

Predictive AI for Proactive Risk Detection in Future Digital Home Care Systems

UNIVERSITY OF TURKU
Department of Computing
Master of Science in Technology Thesis
Master's Degree Programme in Health Technology
May 2026
Rimsha Kaokab

Supervisors:
Tuukka Panula
Johannes Holvitie (Oiva Health)

UNIVERSITY OF TURKU
Department of Computing

RIMSHA KAOKAB: Predictive AI for Proactive Risk Detection in Future Digital
Home Care Systems

Master of Science in Technology Thesis, 80 p., 29 app. p.
Master's Degree Programme in Health Technology
May 2026

Falls and acute health deterioration in elderly care populations are often preceded by gradual, measurable changes in daily activity and sleep behaviour that unfold over days to weeks before formal system alerts are generated. Commercial wearable monitoring systems detect deterioration reactively, triggering alerts only after pre-defined thresholds are crossed. This thesis investigates whether machine learning applied to wrist-worn sensor data can provide earlier, more interpretable risk detection, addressing two questions: whether a model can detect if a user is currently in a behavioural deterioration state and how early (RQ1), and whether a model can predict a critical adverse event within a defined horizon (RQ2).

Using pseudonymised real-world data from 20 elderly care users monitored for approximately seven months via the Vivago[®] wrist-worn activity device, a two-component framework was developed and evaluated under leave-one-user-out cross-validation in collaboration with Oiva Health. RQ1 is addressed through a state classification model (Task G) using change-point-derived labels; RQ2 through an event prediction model (Task C) using 14-day forward-window labels targeting red well-being flags. Task G achieves PR-AUC = 0.829, identifying deteriorating users a mean of 56.0 days before yellow well-being flags, with statistically significant separation from stable users (Mann-Whitney $p = 0.029$). Task C achieves PR-AUC = 0.975 (± 0.020), recall = 0.885, and precision = 0.775, providing exactly 14 days of advance warning before confirmed red flags for all five deteriorating users with zero false alerts for stable users. SHAP analysis reveals complementary feature profiles; Task G is driven by 14-day mean circadian amplitude and sleep duration, while Task C by 14-day downward slopes in daytime activity.

The study demonstrates that label design determines temporal detection behaviour, state-based labels unlock 56 days of lead time compared to zero days under event-window labels using identical data and infrastructure, providing a transferable contribution for proactive well-being risk detection in digital home care.

Keywords: proactive well-being risk detection, wearable activity monitoring, leave-one-user-out cross-validation, state-based labels, SHAP explainability, elderly home care, Vivago

Contents

1	Introduction	1
1.1	Background	1
1.2	Research Gap	2
1.3	Research Objectives and Scope	3
1.4	Contributions	4
1.5	Structure of the Thesis	5
	Use of Artificial Intelligence Tools	6
2	Related Work	7
2.1	Well-being Deterioration and falls in Elderly Populations	7
2.2	Wearable Sensors in Health Monitoring	8
2.3	From Detection to Prediction	10
2.4	Machine Learning for well-being Risk Prediction	11
2.5	Explainability and Clinical Interpretability	13
2.6	Industry and Real-World Deployment Constraints	14
2.7	Summary	16
3	Methodology	17
3.1	Research Design	17
3.2	Data Source and Overview	20
3.2.1	The Vivago Monitoring System	20

3.2.2	Dataset Characteristics	21
3.3	Ethical and Privacy Considerations	25
3.4	Data Preprocessing	25
3.4.1	Alarm and Flag Cleaning	25
3.4.2	Daily Aggregation	26
3.5	Feature Engineering	27
3.5.1	Per-User Baseline Normalisation	27
3.5.2	Rolling Features	27
3.5.3	Circadian Rhythm Metrics	29
3.5.4	Feature Matrix	29
3.6	Prediction Horizon	31
3.7	Label Generation	32
3.7.1	Event-Window Labels (Tasks A–F)	32
3.7.2	State-Based Labels (Task G)	33
3.8	Model Development	35
3.8.1	Model Selection	35
3.8.2	Hyperparameter Tuning	36
3.9	Evaluation Strategy	37
3.9.1	Cross-Validation Protocol	37
3.9.2	Evaluation Metrics	37
3.9.3	Lead Time Validation	38
3.9.4	SHAP Explainability Analysis	39
3.10	Proof-of-Concept Application	39
4	Results	42
4.1	Exploratory Data Analysis	42
4.1.1	Dataset Overview and State Classification	42
4.1.2	Prediction Horizon Analysis	43

4.1.3	Per-User Baseline Normalisation Validation	44
4.1.4	Feature Variability Analysis	45
4.2	Exploratory Label Analysis and Baseline Comparators	46
4.2.1	Label Distribution Across Prediction Tasks	46
4.2.2	Exploratory Task Performance (Tasks A, B, D, E, F)	48
4.2.3	Label Design Observation	49
4.3	RQ2 – Event Prediction (Task C)	50
4.3.1	LOUO Cross-Validation Results	50
4.3.2	Threshold Analysis	50
4.3.3	SHAP Feature Importance	51
4.3.4	Risk Score Trajectories	53
4.3.5	Lead Time Validation	55
4.3.6	Negative Control Validation	56
4.4	RQ1 – Behavioural State Detection (Task G)	57
4.4.1	Change-Point Detection Results	57
4.4.2	LOUO Cross-Validation Results	59
4.4.3	Threshold Analysis	59
4.4.4	SHAP Feature Importance	60
4.4.5	Lead Time Analysis	62
4.4.6	Post-Baseline Validation	64
5	Discussion	66
5.1	Label Design	66
5.2	Task C Performance	67
5.3	Task G Performance	69
5.4	Comparison with Existing Literature	71
5.5	Industry Use	72
5.6	Limitations	74

5.7 Future Work	76
6 Conclusion	78
References	81
Appendices	
A Supplementary Model Performance Figures	A-1
B Full Model Input Feature List	B-1
C Pipeline Reproducibility Notes	C-1
D Proof-of-Concept Dashboard Interface	D-1

List of Figures

3.1	Overview of the machine learning pipeline for proactive wellbeing risk detection	19
3.2	Per-user baseline daytime activity.	22
3.3	Per-user daytime activity timelines.	23
3.4	Flagged User Trajectories	24
3.5	Activity level distribution by Vivago sensor state.	26
3.6	Hierarchical clustering of feature correlations for the top 25 features by variance.	31
4.1	Calendar heatmap of automatic inactivity alarm events.	43
4.2	Mann-Whitney comparison of normalised daytime activity z-scores for pre-alarm versus non-alarm user-days across three prediction horizons.	44
4.3	Daytime activity z-score distributions after per-user normalisation. . .	45
4.4	Label class distribution across Tasks A through F.	47
4.5	Task G state label structure.	47
4.6	Precision, recall and F1 vs decision threshold for the Task C XGBoost model (LOUO out-of-fold predictions).	51
4.7	Mean absolute SHAP values for the top 15 Task C features.	52
4.8	SHAP beeswarm plot for Task C.	52
4.9	Task C risk trajectory for User_003.	54

4.10	Task C risk trajectory for User_011.	54
4.11	Task C risk trajectory for User_012.	54
4.12	Task C risk trajectory for User_014.	55
4.13	Task C risk trajectory for User_020.	55
4.14	Task C risk trajectories for the three confirmed alarm-free stable users (negative control).	57
4.15	Change-point detection validation for all five deteriorating users. . . .	58
4.16	Precision, recall and F1 vs decision threshold for the Task G XGBoost model (LOUO out-of-fold predictions).	60
4.17	Mean absolute SHAP values for the top 15 Task G features.	61
4.18	SHAP beeswarm for Task G.	62
4.19	Task G risk trajectory for User_003.	63
4.20	Task G risk trajectory for User_011.	63
4.21	Task G risk trajectory for User_012.	64
4.22	Task G risk trajectory for User_020.	64
A.1	Precision-recall curve for Logistic Regression on Task A.	A-2
A.2	Precision-recall curve for Random Forest on Task A.	A-3
A.3	Precision-recall curve for XGBoost (tuned) on Task A.	A-4
A.4	Precision-recall curve for Logistic Regression on Task B.	A-5
A.5	Precision-recall curve for Random Forest on Task B.	A-6
A.6	Precision-recall curve for XGBoost (tuned) on Task B.	A-7
A.7	Precision-recall curve for Logistic Regression on Task D.	A-8
A.8	Precision-recall curve for Random Forest on Task D.	A-9
A.9	Precision-recall curve for XGBoost (tuned) on Task D.	A-10
A.10	Precision-recall curve for Logistic Regression on Task E.	A-11
A.11	Precision-recall curve for Random Forest on Task E.	A-12
A.12	Precision-recall curve for XGBoost (tuned) on Task E.	A-13

A.13	Precision-recall curve for Logistic Regression on Task F.	A-14
A.14	Precision-recall curve for Random Forest on Task F.	A-15
A.15	Precision-recall curve for XGBoost (tuned) on Task F.	A-16
D.1	Proof-of-concept dashboard showing the stable user view (User_016).	D-3
D.2	Proof-of-concept dashboard showing the deteriorating user view (User_003).	D-4
D.3	Proof-of-concept dashboard showing the SHAP feature explanation panel for User_003.	D-4

List of Tables

3.1 Cohort baseline feature means computed over the first 30 observation days per user.	21
3.2 Event-window prediction tasks and label definitions.	32
4.1 Cohort baseline means computed over the first 30 observation days per user.	42
4.2 Mann-Whitney U test results for pre-alarm vs non-alarm daytime activity z-scores across candidate prediction horizons.	43
4.3 Daytime activity z-score summary by user group after per-user normalisation.	45
4.4 Top 15 features by variability score. CV is used for level features; standard deviation is used for zero-centred features.	46
4.5 Label distribution across all seven prediction tasks (3,265 training-eligible rows).	47
4.6 LOUO CV results for all three models across all six event-window prediction tasks.	48
4.7 Task C XGBoost threshold analysis at threshold = 0.4. Metrics computed on pooled LOUO out-of-fold predictions across all five evaluable users.	50
4.8 Task C lead time analysis. All methods were restricted to post-baseline dates (day_number > 30).	55

4.9	Task C vs the first red flag: per-user lead times at threshold 0.4. . . .	56
4.10	Detected change-point dates and lead times before yellow well-being flags.	59
4.11	Task G XGBoost LOUO CV results.	59
4.12	Task G lead time results (post-baseline dates only).	63
4.13	Task G post-baseline validation: risk score comparison (days 31–60 and 31–90).	64
4.14	Individual Task G risk scores during days 31–60 (post-baseline). . . .	65
B.1	Complete list of 49 model input features.	B-1
C.1	Key software library versions used in the pipeline.	C-1
C.2	Pipeline notebook execution order and primary outputs.	C-3

1 Introduction

1.1 Background

Falls and acute health deterioration represent one of the most pressing public health challenges among older adults. Globally, approximately one in four adults aged 65 years or older experiences at least one fall each year, with incidence rates exceeding 50% in institutional and residential care settings [1], [2]. The consequences extend well beyond immediate physical injury: falls are a leading cause of hospitalisation, functional decline, loss of independence, and mortality in elderly populations [3], [4]. In Finland, where an ageing population places increasing demand on home care and supported living services, the early identification of residents at elevated risk of deterioration or adverse events is a recognised priority for digital health infrastructure.

