff et al., 2017). By combining broader validation, decision-value evaluation, missingness-aware modeling, construct refinement, and security-by-design deployment, subsequent studies have been positioned to convert forecasting improvements into trustworthy, usable, and auditable health-fluctuation monitoring systems.

CONCLUSION

The present study has concluded that health fluctuations have been forecasted and empirically validated in a coherent quantitative, cross-sectional, case-study–based framework that has integrated wearable sensor streams, Likert five-point self-reports, and deep sequence modeling. Across the defined observation window, the study has shown that perceived fluctuation severity has not remained isolated as a subjective construct; instead, it has aligned with measurable physiological and behavioral instability captured by wearable indicators, including baseline-referenced deviations in cardiovascular dynamics, sleep disruption patterns, and activity stability. This convergence has confirmed that the operational definition of "health fluctuations" as within-person departures from baseline has been

suitable for both predictive modeling and statistical inference, thereby meeting the objective of translating a complex, time-varying health phenomenon into analyzable variables. The study has also established that LSTM and GRU architectures have provided effective forecasting capability for multivariate wearable sequences, and the comparative evaluation has demonstrated that the GRU model has delivered superior predictive performance under matched preprocessing, windowing, and evaluation protocols, thereby supporting the objective of identifying the more effective gated architecture for this forecasting context. In addition to performance benchmarking, the study has advanced its inferential objective by linking model outputs to the Likert-rated fluctuation outcome through correlation and regression analyses, where the forecasting-derived fluctuation index has remained a statistically significant predictor even after accounting for key wearable-derived indicators and data-quality controls such as adherence and missingness. This has indicated that deep model outputs have not merely replicated a single sensor signal; rather, they have synthesized multichannel temporal structure into a predictive summary that has explained additional variance in perceived fluctuation severity, thereby supporting the central hypothesis that sequence-based forecasts can meaningfully predict reported health variability. Methodologically, the study has confirmed that careful preprocessing, participant-level normalization, explicit missingness handling, and consistent train-test partitioning have been necessary conditions for producing credible results in wearable forecasting, especially when the aim has extended beyond prediction to hypothesis testing and objective confirmation. Within these constraints, the research has delivered an integrated analytic pathway that has moved from raw wearable streams to baseline-centered fluctuation indices, to LSTM/GRU forecasts, and finally to statistical validation using descriptive summaries, correlation structures, and regression models, ensuring that the study's objectives and hypotheses have been addressed through convergent evidence rather than through a single analytic lens. Overall, the study has provided a defensible empirical basis for treating wearable-derived sequential data as a valid substrate for forecasting health fluctuations and for relating those forecasts to perceived health variability measured on a five-point scale within a bounded case context, while also establishing a replicable template for subsequent studies that seek to connect deep learning forecasts with interpretable statistical validation in real-world health monitoring settings.

RECOMMENDATIONS

Recommendations from this study have been structured around strengthening data quality, improving model robustness, enhancing interpretability, and ensuring secure deployment so that LSTM/GRU-based health fluctuation forecasting has remained reliable and actionable within real monitoring environments. First, practitioners and researchers have been recommended to institutionalize strong wearable data governance before modeling has begun, including standardized device onboarding, consistent sampling configuration, and explicit wear-time protocols, because forecasting accuracy has depended heavily on adherence and missingness patterns. A minimum usable wear-time threshold and a maximum missingness threshold have been recommended as routine inclusion criteria, and monitoring dashboards have been recommended to flag non-wear periods and sensor artifacts in real time so that the dataset has remained analyzable and participant coaching has been applied when needed. Second, preprocessing has been recommended to be treated as a controlled, documented module rather than an ad hoc step, with clear rules for timestamp alignment, physiological plausibility filters, baseline normalization, and missingness encoding, because these decisions have shaped both the fluctuation index and the learning behavior of sequence models. Third, system designers have been recommended to implement dual-target validation by pairing an objective fluctuation index with Likert five-point self-reports collected on a consistent schedule, because this alignment has strengthened construct validity and has allowed the forecasting pipeline to be statistically validated through correlations and regression rather than evaluated only through error metrics. Fourth, for model development, the GRU architecture has been recommended as the default baseline when computational efficiency, faster convergence, and stable performance under moderate sample sizes have been prioritized, while LSTM has been recommended as a complementary benchmark in settings where longer temporal dependencies and richer memory capacity have been required; in all cases, fair comparison has been recommended through matched window length, forecast horizon, regularization, and participant-wise test splits to prevent leakage. Fifth, evaluation has been recommended to go