The difficulty is not detecting falls; it is seeing what comes before them. In the weeks and months preceding a deterioration alert, activity levels shift, sleep patterns fragment, and circadian rhythms weaken[5], [6]. These changes are real and measurable, but they unfold slowly enough that routine clinical observation misses them. A care worker checking on a resident twice a day cannot track a two-count-per-minute weekly decline in daytime activity. Continuous wrist-worn accelerometry can do so without requiring the user to do anything differently and without adding to staff workload[7], [8].

The Vivago[®] wrist-worn activity monitor, deployed in Finnish care settings in collaboration with Oiva Health, exemplifies this class of device. It provides continuous minute-level activity and sleep data alongside a built-in alert system that flags sustained changes in a user’s daytime activity level or circadian rhythm regularity. These yellow and red well-being alerts, called flags, represent the system’s formalised detection of deterioration. However, the system is by design reactive, where alerts are generated only once an event has already occurred, at which point clinical intervention may be urgently required, but preventive action is no longer possible. The gap between detectable precursor signal and the system-generated alert represents an untapped opportunity for earlier, more actionable risk identification.

1.2 Research Gap

There is a gap between research promise and clinical reality in predictive elderly care[9], [10]. Most published work relies on simulated falls, synthetic datasets, or controlled laboratory conditions, none of which captures what a real care environment looks like, where devices are occasionally removed, users have irregular routines, and adverse events are so rare that standard supervised learning approaches struggle without careful design [11], [12]. Studies that do use longitudinal real-world data face a different set of constraints: short observation windows, small cohorts, and labels that are difficult to define, let alone obtain consistently [13], [14].

A further underexplored dimension is the label design problem. Most predictive approaches frame the target as an upcoming event, predicting whether a fall or alarm will occur within some future time window [15], [16]. This framing is operationally intuitive but methodologically limiting; it suppresses early signal by labelling pre-event days as negative even when the user’s behavioural profile is already diverging from their personal norm. An alternative formulation like classifying whether a user is currently in a deterioration state, rather than predicting an imminent discrete

event, has received comparatively little attention in the wearable health monitoring literature.

This thesis directly addresses these gaps using pseudonymised longitudinal data collected from the Vivago system across 20 elderly care users over a continuous, approximately seven-month observation period. Rather than relying on simulated scenarios, the study works entirely within the constraints of real-world deployment, a heterogeneous cohort with individual differences in baseline activity, incomplete device wear, and sparse clinical annotation. The collaboration with Oiva Health provides both the data and a concrete deployment context that grounds every methodological decision.

1.3 Research Objectives and Scope

The primary objective of this thesis is to design, implement, and evaluate a proof-of-concept machine learning pipeline for proactive well-being risk detection using wrist-worn activity sensor data from the Vivago device. The study addresses two complementary and operationally distinct research questions.

Research Question 1 (RQ1): Can a machine learning model detect whether an elderly care user is currently in a behavioural deterioration state, and how early can this state be identified relative to formal system-generated alerts?

Research Question 2 (RQ2): Can a machine learning model predict the occurrence of a critical adverse event within a defined prediction horizon, and which label design and temporal window are optimal for this task in a real-world care context?

In the following sections, these two experimental setups are referred to as Task G (state-based detection aligned with RQ1) and Task C (event-window-based prediction aligned with RQ2). RQ1 is addressed through a state classification approach in which change-point detection is used to identify the onset of sustained

behavioural decline for each deteriorating user. Days from the detected change-point onward are labelled positive, irrespective of when a subsequent system event occurs. This design removes the forward-window constraint that suppresses early signal in event-based labelling schemes. RQ2 is addressed through a systematic evaluation of six event-window label configurations spanning three prediction horizons and three event types, including automatic inactivity alarms, formal red well-being flags, and combined adverse event labels incorporating manual alarm activations.

The scope of the study is explicitly bounded by the available data and the proof-of-concept nature of the work. Deep learning approaches were excluded on the basis of the small cohort size, which precludes robust training of sequence models such as long short-term memory networks or transformers. Synthetic data augmentation was similarly excluded to preserve the ecological validity of the findings. Three classical machine learning models, namely Logistic Regression, Random Forest, and XGBoost were evaluated under leave-one-user-out (LOUO) cross-validation, the most rigorous evaluation protocol available for this cohort structure. The study does not aim to develop a clinically validated medical device or to establish generalisability beyond this cohort.

1.4 Contributions

This thesis makes five contributions based on two experimental setups: Task G (RQ1) and Task C (RQ2).

A complete, reproducible machine learning pipeline was built and evaluated for proactive well-being risk detection using real Vivago wrist sensor data, covering preprocessing, per-user normalisation, feature engineering, label generation across seven tasks, model training, SHapley Additive exPlanations (SHAP) explainability, and lead time validation.

The central empirical finding concerns label design. State-based training labels

unlock detection that is structurally impossible under event-window labels in which the positive class marks a current deterioration state rather than proximity to a future system event. Task G achieved a Precision-Recall Area Under the Curve (PR-AUC) of 0.829 under Leave-One-User-Out (LOUO) cross-validation and identifies deteriorating users with a mean of 56.0 days before formal yellow well-being flags.

For event prediction specifically, the study establishes that red well-being flags are highly predictable from wrist-worn activity data alone. Task C (XGBoost, PR-AUC = 0.975) provided exactly 14 days of advance warning before confirmed red flag events for all five deteriorating users, with zero false alerts for stable users across the full observation period.

SHAP analysis reveals why each model works: Task C is driven by 14-day downward slopes in daytime activity; task G, by 14-day mean circadian amplitude and sleep duration. These are different features answering different clinical questions; the complementarity confirms that the two models target genuinely distinct aspects of deterioration rather than being incidental.

Finally, a proof-of-concept Streamlit dashboard demonstrates operational deployment of both models simultaneously, presenting risk scores and SHAP explanations in an interface designed for care coordinators.

1.5 Structure of the Thesis

The remainder of the thesis is organised as follows. Chapter 2 surveys the relevant literature on deterioration and falls in elderly populations, wearable sensor technologies, the distinction between risk detection and prediction, machine learning approaches for health risk modelling, and real-world deployment constraints. Chapter 3 describes the dataset, ethical considerations, preprocessing procedures, feature engineering, label generation, modelling framework, and evaluation strategy. Chapter 4 presents the empirical findings across all seven prediction tasks, includ-

ing three-model comparisons, SHAP analysis and lead time validation. Chapter 5 interprets the findings in relation to existing literature, discusses the practical implications for Oiva Health, and addresses the limitations of the study. Chapter 6 summarises the key contributions and proposes directions for future work.

Use of Artificial Intelligence Tools

Artificial intelligence tools used in the preparation of this thesis were to assist with language refinement, structural organisation, and clarity of presentation. Anara was used to search for relevant research sources; Claude (Anthropic) assisted in code debugging and academic writing feedback; and QuillBot was used to improve the clarity and tone of some passages. No confidential, personal, or research data were entered into these tools. All research design, data analysis, and interpretation were carried out independently by the author, who assumes full responsibility for the content of this thesis.

2 Related Work

2.1 Well-being Deterioration and falls in Elderly Populations

Acute health deterioration and falls represent a major and growing public health burden among older adults. Globally, approximately one in four adults aged 65 and above experiences at least one fall each year, and the burden escalates markedly in institutional and residential care settings, where prevalence rates can exceed 50% [1], [17]. The World Health Organization estimated that falls account for approximately 684 000 deaths annually worldwide, with adults over 60 bearing a disproportionate share of fatality and morbidity [2]. Projections based on demographic trends suggest that without effective preventive strategies, the absolute number of fall-related incidents will continue to rise through 2040 and beyond, driven primarily by population ageing rather than by any increase in age-standardised rates [17], [18].

Falls in older adults are often not isolated events. They are better understood as symptomatic expressions of broader physiological vulnerability, arising from a complex interaction of intrinsic and extrinsic risk factors. Intrinsic factors reported consistently across systematic reviews include muscle weakness, balance and gait impairment, cognitive decline, malnutrition, frailty, and polypharmacy [5], [6]. Mobility-related impairments, particularly reduced lower-limb strength and gait instability, have emerged as the most consistent predictors across multiple review methodolo-

gies. Sleep disruption and circadian dysregulation are increasingly recognised as contributors, as poor sleep quality and nocturnal activity impair daytime postural control and increase fall susceptibility, a relationship of particular relevance to home care and ageing-in-place contexts where nocturnal monitoring is typically absent [1], [3].

The consequences of falls extend well beyond the immediate incident. Hip fractures carry high mortality and prolonged functional decline; traumatic brain injuries may permanently alter independence[19]. Recurrent falls are common; approximately half of individuals who fall will do so again within twelve months, and fear of falling often precipitates a cycle of activity avoidance, deconditioning, and social withdrawal that further elevates risk [4], [20]. At the societal level, falls impose substantial economic costs through emergency care, rehabilitation, and long-term institutionalisation. These costs are not confined to acute events: the chronic, progressive nature of underlying deterioration means that detecting and acting on early warning signals carries potentially far greater value than responding to the fall itself.

This thesis takes a broader framing than fall detection alone. Automatic inactivity alarms and formal well-being flags, the primary prediction targets used here, represent the Vivago monitoring system’s own formalised detection of sustained behavioural deterioration, a category that encompasses the physiological trajectories that precede and predispose to falls, acute episodes, and emergency interventions. This framing is consistent with the literature’s recognition that falls are indicators of system-level vulnerability rather than discrete, unpredictable events [2], [6].

2.2 Wearable Sensors in Health Monitoring

Wearable sensors have become established tools for the continuous, non-intrusive monitoring of physical activity, sleep, and physiological parameters in older adults. Inertial measurement units typically incorporate accelerometers and a gyroscope

and are the most widely deployed sensing modality, capturing movement patterns at high temporal resolution without imposing the compliance burden associated with periodic clinical assessments [7], [8]. Wrist-worn devices have emerged as the preferred placement for elderly populations due to their unobtrusiveness and wear comfort, achieving higher adherence compared to waist, ankle, or chest placements in free-living conditions.

The validity of wrist-worn accelerometry for health monitoring in older adults has been examined across multiple systematic reviews. Chen et al. [7] provided evidence on wearable fall risk assessment in older adult centered communities, finding that wrist and lower-limb placements dominate current practice, with accelerometer-derived gait and activity features demonstrating acceptable validity against standard clinical assessments. Importantly, these reviews also highlight that laboratory-based validation frequently does not translate directly to free-living accuracy because motion artefacts and heterogeneous device placement compound the challenge of robust signal extraction in real care environments. Kristoffersson and Björk [21] similarly noted that longitudinal monitoring in diverse elderly cohorts remains comparatively under-represented relative to controlled study designs.

Sleep monitoring using wrist-worn accelerometry represents a growing area of validation research. Binary sleep-wake classification from wrist movement achieves reasonable accuracy when compared against polysomnography, though multi-stage sleep classification typically requires additional sensing modalities such as photoplethysmography [22], [23]. For the purposes of circadian rhythm monitoring, the assessment of the regularity and amplitude of the daily activity cycle, accelerometry alone has been shown to provide meaningful estimates of interdaily stability (IS) and intradaily variability (IV), validated gerontological metrics that quantify the consistency and fragmentation of the circadian pattern respectively [24].

The distinction between wearable and ambient sensing technologies is worth

noting for context. Ambient systems like fixed motion sensors, pressure mats and camera-based solutions can offer accurate event detection in controlled environments but lack the subject-level specificity of wearables and raise additional privacy concerns in home settings [25]. Wearable devices provide individual-level continuous data and are well-suited to the kind of longitudinal behavioural trajectory modelling that underlies this research.

Several methodological challenges remain. Step count validity and activity intensity classification vary substantially across device types and placement sites [26], [27], [28]. Gender, age, and mobility disparities affect sensor accuracy in ways that population-level calibration may not fully address. For the Vivago wrist-worn activity monitor used in this study, the manufacturer validated cut points for sedentary, light, moderate, and vigorous activity intensity were available. In addition, prior published work has reported a regression relationship between accelerometer output and metabolic equivalent of task (MET) values derived from indirect calorimetry, providing a basis for physiological interpretation of activity counts[29].

2.3 From Detection to Prediction

The distinction between detection and risk prediction is fundamental to understanding the contribution of this research and the limitations of most deployed systems. Fall detection systems identify an ongoing or completed fall event and trigger an alert; they are reactive by design [30], [31]. Threshold-based accelerometry approaches, detecting sudden changes in acceleration or orientation, remain the dominant commercial implementation and can achieve high accuracy in controlled conditions, with recent deep learning frameworks reporting accuracies exceeding 90% in laboratory settings [31], [32]. However, accuracy under real-world conditions degrades substantially, and the reactive nature of post-fall detection fundamentally limits its clinical utility; an alert generated after a fall cannot prevent the fall or the

injuries that accompany it [33], [34].

Fall risk prediction occupies a fundamentally different position. Rather than identifying a completed event, it attempts to characterise a user’s current risk state or forecast the likelihood of an upcoming adverse event, enabling preventive intervention [7], [35]. The clinical value this provides is a system that identifies elevated risk two to four weeks before a fall enables targeted physiotherapy, medication review, environmental modification, or care plan adjustment. A system that detects the fall as it happens enables only emergency response.

Despite the clear clinical advantage, the literature consistently identifies fall prediction as underdeveloped relative to detection. Reviews by Gharghan et al. [9] and Li et al. [10] found that the majority of commercial and academic systems remain detection-orientated, with prediction approaches comprising a small and methodologically heterogeneous minority. Challenges include the rarity of fall events (making supervised learning inherently difficult), the heterogeneity of older adult behaviour, and the absence of prospective longitudinal validation in free-living settings. This thesis addresses these challenges directly by working with confirmed longitudinal Vivago data from a real care environment and by reframing the prediction target as behavioural deterioration state rather than discrete fall events.

2.4 Machine Learning for well-being Risk Prediction

Machine learning approaches to health risk prediction in older adults span classical statistical models, ensemble methods, and deep learning architectures. Classical approaches like Logistic Regression, Support Vector Machines, and Random Forests rely on engineered features extracted from time-series or clinical data. These models are interpretable, computationally tractable and have demonstrated competitive

performance in small-to-medium cohort studies. A prospective cohort study by Shao et al. [15] evaluated seven machine learning classifiers for six-month fall risk prediction in nursing home residents using multimodal clinical features, finding XGBoost competitive with the best-performing models. Jahandideh et al. [36] applied Random Forest to hospital electronic health records, achieving high specificity for fall prediction without sensor dependency.

Deep learning models like Convolutional Neural Networks, Long Short-Term Memory networks, and Transformer architectures offer the potential for end-to-end learning from raw sensor signals without explicit feature engineering. Roshdibenam et al. [37] demonstrated CNN-based fall risk prediction from IMU data collected during Timed Up and Go tests. Lin et al. [38] reported LSTM-based prediction achieving 0.90 accuracy on combined activity and sleep data from the MobiAct public dataset. Guo et al. [39] applied Transformer architectures to clinical time series, with results suggesting advantages over classical methods in data-rich hospital settings.

However, the applicability of deep learning approaches to the real-world scenario addressed by this thesis is limited. With 20 users and a positive class comprising a small fraction of training rows of 3.4% for the primary event prediction task, the data volume is insufficient to train sequence models robustly without substantial risk of overfitting. This is a well-recognised problem in wearable health data modelling so gradient-boosted trees and ensemble methods consistently outperform deep learning approaches when the number of labelled subjects is small [14], [15]. The decision to restrict this study to Logistic Regression, Random Forest, and XGBoost is therefore methodologically grounded.

A particularly relevant methodological issue is the choice of training labels. Most published approaches define the prediction target as whether a fall or other adverse event will occur within a fixed forward time window [16], [35]. This framing is oper-

ationally clear but creates a structural problem that days that precede an event by more than the window length are labelled negative, even if the user’s behavioural profile is already deteriorating. The resulting training signal suppresses early detection capacity. An alternative of labelling the current deterioration state rather than an upcoming event addresses this problem directly but requires a principled method for determining deterioration onset from continuous sensor data. This study implements both approaches, using event-window labels for the primary event prediction task and change-point-derived state labels for the early state detection task, and demonstrates empirically that the label design substantially determines the model’s temporal detection behaviour.

The role of personalised normalisation in this domain has received limited explicit attention in the published literature, yet it is critical for heterogeneous elderly cohorts. Cohort-level activity norms are poor representations of individual baselines: a user who is habitually sedentary will appear low-risk relative to a population norm even if their personal activity is declining precipitously. Within-subject normalisation involves computing each user’s personal baseline and expressing subsequent measurements as deviation from that baseline, which is the correct approach for detecting deterioration within individuals rather than comparing across them [7]. This study implements per-user z-score normalisation with a 30-day baseline window, demonstrating that raw absolute activity values show no statistically significant difference between pre-event and non-event periods ($p = 0.997$), whereas normalised features show significant separation at the 14-day horizon ($p = 0.033$).

2.5 Explainability and Clinical Interpretability

The deployment of machine learning in clinical and care settings introduces requirements beyond predictive accuracy. Care coordinators and clinicians need to understand why a model generates a given alert, both to assess its credibility and to

determine an appropriate care response. Black-box predictions carry limited actionability; a risk score without an explanation does not tell a care coordinator whether to increase monitoring frequency, review medications, or schedule a physiotherapy assessment.

SHAP values provide a theoretically grounded framework for attributing model predictions to individual input features [39], [40]. SHAP analysis has been applied in clinical risk modelling to identify which physiological or behavioural variables are driving predictions for individual patients, supporting the kind of targeted intervention planning that regulatory and clinical governance frameworks increasingly require [11].

This study applies SHAP analysis to both primary models. The event prediction model (Task C) is shown to be dominated by 14-day downward slopes in daytime activity and mean activity; these features are directly interpretable as a declining trend in physical engagement. The state detection model (Task G) is dominated by 14-day mean levels of circadian amplitude and sleep duration, which are interpretable as chronic behavioural profile differences between deteriorating and stable users. The consistency of these feature profiles across all five deteriorating users, each evaluated independently, suggests that the models have learned generalisable deterioration signatures rather than user-specific patterns.

2.6 Industry and Real-World Deployment Constraints

Translating predictive models from academic research to operational deployment in digital home care faces a set of practical constraints that are frequently under-acknowledged in the literature. Data privacy regulations, particularly the General Data Protection Regulation (GDPR) applicable across the European Union, impose strict requirements on the collection, storage, and processing of personal health data from wearable devices [12], [41], [42]. Pseudonymisation and data minimi-

sation strategies are essential, and any pipeline intended for deployment must be designed with these constraints built in rather than retrofitted.

Label scarcity is a separate problem and, in some ways, a harder one. Confirmed adverse events like falls, acute deteriorations, emergency interventions are rare, almost by definition. Most residents in a well-managed care facility go months without a serious incident. Over a six-to-twelve-month observation window, any individual user might contribute only a handful of positive examples. This rarity creates severe class imbalance that standard supervised learning approaches handle poorly without explicit accommodation [13]. The use of class weighting (`class_weight='balanced'` for Logistic Regression and Random Forest, `scale_pos_weight` for XGBoost) and the selection of precision-recall area under the curve (PR-AUC) rather than receiver operating characteristic - area under the curve (ROC-AUC) as the primary evaluation metric reflect this constraint directly. ROC-AUC can produce misleadingly optimistic estimates of classifier performance in highly imbalanced settings[14].

Free-living sensor data introduces noise sources absent from laboratory recordings, irregular device wear, motion artefacts, missing data segments, and comorbidity-related signal contamination. Models trained on controlled datasets frequently lose predictive power when applied to real care environments [43]. The data quality filtering applied in this study, like excluding days with device completeness below 60% and applying a seven-day post-event cooldown to prevent recovery-period contamination of the negative class, reflects this practical reality.

Clinical acceptance poses a further barrier. Care staff and coordinators may be reluctant to act on algorithmic alerts if those alerts are frequent, unexplained, or inconsistent with their own observations [44]. Alert fatigue is a documented risk in clinical AI deployment: systems that generate too many false positives lose credibility and may be disregarded even when alerts are genuine. The finding that all three confirmed alarm-free users (User_004, User_016, User_018) maintained

risk scores below 0.30 throughout the full seven-month observation period speaks directly to this concern.

Proof-of-concept evaluations in realistic settings serves as a critical bridge, providing controlled environments to test methodological assumptions, validate ethical and technical constraints, and assess user acceptance before committing to large-scale deployment [11], [45]. This study includes a pipeline which is designed and assessed as a deployable proof-of-concept for Oiva Health’s digital home care platform, with the practical constraints of GDPR compliance, class imbalance, real-world noise, and care-staff interpretability built into every methodological decision.

2.7 Summary

Several patterns emerge from this review. The majority of existing work focuses on fall detection rather than prediction and on laboratory settings rather than deployed care environments. Where longitudinal real-world data are used, prediction targets are almost always event-based, which structurally suppresses early detection capacity, as discussed above. The label design question has attracted surprisingly little attention; most studies treat it as a data engineering decision rather than a scientific one. Real-world wearable data from deployed commercial devices feature in only a small minority of studies because of their attendant incompleteness, heterogeneity, and privacy constraints. And while explainability is widely acknowledged as necessary for clinical deployment, it is rarely implemented in wearable-based deterioration prediction in any rigorous way.

These gaps collectively justify the approach taken in this research: a real-world, privacy-aware pipeline evaluated under the strict cross-validation protocol available for small cohorts, addressing two complementary clinical questions through two methodologically distinct label designs, with SHAP-based explainability built into both primary models. The methodology is described in Chapter 3.

3 Methodology

3.1 Research Design

This study follows an empirical, quantitative research design orientated toward proof-of-concept development. The aim is to evaluate whether machine learning applied to longitudinal wrist-worn activity sensor data can support proactive well-being risk detection in elderly care, addressing the two research questions articulated in Chapter 1. The work was conducted in collaboration with Oiva Health, which provided access to pseudonymised real-world monitoring data and the clinical context required to define meaningful prediction targets.