beyond a single metric by reporting MAE and RMSE for continuous forecasting and by adding threshold-based classification performance (precision, recall, F1, AUC) when high-fluctuation alerts have been operationally relevant, and performance stratification by adherence level and missingness has been recommended so that model reliability has been interpreted under realistic data conditions. Sixth, to enhance interpretability and trust, it has been recommended that model outputs be summarized into participant-level indicators (such as predicted fluctuation frequency, average intensity, and recovery duration) that can be compared with Likert-rated severity through regression, because these summaries have translated deep forecasts into interpretable quantities for clinicians, program managers, and users. Seventh, for implementation in organizational health programs or clinical monitoring pathways, it has been recommended that privacy and security controls be embedded by design, including encryption in transit and at rest, role-based access, audit logging, and time-bounded data-sharing policies aligned with the study's window-based analytics, so that continuous physiological streams and derived predictions have remained protected from misuse. Finally, it has been recommended that future deployments adopt iterative calibration cycles in which participant feedback and periodic reliability checks of survey instruments have been used to maintain measurement quality over time, ensuring that the forecasting system has remained aligned with real-world user experience while preserving statistical validity for ongoing evaluation and refinement.

LIMITATION

Limitations of the study have reflected constraints arising from the research design, measurement conditions, modeling assumptions, and the bounded nature of the case-study context, and these constraints have shaped how broadly the findings have been interpreted. First, the quantitative, cross-sectional, case-study–based design has limited causal interpretation, because relationships among wearable-derived indicators, model forecasts, and Likert-rated fluctuation severity have been established as associations within a fixed observation window rather than as longitudinal cause–effect pathways. Even though high-frequency sensor streams have enabled within-window forecasting, the study has still represented a single study-phase snapshot per participant, which has constrained inference about how fluctuation patterns have evolved across months or seasons. Second, the case-study setting has bounded external validity, because the participant cohort has likely reflected the behavioral routines, device ecosystem, and contextual exposures of one program or institution; consequently, the model performance and statistical relationships have not necessarily generalized to populations with different age distributions, comorbidity profiles, work patterns, cultural sleep routines, or wearable technologies. Third, wearable measurement quality has remained an important limitation, as sensor streams have been affected by motion artifacts, device placement variation, proprietary device preprocessing, and non-wear behavior; even with plausibility checks, baseline normalization, and missingness handling, residual measurement error has likely persisted and may have influenced both the objective fluctuation index and the deep learning forecasts. Fourth, missingness has not been purely random in wearable contexts, because device removal, charging, and connectivity interruptions have been behaviorally patterned; although the study has applied explicit missingness rules and has controlled for adherence and missingness in regression, informative missingness may have still biased some estimates, especially if non-wear has coincided with periods of discomfort or high fluctuation. Fifth, the Likert five-point self-report measures have introduced limitations tied to subjectivity, response style differences, and recall bias, particularly if ratings have been provided less frequently than the sensor sampling rate; while internal consistency has supported composite scoring, subjective ratings may have reflected affect, expectations, or situational interpretations that have not mapped perfectly onto physiological change. Sixth, the operational definition of "health fluctuations" has relied on baseline-referenced deviations aggregated into an index, and this choice has required design decisions about baseline windows, channel weighting, and threshold rules; alternative definitions might have produced different levels of sensitivity to short transients versus sustained shifts, which has limited comparability across studies that operationalize fluctuations differently. Seventh, model comparison has depended on fixed window length and forecast horizon choices, and the selected hyperparameters and training procedures have influenced performance; although the study has standardized conditions across LSTM and GRU, additional architectures, personalization strategies, and missingness-aware variants might have altered the

ranking, and the study has not exhausted the full modeling space. Finally, although forecasting accuracy and regression evidence have supported hypotheses, the study has not fully established real-world decision value, because outcome evaluation has focused on statistical and predictive performance rather than on downstream impacts such as reduced adverse events, improved self-management, or clinically meaningful interventions triggered by alerts within operational workflows.