The design is explicitly bounded as a proof-of-concept rather than a clinical validation study. For this study, no new data were collected, and no experimental intervention was performed. The pipeline operates entirely on observational data generated through the routine deployment of the Vivago monitoring system. This framing preserves ecological validity, as the data reflects genuine care-environment conditions, including device non-wear, incomplete records, and the sparse clinical event labels characteristic of real elderly care populations, while acknowledging the constraints that arise from a cohort of limited size.

Deep learning was not considered for this study because the data cannot support it. With 20 users and a positive class rate as low as 3.4% for the primary prediction task, sequence models like LSTMs, Transformers, and similar architectures would

overfit to the training users rather than learning generalisable deterioration patterns. Classical machine learning methods are the right tools at this scale, a conclusion well-supported in the literature [14], [15]. Three models were therefore evaluated: Logistic Regression as a linear baseline, Random Forest as an ensemble comparator, and XGBoost as the primary model.

The evaluation protocol is Leave-One-User-Out (LOUO) cross-validation, in which each of the 20 users serves as the test set exactly once while the remaining 19 users form the training set. This protocol is the most stringent available given the cohort structure: it ensures that no information about the test user appears in training, directly simulating the deployment scenario in which the model is applied to a new user whose data was not observed during training. The primary evaluation metric is the area under the precision-recall curve (PR-AUC), selected in preference to ROC-AUC because of the substantial class imbalance present across all prediction tasks. Figure 3.1 provides an overview of the complete pipeline, from raw Vivago activity data through to the proof-of-concept dashboard.

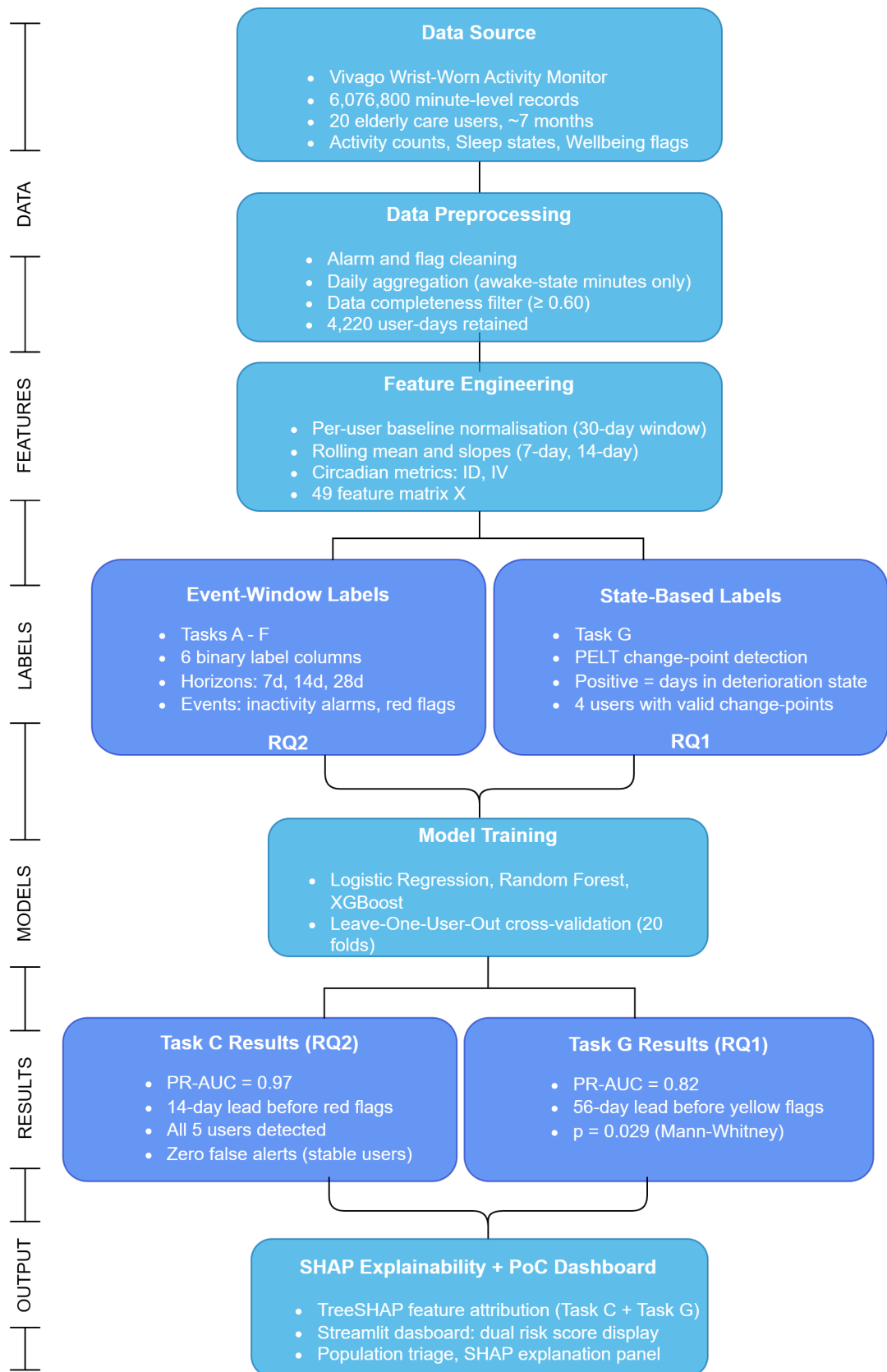


Figure 3.1: Overview of the machine learning pipeline for proactive wellbeing risk detection

3.2 Data Source and Overview

3.2.1 The Vivago Monitoring System

The Vivago[®] wrist-worn activity monitor, developed by Vivago Oy (Espoo, Finland), is a commercial device designed for continuous health monitoring in elderly care settings. It incorporates a piezoelectric accelerometer with a dynamic range of $\pm 4G$ and a body-motion sampling rate of 40 samples per second. Previous validation studies report that the device is sensitive to movements in the 0.5–10 Hz frequency range, although the underlying signal-processing implementation is not publicly described [29]. The raw acceleration signal is processed onboard into a single integer activity count per minute, where a value of zero indicates no detected movement. Published intensity cut-points validated for this device classify activity as sedentary (0–15 counts per minute), light intensity (16–40), moderate intensity (41–85), and vigorous intensity (above 85) [29]. A validated MET regression equation has been published for this sensor: $MET = 0.0949 \times \text{counts}/\text{min} + 1.9145$ ($R^2 = 0.82$, $r = 0.89$, $p < 0.001$) [29]. The device is more sensitive to low-intensity movement than traditional actigraphs, a property advantageous in elderly care populations where most physical activity falls in the sedentary-to-light intensity range[46].

Alongside the activity count, the Vivago system classifies each minute-epoch into one of four mutually exclusive states: awake (state 0), gap (state 1), sleep (state 2), and off-wrist (state 3). These state annotations are generated by the device’s onboard algorithms and provide the basis for distinguishing genuine activity from non-wear and sleep periods during feature engineering.

The Vivago system also generates automated well-being alerts derived from long-term trend analysis of activity and sleep patterns. Yellow well-being flags indicate that a user’s monitoring profile has changed to need increased attention; red well-being flags indicate a significant, potentially critical change, triggered specifically by

a weak circadian rhythm or low daytime activity. Automatic inactivity alarms are generated when the system detects a sustained period of inactivity. Manual alarm button presses by the user or caregiver were also recorded and treated as proxy indicators of acute distress or fall-related incidents.

3.2.2 Dataset Characteristics

The dataset comprises pseudonymised continuous monitoring records for 20 elderly care users. The raw activity data file contains 6,076,800 minute-level rows, yielding 4,220 user-days after daily aggregation. Device adherence was high; the mean data completeness, defined as the proportion of awake and sleep minutes per day relative to 1,440 minutes, was 0.943. Cohort baseline means computed over the first 30 observation days per user are summarised in Table 3.1. Individual baselines ranged from 20.4 to 46.5 counts per minute across users, reflecting substantial inter-user heterogeneity that motivated the per-user normalisation approach described in Section 3.5.1.

Table 3.1: Cohort baseline feature means computed over the first 30 observation days per user.

Feature	Cohort mean
mean_activity (counts/min)	29.5
daytime_activity (counts/min)	35.4
sleep_hours	7.5
circadian_amplitude	31.8
peak_activity_hour	13.8
data_completeness	0.943

Five users experienced confirmed well-being deterioration during the observation period, evidenced by at least one yellow followed by one or more red well-being flags: User_003, User_011, User_012, User_014, and User_020. Three users had zero alarm events of any type across the full observation period (User_004, User_016, User_018) and served as the confirmed negative control group in trajectory valida-

tion.

A total of 128 alarm events were recorded across all users prior to filtering. After excluding battery-related and device-status alarms, 59 clinically relevant alarm records remained: 29 automatic inactivity alarms, 19 manual alarm button presses, and 11 well-being change notifications. Fifteen well-being flag records were retained after cleaning: 5 yellow flags and 10 red flags.

The inter-user heterogeneity in absolute activity levels in individual baselines ranges from 20.4 to 46.5 counts per minute, spanning light to moderate intensity per the published Vivago cut-points. (Figure 3.2) confirms that population-level activity norms cannot serve as meaningful deterioration thresholds, motivating the per-user normalisation approach described in Section 3.5.1.

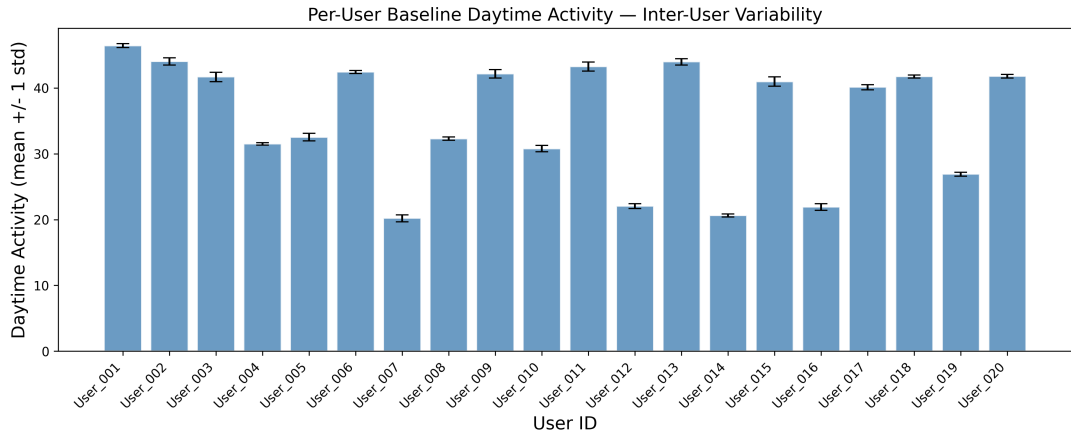


Figure 3.2: Per-user baseline daytime activity.

Per-user daytime activity timelines (Figure 3.3) reveal pronounced differences between stable and deteriorating users, with the five flagged users showing sustained downward trends not present in stable users, illustrating the inter-user heterogeneity that motivated per-user normalisation.

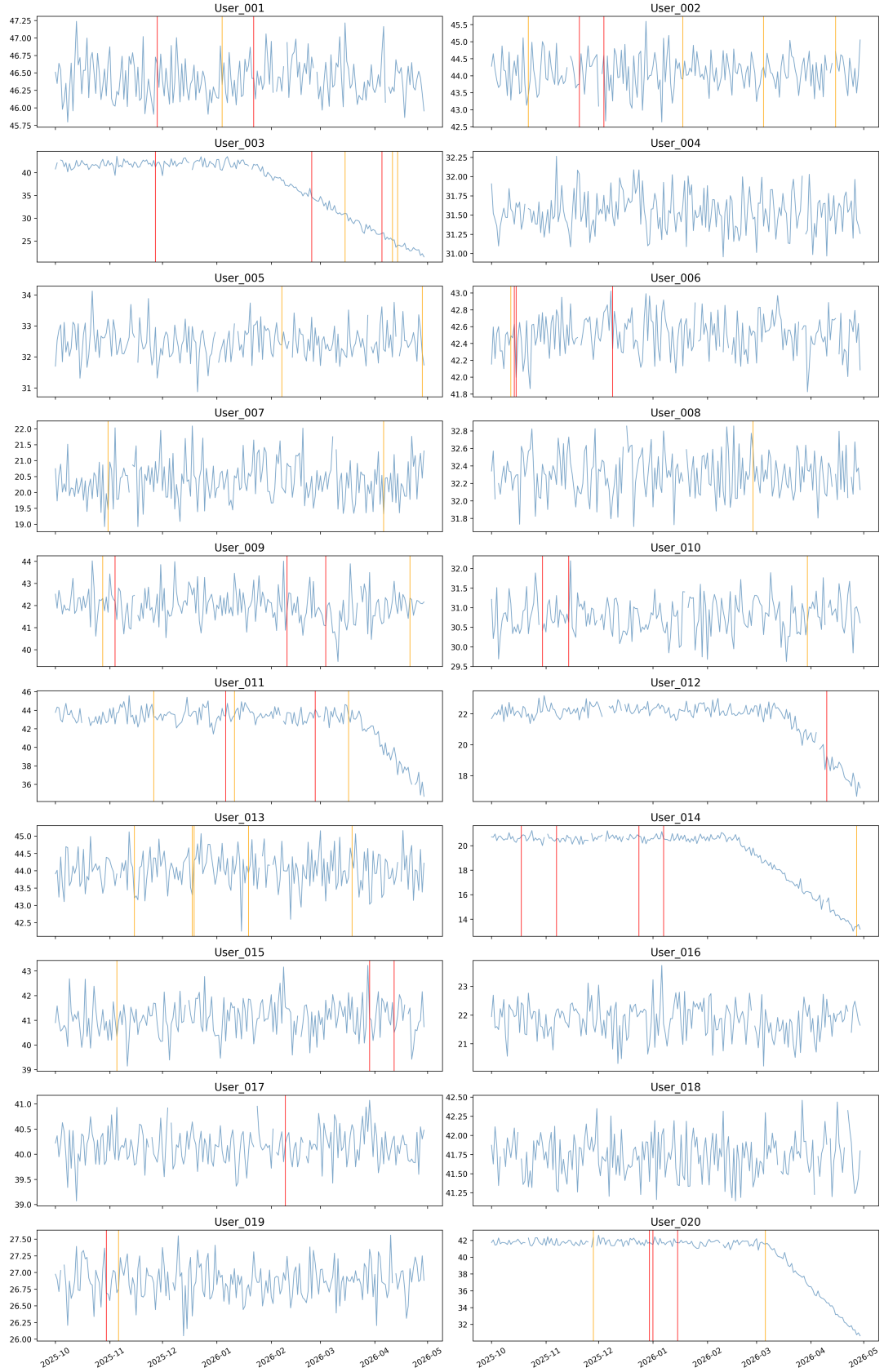


Figure 3.3: Per-user daytime activity timelines.

Activity trajectories for the five flagged users, overlaid with alarm and flag events (Figure 3.4), confirm that alarm events cluster during or after the declining phase rather than at its onset.

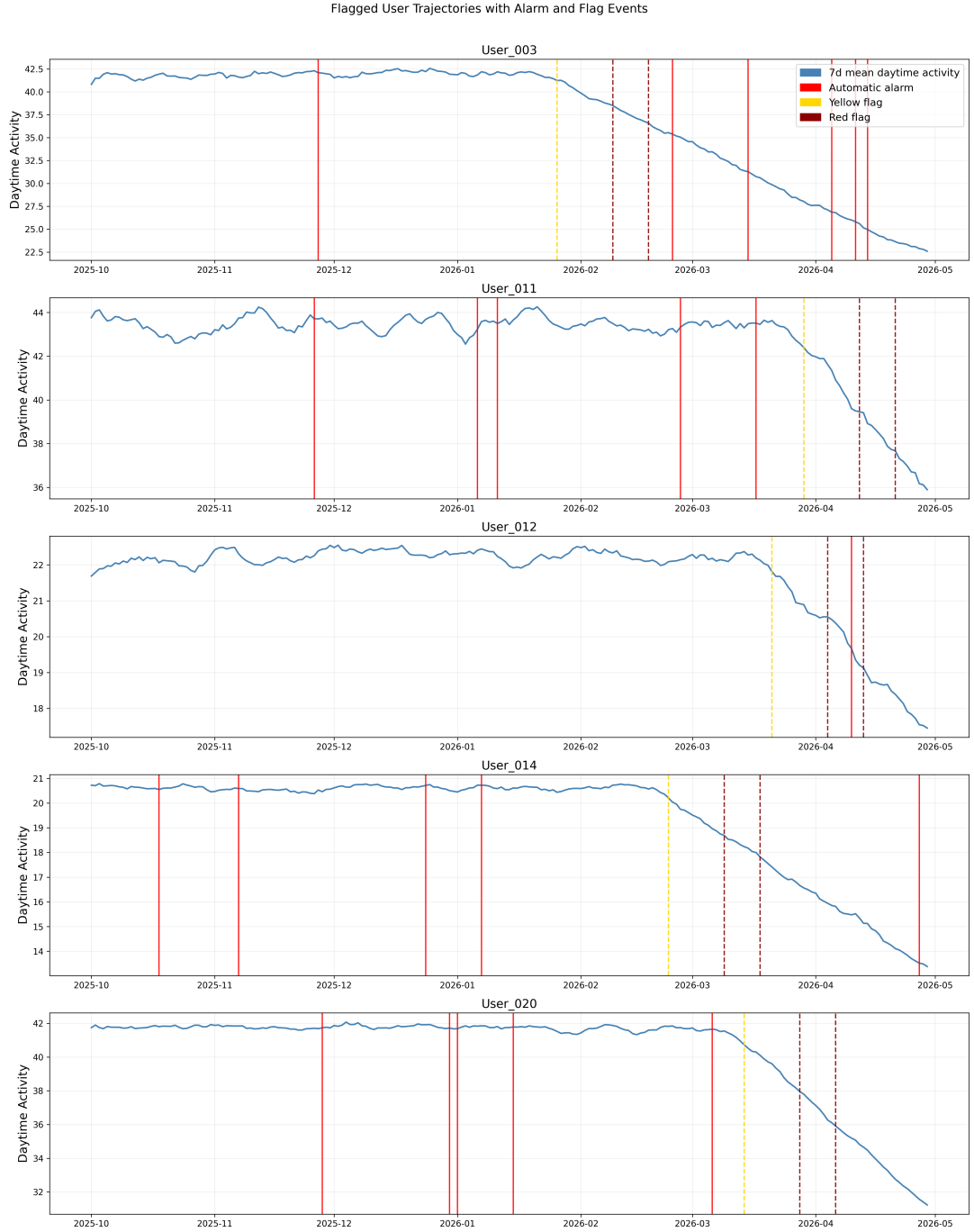


Figure 3.4: Flagged User Trajectories

3.3 Ethical and Privacy Considerations

All data used in this study were provided by Oiva Health under a research collaboration agreement. Prior to access, the dataset was pseudonymised by replacing direct identifiers with anonymous user codes (User_001 through User_020). No name, date of birth, address, or other directly identifying information was present in any file provided to the researcher. The pseudonymised dataset was stored exclusively in secured computing environments, and no raw data were uploaded to public repositories or shared with third parties.

The study uses retrospectively collected, pseudonymised and temporally shifted operational monitoring data. No new data were collected, and no experimental interventions were performed. The use of the data for research purposes was covered by the data sharing agreement between the researcher and Oiva Health, which specified the scope of analysis and publication.

Consistent with the data minimisation principle of the GDPR, user identifiers were referenced only to the extent necessary for the leave-one-user-out cross-validation protocol. All analytical outputs are reported at the aggregate or anonymised user level.

3.4 Data Preprocessing

3.4.1 Alarm and Flag Cleaning

Preprocessing began with the alarm events and well-being flags files. From the alarm records, alarm types classified as `battery_low` or `device_status` were excluded entirely, as these reflect device operational states rather than user health events. The remaining alarm events were retained and categorised as `automatic_inactivity`, `manual_alarm`, or `well-being_change`.

3.4.2 Daily Aggregation

The minute-level activity records were aggregated to a single row per user per calendar day. Activity magnitude features were computed exclusively from awake-state minutes (state 0), ensuring that sleep and non-wear periods did not contribute to activity-level aggregates. Key daily aggregates included mean activity, daytime activity, night activity, maximum activity count, active minutes above the sedentary threshold, sleep hours, sleep onset hour, sleep fragmentation, hourly activity coefficient of variation, and data completeness. The state classifier cleanly separated behavioural states, as shown in (Figure 3.5). Awake-state minutes (state 0) show the widest distribution and highest counts; sleep-state minutes (state 2) are concentrated near zero; gap and off-wrist minutes are uniformly zero.

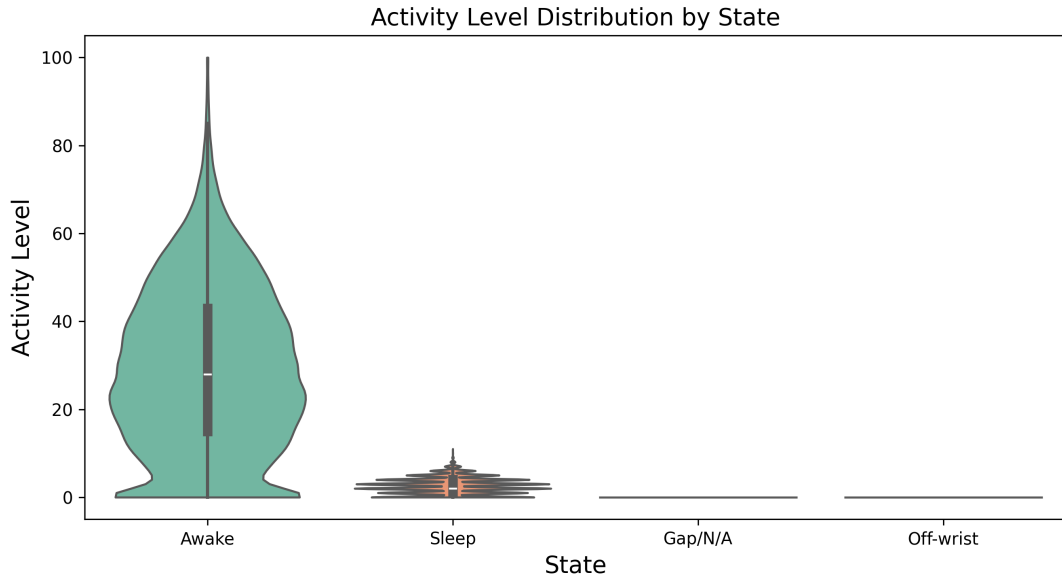


Figure 3.5: Activity level distribution by Vivago sensor state.