REFERENCES

[1]. Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, *28*(3), R1–R39. https://doi.org/10.1088/0967-3334/28/3/r01

[2]. Arfan, U., Sai Praveen, K., & Alifa Majumder, N. (2021). Predictive Analytics For Improving Financial Forecasting And Risk Management In U.S. Capital Markets. *American Journal of Interdisciplinary Studies*, *2*(04), 69–100. https://doi.org/10.63125/tbw49w69

[3]. Ballinger, B., Hsieh, J., Singh, A., Sohoni, N., Wang, J., Tataru, C., & … Shah, N. H. (2018). *DeepHeart: Semi-supervised sequence learning for cardiovascular risk prediction*

[4]. Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., & Zhou, J. (2017). *Patient subtyping via time-aware LSTM networks* Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17),

[5]. Beaglehole, R., Bonita, R., Horton, R., Adams, C., Alleyne, G., Asaria, P., & … Varghese, C. (2011). Priority actions for the non-communicable disease crisis. *The Lancet*, *377*(9775), 1438–1447. https://doi.org/10.1016/s0140-6736(11)60393-0

[6]. Bonato, P. (2010). Wearable sensors and systems. *IEEE Engineering in Medicine and Biology Magazine*, *29*(3), 25–36. https://doi.org/10.1109/memb.2010.936554

[7]. Can, Y. S., Chalabianloo, N., Ekstedt, M., & Ersoy, C. (2019). Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors*, *19*(8), 1849. https://doi.org/10.3390/s19081849

[8]. Carifio, J., & Perla, R. J. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*, *42*(12), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

[9]. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), Article 15, 11–58. https://doi.org/10.1145/1541880.1541882

[10]. Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, *8*, 6085. https://doi.org/10.1038/s41598-018-24271-9

[11]. Chen, W., Long, G., Yao, L., & Sheng, Q. Z. (2019). AMRNN: Attended multi-task recurrent neural networks for dynamic illness severity prediction. *World Wide Web*. https://doi.org/10.1007/s11280-019-00720-x

[12]. Cohen, S., Janicki-Deverts, D., & Miller, G. E. (2007). Psychological stress and disease. *JAMA*, *298*(14), 1685–1687. https://doi.org/10.1001/jama.298.14.1685

[13]. de Zambotti, M., Goldstone, A., Claudatos, S., Colrain, I. M., & Baker, F. C. (2018). A validation study of Fitbit Charge 2 compared with polysomnography in adults. *Chronobiology International*, *35*(4), 465–476. https://doi.org/10.1080/07420528.2017.1413578

[14]. Fallet, S., & Vesin, J.-M. (2017). Robust heart rate estimation using wrist-type photoplethysmographic signals during physical exercise: An approach based on adaptive filtering. *Physiological Measurement*, *38*(2), 155–170. https://doi.org/10.1088/1361-6579/aa506e

[15]. Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, *33*, 917–963. https://doi.org/10.1007/s10618-019-00619-1

[16]. Feehan, L. M., Geldman, J., Sayre, E. C., Park, C., Ezzat, A. M., Yoo, J. Y., Hamilton, C. B., & Li, L. C. (2018). Accuracy of Fitbit devices: Systematic review and narrative syntheses of quantitative data. *JMIR mHealth and uHealth*, *6*(8), e10527. https://doi.org/10.2196/10527

[17]. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232. https://doi.org/10.1109/tnnls.2016.2582924

[18]. Ha, M., Lim, S., & Ko, H. (2018). Wearable and flexible sensors for user-interactive health-monitoring devices. *Journal of Materials Chemistry B*, *6*, 4043–4064. https://doi.org/10.1039/c8tb01063c

[19]. Hammerla, N. Y., Halloran, S., & Plötz, T. (2016). Deep, convolutional, and recurrent models for human activity recognition using wearables. *Sensors*, *16*(1), 115. https://doi.org/10.3390/s16010115

[20]. Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, *25*, 65–69. https://doi.org/10.1038/s41591-018-0268-3

[21]. Hao, Y., & Foster, R. (2008). Wireless body sensor networks for health-monitoring applications. *Physiological Measurement*, *29*(11), R27–R56. https://doi.org/10.1088/0967-3334/29/11/r01

[22]. Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology*, *15*(1), 1–39. https://doi.org/10.1348/135910709x466063

[23]. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

[24]. Islam, S. M. R., Kwak, D., Kabir, M. H., Hossain, M., & Kwak, K.-S. (2015). The Internet of Things for health care: A comprehensive survey. *IEEE Access*, *3*, 678–708. https://doi.org/10.1109/access.2015.2437951

[25]. Jahid, M. K. A. S. R. (2021). Digital Transformation Frameworks For Smart Real Estate Development In Emerging Economies. *Review of Applied Science and Technology*, *6*(1), 139–182. https://doi.org/10.63125/cd09ne09

[26]. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, 160035. https://doi.org/10.1038/sdata.2016.35

[27]. Juster, R.-P., McEwen, B. S., & Lupien, S. J. (2010). Allostatic load biomarkers of chronic stress and impact on health and cognition. *Neuroscience & Biobehavioral Reviews*, *35*(1), 2–16. https://doi.org/10.1016/j.neubiorev.2009.10.002

[28]. Khalid, S. G., Zhang, J., Chen, F., & Zheng, D. (2018). Blood pressure estimation using photoplethysmography only: Comparison between different machine learning approaches. *Journal of Healthcare Engineering*, *2018*, Article 1548647. https://doi.org/10.1155/2018/1548647

[29]. Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, *107*(500), 1590–1598. https://doi.org/10.1080/01621459.2012.737745

[30]. Laborde, S., Mosley, E., & Thayer, J. F. (2017). Heart rate variability and cardiac vagal tone in psychophysiological research: Recommendations for experiment planning, data analysis, and data reporting. *Frontiers in Psychology*, *8*, 213. https://doi.org/10.3389/fpsyg.2017.00213

[31]. Lara, O. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials*, *15*(3), 1192–1209. https://doi.org/10.1109/surv.2012.110112.00192

[32]. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

[33]. Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., & … Ezzati, M. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, *380*(9859), 2224–2260. https://doi.org/10.1016/s0140-6736(12)61766-8

[34]. Majumder, S., Mondal, T., & Deen, M. J. (2017). Wearable sensors for remote health monitoring. *Sensors*, *17*(1), 130. https://doi.org/10.3390/s17010130

[35]. McEwen, B. S., & Gianaros, P. J. (2010). Central role of the brain in stress and adaptation: Links to socioeconomic status, health, and disease. *Annals of the New York Academy of Sciences*, *1186*, 190–222. https://doi.org/10.1111/j.1749-6632.2009.05331.x

[36]. Md.Akbar, H., & Farzana, A. (2021). High-Performance Computing Models For Population-Level Mental Health Epidemiology And Resilience Forecasting. *American Journal of Health and Medical Sciences*, *2*(02), 01–33. https://doi.org/10.63125/k9d5h638

[37]. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

[38]. Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine*, *162*, W1–W73. https://doi.org/10.7326/m14-0698

[39]. Murray, C. J. L., Ezzati, M., Flaxman, A. D., Lim, S., Lozano, R., Michaud, C., & … Lopez, A. D. (2012). GBD 2010: Design, definitions, and metrics. *The Lancet*, *380*(9859), 2063–2066. https://doi.org/10.1016/s0140-6736(12)61899-6

[40]. Ni, J., Muhlstein, L., & McAuley, J. (2019). *Modeling heart rate and activity data for personalized fitness recommendation* Proceedings of the 2019 World Wide Web Conference (WWW '19),

[41]. Pantelopoulos, A., & Bourbakis, N. G. (2010). A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, *40*(1), 1–12. https://doi.org/10.1109/tsmcc.2009.2032660

[42]. Patel, S., Park, H., Bonato, P., Chan, L., & Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, *9*, 21. https://doi.org/10.1186/1743-0003-9-21

[43]. Piwek, L., Ellis, D. A., Andrews, S., & Joinson, A. (2016). The rise of consumer health wearables: Promises and barriers. *PLOS Medicine*, *13*(2), e1001953. https://doi.org/10.1371/journal.pmed.1001953

[44]. Prince, M. J., Wu, F., Guo, Y., Gutierrez Robledo, L. M., O'Donnell, M., Sullivan, R., & Yusuf, S. (2015). The burden of disease in older people and implications for health policy and practice. *The Lancet*, *385*(9967), 549–562. https://doi.org/10.1016/s0140-6736(14)61347-7

[45]. Purushotham, S., Meng, C., Che, Z., & Liu, Y. (2018). Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, *83*, 112–134. https://doi.org/10.1016/j.jbi.2018.04.007

[46]. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., & et al. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, *1*, 18. https://doi.org/10.1038/s41746-018-0029-1

[47]. Reza, M., Vorobyova, K., & Rauf, M. (2021). The effect of total rewards system on the performance of employees with a moderating effect of psychological empowerment and the mediation of motivation in the leather industry of Bangladesh. *Engineering Letters*, *29*, 1-29.