Days with data completeness below 0.60 were excluded from training. This threshold was selected to eliminate days on which the monitoring record was too incomplete to produce reliable daily aggregates while retaining the large majority of available data.

3.5 Feature Engineering

3.5.1 Per-User Baseline Normalisation

One user in this cohort walks 46 counts per minute on an average day; another manages 20. Both are entirely normal for those individuals. A population-level activity threshold applied to both would either generate constant false alerts for the low-activity user or miss genuine deterioration in the high-activity one. The solution is to compare each person to themselves. Per-user normalisation expresses daily measurements as deviation from each user’s personal 30-day baseline, so the model learns to detect within-person decline rather than between-person differences.

The baseline for each user was defined as the first 30 observation days, sorted chronologically. Baseline statistics such as mean and standard deviation were computed for four primary features: `mean_activity`, `daytime_activity`, `sleep_hours`, and `circadian_amplitude`. A minimum standard deviation floor of 0.5 was applied to prevent division by near-zero values:

$$\text{safe_std} = \max(\text{baseline_std}, 0.5)$$

From these baseline statistics, three normalised feature variants were generated for each base feature: a winsorised z-score clipped to $[-5, 5]$, a percentage of baseline mean (`pct_baseline`), and an absolute baseline deviation (`baseline_dev`). Raw z-scores were set to NaN for the first 30 days of each user’s record to prevent data leakage. The z-score distribution after normalisation is presented in Section 4.1.3.

3.5.2 Rolling Features

Deterioration in elderly care populations typically unfolds over days to weeks rather than abruptly. To capture this temporal dimension, rolling statistics were computed for the primary activity and sleep features across two time windows: 7 days and

14 days. Rolling means and linear slopes were computed separately for each user using a grouped operation, ensuring that no information from one user’s record contributed to another user’s rolling statistics. A minimum of four observations was required within a rolling window before a value was computed, preventing edge-period artefacts from the start of each user’s record.

The 14-day window was selected as the primary rolling horizon based on the prediction horizon analysis described in Section 3.6, which confirmed statistically significant separation between pre-alarm and non-alarm normalised activity values at this horizon. The 7-day window was included as a complementary shorter-term feature set capturing recent trend direction and short-term fluctuation not fully represented in the 14-day statistics. Although the 7-day prediction horizon did not achieve statistical significance in the window analysis, this test evaluated labelling horizon suitability rather than feature informativeness; a 7-day slope reflects a different temporal scale of behavioural change than a 14-day slope, and the two are complementary rather than redundant. Feature variability analysis confirmed this empirically; `sleep_fragmentation_7d_slope` was the highest-variability feature in the full 49-feature matrix.

A 28-day rolling window was not included for two reasons. First, combined with the 30-day baseline warm-up exclusion, a 28-day rolling requirement would delay the first valid feature value to day 58 per user, further reducing the already-limited training set beyond the exclusions already applied. Second, a 28-day rolling mean captures substantially the same behavioural trend as the 14-day mean with greater temporal smoothing, producing high inter-feature correlation and marginal information gain relative to the increased feature count and reduced training coverage.

Rolling linear slopes were computed by fitting a first-order linear regression to the window of observations for each day, producing a coefficient that quantifies the rate and direction of change. Positive slopes indicate increasing values; negative

slopes indicate decline.

3.5.3 Circadian Rhythm Metrics

Two validated gerontological metrics for characterising circadian rhythm regularity were computed from the minute-level activity records using a rolling 7-day window:

Interdaily Stability (IS) quantifies the regularity of the 24-hour activity pattern across days:

$$\text{IS} = \frac{n \sum_{h=0}^{23} \bar{x}_h^2}{\sum_{i=1}^N x_i^2}$$

where $\bar{x}_h$ is the mean activity at hour h averaged across all days in the window, x_i is each individual minute-level observation, n is the number of hours in a day, and N is the total number of observations. Higher IS values indicate a more consistent daily rhythm.

Intradaily Variability (IV) quantifies the fragmentation of the activity rhythm within days:

$$\text{IV} = \frac{N \sum_{i=1}^{N-1} (x_{i+1} - x_i)^2}{(N - 1) \sum_{i=1}^N x_i^2}$$

Higher IV values indicate more fragmented activity patterns, which have been associated with circadian disruption and cognitive decline in elderly populations [24]. Both metrics were set to NaN when fewer than three full days of data were present in the rolling window.

3.5.4 Feature Matrix

The final feature matrix $\mathbf{X}$ was assembled from the above feature engineering steps. Columns excluded from model input included raw metadata (user identifier, date, day number), all raw z-score columns (the winsorised `_zscore_capped` versions were retained), and all baseline mean and standard deviation columns. The following four daily device quality columns, `off_wrist_minutes`, `gap_minutes`, `wear_minutes`,

and raw `data_completeness` were excluded on the grounds that they reflect device management behaviour rather than health-relevant signals; a sensitivity analysis confirmed that including these columns led to spurious early detection artefacts. Rolling versions of data completeness were retained, as they capture chronic wear pattern trends. The resulting feature matrix contained 49 columns. Missing values arising from the baseline warm-up window were imputed using per-user median values, with a global median fallback for any remaining missingness.

The feature correlation structure (Figure 3.6) reveals the following broad clusters: rolling slope features, normalised level features, and circadian and sleep features. The separation between slope and level clusters is consistent with the complementary SHAP profiles observed for the two primary models discussed in Chapter 4.

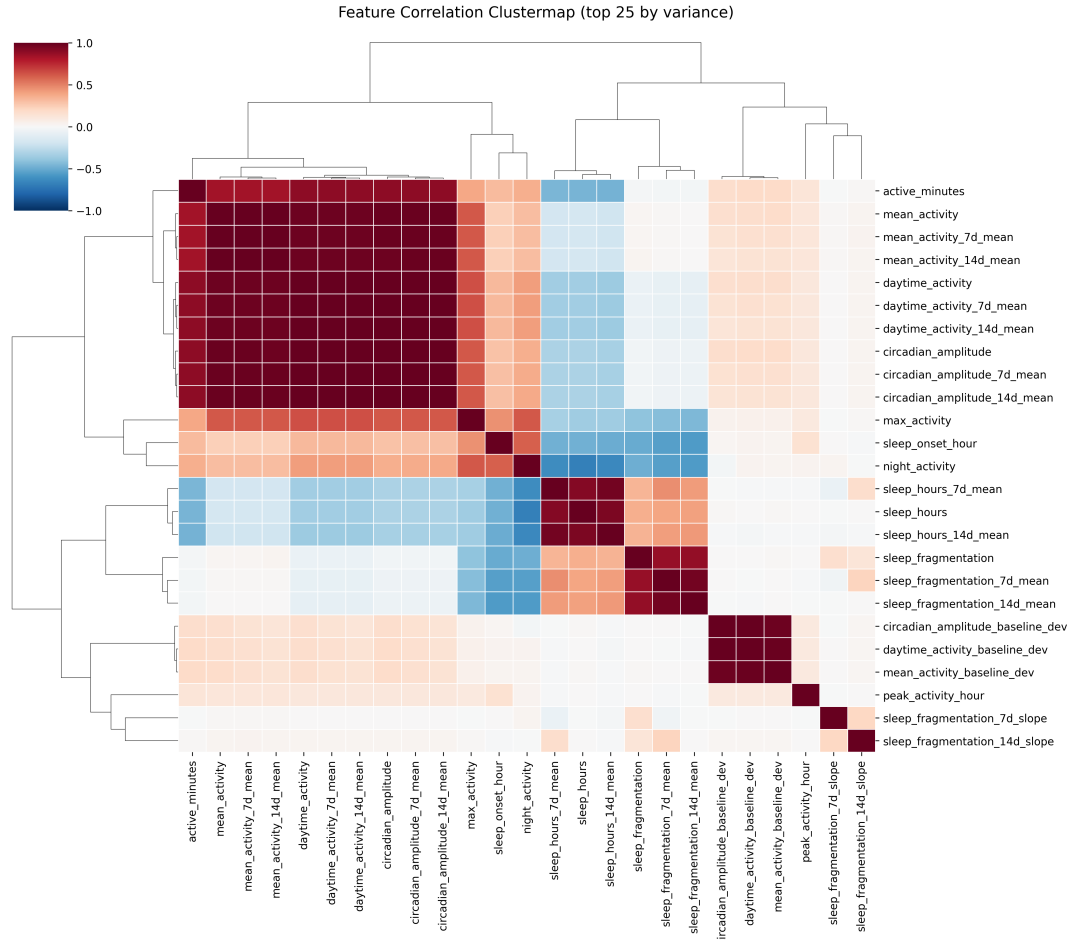


Figure 3.6: Hierarchical clustering of feature correlations for the top 25 features by variance.

3.6 Prediction Horizon

Before defining prediction labels, the statistical significance of the behavioural signal preceding alarm events was assessed across three candidate window lengths. For each window size $w \in \{7, 14, 28\}$ days, all user-days falling within w days before an automatic inactivity alarm were classified as pre-alarm; all other days were classified as non-alarm. A one-tailed Mann-Whitney U test was applied to the normalised `daytime_activity_zscore_capped` values for the two groups. The results of this analysis and the corresponding figure are reported in section 4.1.2.

The 14-day window was selected as the primary label horizon. Although the 28-day window also achieves significance, the 14-day horizon produces a cleaner training signal: a 28-day positive window labels days up to four weeks before an alarm as positive, including periods during which behavioural indicators may not yet reflect the trajectory leading to the event. The empirical confirmation of this choice from the model comparison results is discussed in Chapter 4.

3.7 Label Generation

3.7.1 Event-Window Labels (Tasks A–F)

Six binary event-window label columns were generated, each defining whether a specified type of adverse event occurred within a defined forward horizon for each user-day. The labelling logic assigned a value of 1 to a given day if the corresponding event type occurred within days 1 through h following that day (exclusive of the day itself), and 0 otherwise. Days assigned a positive label under a narrower horizon are a strict subset of those labelled positive under a wider horizon for the same event type.

Table 3.2: Event-window prediction tasks and label definitions.

Task	Label column	Event type		Horizon	Positive rate
A	label_inactivity_7d	Automatic alarm	inactivity	7 days	4.6%
B	label_inactivity_14d	Automatic alarm	inactivity	14 days	7.9%
C	label_redflag_14d	Red well-being flag		14 days	3.4%
D	label_redflag_28d	Red well-being flag		28 days	5.1%
E	label_combined_alarm_14d	Inactivity alarm	or manual	14 days	13.5%
F	label_combined_alarm_28d	Inactivity alarm	or manual	28 days	23.3%

Task C was designated the primary RQ2 model, which was predicting a red well-

being flag within 14 days. Red well-being flags represent formally confirmed, sustained deterioration as assessed by the Vivago system, providing a clinically meaningful and unambiguous positive class. Tasks A, B, D, E, and F were evaluated as exploratory comparators to contextualise the performance advantage of Task C and demonstrate the effect of label design on model behaviour.

Three categories of rows were excluded from training across all tasks: days with data completeness below 0.60 (79 rows), days falling within the first 30 days of each user’s record (600 rows), and days falling within seven days following any positive alarm event of types `automatic_inactivity` or `manual_alarm` (313 rows). This post-event cooldown window prevents recovery-period days from being used as negative training examples. After applying all three exclusions, 3,265 training-eligible rows remained from 4,220 total user-days.

3.7.2 State-Based Labels (Task G)

The event-window label design in Tasks A through F constrains the model’s temporal detection behaviour as, days preceding a system event by more than the window length are labelled negative regardless of the user’s behavioural state. This structural feature suppresses early deterioration signal and limits the model to detecting changes only in the period immediately before an event occurs.

Task G was designed to remove this constraint by labelling the deterioration state directly. Rather than asking, "Will a system event occur within the next h days?" The Task G labels encode the question: "Is this user currently in an active behavioural deterioration phase?"

Deterioration onset dates were identified for each of the five deteriorating users using the ruptures library’s PELT (Pruned Exact Linear Time) algorithm [47] applied to each user’s smoothed daytime activity time series (`daytime_activity_7d_mean`). The Radial Basis Function (RBF) cost function was used, with the penalty

parameter controlling sensitivity: higher penalties favour fewer, larger structural breaks; lower penalties allow detection of smaller or more gradual transitions.

User-specific penalty parameters were applied where the default setting of 5 did not produce a plausible result. A lower penalty of 2 was used for User_003 and User_014, whose gradual or complex activity trajectories required greater sensitivity. Change-point dates were validated visually against Figure 4.15 in Chapter 4, which plots the detected change-point alongside the smoothed activity trajectory, yellow flag, and red flag for each user.

For User_003, whose 10-week linear decline did not produce a clean structural break under any automated penalty because PELT identifies step-function breaks more reliably than gradual linear slopes, the change-point date was set to 15th November 2025 based on visual inspection of the smoothed activity trajectory as the point at which the rolling mean visibly transitions from a stable plateau into sustained decline. This manual override was documented and applied consistently.

For User_011, the PELT algorithm detected a change-point on 25th March 2026, four days before the yellow flag. Visual inspection of the smoothed trajectory reveals an earlier, smaller decline around February 2026 that the algorithm did not detect at the penalty setting used. This reflects a known characteristic of PELT because with a penalty parameter that prevents over-segmentation, the algorithm identifies the dominant structural break, here the acute late-March decline, rather than earlier smaller changes. User_011 is therefore characterised as a dynamic detection case in this study.

User_014 was excluded from Task G entirely. Recurring automatic inactivity and manual alarm events throughout the nominally stable phase prevented reliable identification of a valid deterioration onset date before the yellow flag, regardless of penalty setting.

The inherent subjectivity of the manual override for User_003 and the depen-

dence of automated results on penalty parameter selection are acknowledged as limitations of the change-point label generation approach. The specific change-point dates identified and their lead times relative to formal system events are reported in Table 4.10 in Chapter 4.

3.8 Model Development

3.8.1 Model Selection

Three models were evaluated across all prediction tasks: Logistic Regression, Random Forest, and XGBoost. Logistic Regression (`class_weight='balanced'`) served as the linear baseline, a useful lower bound that shows what is achievable without non-linear feature interactions. Random Forest extended this with an ensemble approach, capturing feature interactions through tree structures while remaining interpretable through feature importance scores. XGBoost was the primary model. It was selected partly for its strong performance in prior studies involving imbalanced classification problems and clinical prediction tasks [48], [49], [50] and partly for its native `scale_pos_weight` parameter, which handles imbalanced classes more directly than the balanced weighting approach used by the other two.

For XGBoost, the class imbalance parameter was computed separately for each prediction task as:

$$\text{scale_pos_weight} = \frac{N_{\text{neg}}}{N_{\text{pos}}}$$

where N_{neg} and N_{pos} are the counts of negative and positive training examples, respectively. This parameter ranged from approximately 3.3 for Task F (23.3% positive rate) to approximately 28.2 for Task C (3.4% positive rate). For the leave-one-user-out folds, `scale_pos_weight` was recomputed within each fold using the training subset rather than the global dataset to account for fold-level variation in class composition.

3.8.2 Hyperparameter Tuning

Hyperparameter optimisation was performed for XGBoost using `RandomizedSearchCV` with 30 iterations and `GroupKFold(5)` cross-validation, with user identifiers as the grouping variable to prevent user-level data leakage during tuning. The scoring metric was average precision (equivalent to PR-AUC). The search space covered learning rate, maximum tree depth, number of estimators, column subsampling ratio, row subsampling ratio, and minimum child weight.

Tuning was conducted on Task B (`label_inactivity_14d`), which provides a larger positive class than the primary Task C (7.9% versus 3.4%) and therefore a more stable optimisation landscape. The best cross-validated PR-AUC achieved during tuning on Task B was 0.160, which was approximately twice the Task B no-skill baseline of 0.079, reflecting a meaningful improvement on a genuinely difficult prediction task. Tuning CV scores are lower than final LOUO scores because `GroupKFold(5)` trains on four-fifths of users per fold, whereas LOUO trains on nineteen-twentieths; the tuning identifies the best generalising hyperparameter configuration rather than the highest absolute score.

The best configuration identified was: `max_depth=3`, `learning_rate=0.057`, `n_estimators=356`, `colsample_bytree=0.955`, `subsample=0.694`, `min_child_weight=3`. The shallow tree depth combined with a large number of estimators and slow learning rate is consistent with best practice for small-N datasets; it regularises the model against overfitting while preserving the capacity to capture complex feature interactions through additive ensemble structure. The tuned hyperparameters were subsequently applied to all six event-window tasks with per-task `scale_pos_weight` adjustment and to Task G with its own `scale_pos_weight`.

3.9 Evaluation Strategy

3.9.1 Cross-Validation Protocol

The primary evaluation protocol was Leave-One-User-Out (LOUO) cross-validation. For each of the 20 users, the full dataset was split into a training set comprising the remaining 19 users and a test set comprising the held-out user. The model was retrained from scratch for each fold with a fold-specific `scale_pos_weight` computed from the training subset. Folds in which the test user had no positive labels, along with those in which the training set contained only one class, were skipped, as PR-AUC is undefined in the absence of positive test examples.

LOUO was selected as the evaluation protocol because it directly simulates the deployment scenario where the model is applied to a new user whose data was not observed during training. With 20 users, each serving as the test set exactly once, LOUO provides the most stringent available estimate of generalisation performance for this cohort structure.

3.9.2 Evaluation Metrics

The primary evaluation metric was the area under the precision-recall curve (PR-AUC), computed using `average_precision_score` from scikit-learn. This metric is appropriate for imbalanced classification tasks because it explicitly quantifies performance in the positive class without the masking effect that large true-negative counts introduce into ROC-AUC. No-skill baselines were included in all PR curve figures to contextualise the magnitude of improvement over random prediction.

Secondary metrics reported at the operating threshold of 0.4 included recall (sensitivity), precision (positive predictive value), and F1 score. The threshold of 0.4 was selected based on threshold analysis of LOUO out-of-fold predictions for Task C (Section 4.3.2), which confirmed that $\text{recall} = 0.893$ at this operating point,

satisfying the pre-specified clinical minimum of 0.70 with a substantial margin. For Task G, PR-AUC is reported as the primary and sole summary metric; the rationale for this decision is described in Section 4.4.2.

3.9.3 Lead Time Validation

For the primary event prediction model (Task C) and the state detection model (Task G), lead time analysis was conducted to quantify how many days in advance each model first exceeded the alert threshold relative to formal Vivago well-being events. Risk scores were computed for all 20 users using models trained on the full 3,265-row training dataset. To ensure consistency with the training exclusion criteria, lead time searches were restricted to days with `day_number` greater than 30, eliminating artefactual early alerts arising from baseline-period days on which z-score features were NaN-imputed to zero.

Three complementary lead time methods were evaluated for Task C. Method (a) measured the first alert before the yellow well-being flag at the operational threshold of 0.4, establishing whether the model detects deterioration earlier. Method (b) repeated this analysis at the lower threshold of 0.2 to assess whether a sub-threshold signal exists at an earlier date for any user. Method (c) measured the first alert before the first confirmed red well-being flag at threshold 0.4, which is the primary clinically actionable lead time result. Methods (a) and (b) contextualise the Task C findings by demonstrating that the model fires concurrently with the yellow flag at threshold 0.4, which motivated the state-based label design of Task G.

For Task G, lead time was measured relative to both the yellow well-being flag and the algorithmically detected change-point. The latter quantifies whether multi-feature machine learning detects deterioration onset earlier than single-feature statistical change detection applied to the same smoothed activity series.

3.9.4 SHAP Explainability Analysis

SHAP (SHapley Additive exPlanations) analysis was applied to the final Task C and Task G XGBoost models trained on the full dataset using TreeSHAP [39]. Global feature importance was summarised as mean absolute SHAP value across all training rows. Beeswarm plots were generated to visualise the directional effects of each top feature, the relationship between feature value (high or low) and the corresponding increase or decrease in predicted risk. Per-user SHAP analysis was additionally performed for the five deteriorating users to assess the consistency of feature importance rankings across individuals with distinct deterioration trajectories.

3.10 Proof-of-Concept Application

A standalone proof-of-concept dashboard was implemented using the Streamlit framework to demonstrate the operational deployment of the two primary models. The application loads three artefacts: the serialised Task C XGBoost model (`model.joblib`), the serialised Task G XGBoost model (`model_taskG.joblib`), and the ordered list of 49 feature column names (`feature_cols.joblib`). No feature scaling is applied at inference time, as XGBoost handles feature magnitude internally.

The dashboard is structured around five sequential components in the main page body, with a configuration sidebar on the left. A population overview panel at the top of the main page presents all 20 monitored users in an expandable table ranked by current Task C event risk score, displaying the Task G state detection score alongside each entry. Three summary metrics display the cohort-level count of high-risk, monitoring, and stable users, enabling care coordinators to perform population-level triage without navigating individual records.

Below the population overview, the individual risk assessment panel presents two side-by-side colour-coded status displays addressing the two research questions

simultaneously. The RQ2 panel (Task C) displays the probability of a critical adverse event within 14 days, with status levels of stable (below 0.30), monitoring (0.30–0.60), and high risk (above 0.60). The RQ1 panel (Task G) displays the probability that the user is currently in an active deterioration state, using thresholds of 0.20 for a monitoring alert and 0.40 for a confirmed deteriorating status. The lower Task G threshold reflects the label length imbalance in LOUO calibration described in Section 3.7.2, as users with shorter positive windows produce systematically lower calibrated probabilities, requiring a reduced operating threshold to achieve clinically useful sensitivity. Each panel displays a plain-language interpretation of the current score alongside the numerical value.

The risk score timeline displays both model outputs as simultaneous lines over a user-selectable date range, with colour-coded background bands at 0.30 and 0.60 providing immediate visual context. The default view covers the most recent 60 days. For users with confirmed well-being flags, a "jump to deterioration event" button in the sidebar shifts the view window to the 30-day period surrounding the red flag date, enabling retrospective demonstration of model behaviour at the time of actual clinical events.