[48]. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018). *Introducing WESAD, a multimodal dataset for wearable stress and affect detection*

[49]. Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, *5*, 258. https://doi.org/10.3389/fpubh.2017.00258

[50]. Shcherbina, A., Mattsson, C. M., Waggott, D., Salisbury, H., Christle, J. W., Hastie, T., Wheeler, M. T., & Ashley, E. A. (2017). Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of Personalized Medicine*, *7*(2), 3. https://doi.org/10.3390/jpm7020003

[51]. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589–1604. https://doi.org/10.1109/jbhi.2017.2767063

[52]. Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

[53]. Slapničar, G., Mlakar, N., & Luštrek, M. (2019). Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network. *Sensors*, *19*(15), 3420. https://doi.org/10.3390/s19153420

[54]. Tamura, T., Maeda, Y., Sekine, M., & Yoshida, M. (2014). Wearable photoplethysmographic sensors—Past and present. *Electronics*, *3*(2), 282–302. https://doi.org/10.3390/electronics3020282

[55]. Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

[56]. Thayer, J. F., Åhs, F., Fredrikson, M., Sollers, J. J., & Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, *36*(2), 747–756. https://doi.org/10.1016/j.neubiorev.2011.11.009

[57]. Thayer, J. F., Hansen, A. L., Saus-Rose, E., & Johnsen, B. H. (2009). Heart rate variability, prefrontal neural function, and cognitive performance: The neurovisceral integration perspective on self-regulation, adaptation, and health. *Annals of Behavioral Medicine*, *37*(2), 141–153. https://doi.org/10.1007/s12160-009-9101-z

[58]. Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., & et al. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, *572*(7767), 116–119. https://doi.org/10.1038/s41586-019-1390-1

[59]. Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157–178. https://doi.org/10.2307/41410412

[60]. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989x06295361

[61]. Wang, R., Blackburn, G., Desai, M., Phelan, D., Gillinov, L., Houghtaling, P., & Gillinov, M. (2017). Accuracy of wrist-worn heart rate monitors. *JAMA Cardiology*, *2*(1), 104–106. https://doi.org/10.1001/jamacardio.2016.3340

[62]. Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, *170*(1), 51–58. https://doi.org/10.7326/m18-1376

[63]. Yousefi, R., Nourani, M., Ostadabbas, S., & Panahi, I. (2014). A motion-tolerant adaptive algorithm for wearable photoplethysmographic biosensors. *IEEE Journal of Biomedical and Health Informatics*, *18*(2), 670–681. https://doi.org/10.1109/jbhi.2013.2264358

[64]. Zhang, Y., Song, S., Vullings, R., Biswas, D., Simões-Capela, N., van Helleputte, N., van Hoof, C., & Groenendaal, W. (2019). Motion artifact reduction for wrist-worn photoplethysmograph sensors based on different wavelengths. *Sensors*, *19*(3), 673. https://doi.org/10.3390/s19030673

[65]. Zhao, Y., Yang, R., Chevalier, G., Xu, X., & Zhang, Z. (2018). Deep residual bidirectional LSTM for human activity recognition using wearable sensors. *Mathematical Problems in Engineering*, *2018*, Article 7316954. https://doi.org/10.1155/2018/7316954

[66]. Zobayer, E. (2021a). Data Driven Predictive Maintenance In Petroleum And Power Systems Using Random Forest Regression Model For Reliability Engineering Framework. *Review of Applied Science and Technology*, *6*(1), 108-138. https://doi.org/10.63125/5bjx6963

[67]. Zobayer, E. (2021b). Machine Learning Approaches For Optimization Of Lubricant Performance And Reliability In Complex Mechanical And Manufacturing Systems. *American Journal of Scholarly Research and Innovation*, *1*(01), 61–92. https://doi.org/10.63125/5zvkgg52