The SHAP feature explanation panel displays the top ten feature contributions driving the current Task C risk score as a colour-coded horizontal bar chart, computed for the most recent day in the view window using TreeSHAP. Red bars indicate features increasing the predicted risk score; blue bars indicate features decreasing it. This panel translates the model output into specific, named behavioural indicators that a care coordinator can assess and communicate to colleagues.

A risk score history table below the SHAP panel displays Task C and Task G scores alongside raw feature values (daytime activity, sleep hours, and data completeness) for each day in the selected view window, sorted by date descending and supporting detailed inspection of individual monitoring days.

The sidebar provides user selection with current status labels, the deterioration event jump button, date range controls for the timeline and history table, and threshold reference guidelines. The application was tested locally against the pseudonymised Vivago dataset and requires no external API calls or network access at inference time. It was not deployed to a public server during the thesis period, consistent with the GDPR data handling requirements applicable to the pseudonymised dataset.

4 Results

4.1 Exploratory Data Analysis

4.1.1 Dataset Overview and State Classification

After daily aggregation, 4,220 user-days were obtained from 6,076,800 minute-level records. Device adherence was high: mean data completeness was 0.943, and fewer than 3% of user-days fell below the 0.60 completeness threshold. Cohort baseline means are summarised in Table 4.1.

Table 4.1: Cohort baseline means computed over the first 30 observation days per user.

Feature	Cohort mean
mean_activity (counts/min)	29.5
daytime_activity (counts/min)	35.4
sleep_hours	7.5
circadian_amplitude	31.8
peak_activity_hour	13.8
data_completeness	0.943

No temporal clustering of alarm events was observed across the observation period (Figure 4.1), confirming that alarm distribution did not introduce a seasonal or temporal bias that could affect cross-validation results.

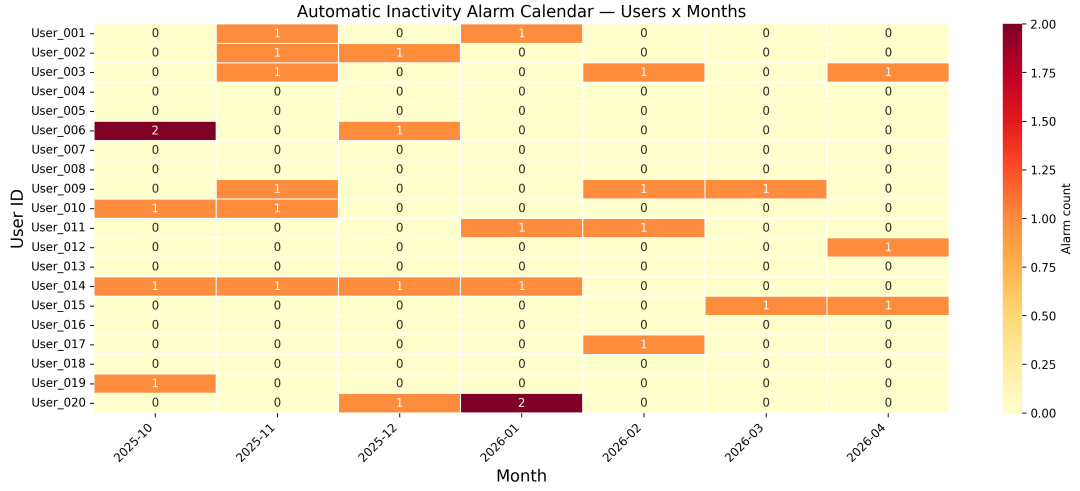


Figure 4.1: Calendar heatmap of automatic inactivity alarm events.

4.1.2 Prediction Horizon Analysis

Mann-Whitney U tests applied to normalised daytime activity z-scores across three candidate prediction horizons produced the results shown in Table 4.2 and Figure 4.2.

Table 4.2: Mann-Whitney U test results for pre-alarm vs non-alarm daytime activity z-scores across candidate prediction horizons.

Window	Pre-alarm mean	Non-alarm mean	p-value
Raw (no normalisation)	—	—	0.997 (not significant)
7-day	−0.525	−0.288	0.162 (not significant)
14-day	−0.534	−0.274	0.033 (significant)
28-day	−0.473	−0.254	0.014 (significant)

The raw comparison without normalisation yielded $p = 0.997$, confirming that cross-user variability in absolute activity levels completely masks within-user dete-

rioration signal. The 14-day and 28-day windows achieve significance ($p = 0.033$ and $p = 0.014$); the 7-day window does not ($p = 0.162$). As the 14-day window was the shortest horizon to achieve statistical significance, motivating its selection as the primary label horizon described in Section 3.6.

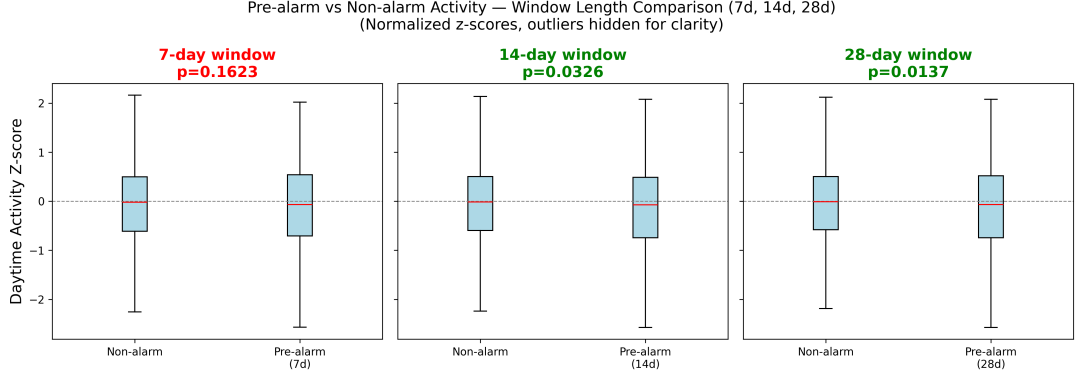


Figure 4.2: Mann-Whitney comparison of normalised daytime activity z-scores for pre-alarm versus non-alarm user-days across three prediction horizons.

4.1.3 Per-User Baseline Normalisation Validation

Z-score normalisation produced clear group-level separation. Across all normalised rows, the daytime activity z-score had a mean of -0.780 and standard deviation of 3.567 , with 213 rows (3.5% of the normalised dataset) winsorised at the ± 5 boundary prior to model training. The two-panel distribution (Figure 4.3) shows what drives this spread: stable users cluster tightly near zero (mean = $+0.049$, SD = 0.892), while deteriorating users produce a heavily left-skewed distribution (mean = -1.227 , SD = 4.312). The extreme negative values in the deteriorating group are a real signal rather than an artefact; rows with $z \leq -2$ had a mean daytime activity z-score of -9.854 , reflecting sustained, severe below-baseline activity during the deterioration phase. These rows were retained and winsorised rather than excluded, preserving the deterioration signal while preventing outlier influence on model training.

Table 4.3: Daytime activity z-score summary by user group after per-user normalisation.

Group	Mean z	Std z	Notes
Stable users	+0.049	0.892	Near-symmetric, centred at zero
Deteriorating users	-1.227	4.312	Left-skewed; rows ≤ -2 have mean -9.854
All rows: mean = -0.780 , SD = 3.567 ; 213 rows winsorised at ± 5 (3.5%)			

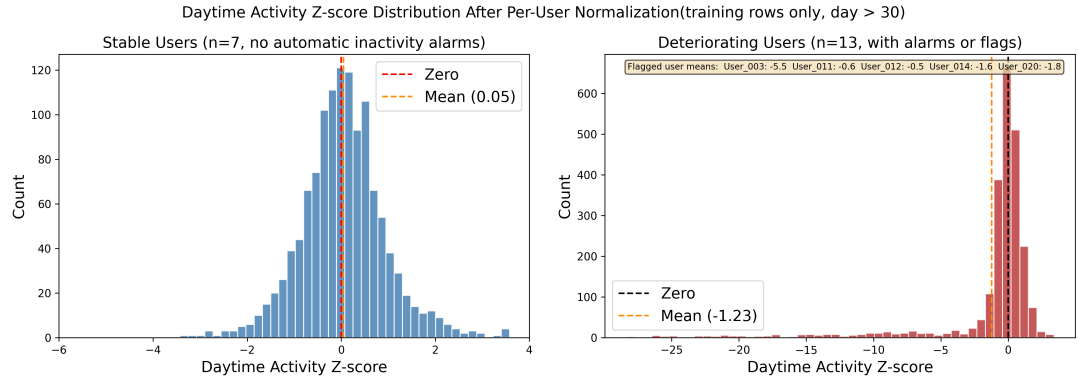


Figure 4.3: Daytime activity z-score distributions after per-user normalisation.

4.1.4 Feature Variability Analysis

Feature variability was assessed using the coefficient of variation (CV) for level features with non-zero means and standard deviation for zero-centred features. The top 15 features by variability score are presented in Table 4.4. Sleep fragmentation slope features showed the highest variability, followed by baseline deviation features for circadian amplitude and daytime activity, consistent with their empirical importance in the SHAP analysis reported in Sections 4.3.3 and 4.4.4.

Table 4.4: Top 15 features by variability score. CV is used for level features; standard deviation is used for zero-centred features.

Feature	Scoring method	Score
sleep_fragmentation_7d_slope	std	5.804
sleep_fragmentation_14d_slope	std	2.108
circadian_amplitude_baseline_dev	std	2.012
daytime_activity_baseline_dev	std	1.997
mean_activity_baseline_dev	std	1.618
circadian_amplitude_zscore_capped	std	1.485
daytime_activity_zscore_capped	std	1.471
mean_activity_zscore_capped	std	1.428
IV	CV	0.452
sleep_hours_zscore_capped	std	0.323
circadian_amplitude	CV	0.279
circadian_amplitude_7d_mean	CV	0.278
circadian_amplitude_14d_mean	CV	0.277
daytime_activity	CV	0.257
daytime_activity_7d_mean	CV	0.256

4.2 Exploratory Label Analysis and Baseline Comparators

4.2.1 Label Distribution Across Prediction Tasks

The label distribution across all seven prediction tasks is summarised in Table 4.5. Positive rates varied from 3.4% for Task C (red flag, 14 days) to 23.3% for Task F (combined alarm, 28 days). Task G contained 320 positive training rows (9.8%), drawn from four users with valid change-point dates. The label length imbalance in task G presented by User_003 contributing 166 positive days versus User_011’s 36 days, is the primary cause of the threshold calibration constraint discussed in Section 4.4.2.

Table 4.5: Label distribution across all seven prediction tasks (3,265 training-eligible rows).

Task	Label column	Role	Positives	Rate
A	label_inactivity_7d	Exploratory	150	4.6%
B	label_inactivity_14d	Exploratory	259	7.9%
C	label_redflag_14d	Primary (RQ2)	112	3.4%
D	label_redflag_28d	Exploratory	167	5.1%
E	label_combined_alarm_14d	Exploratory	442	13.5%
F	label_combined_alarm_28d	Exploratory	760	23.3%
G	label_change-point	Primary (RQ1)	320	9.8%

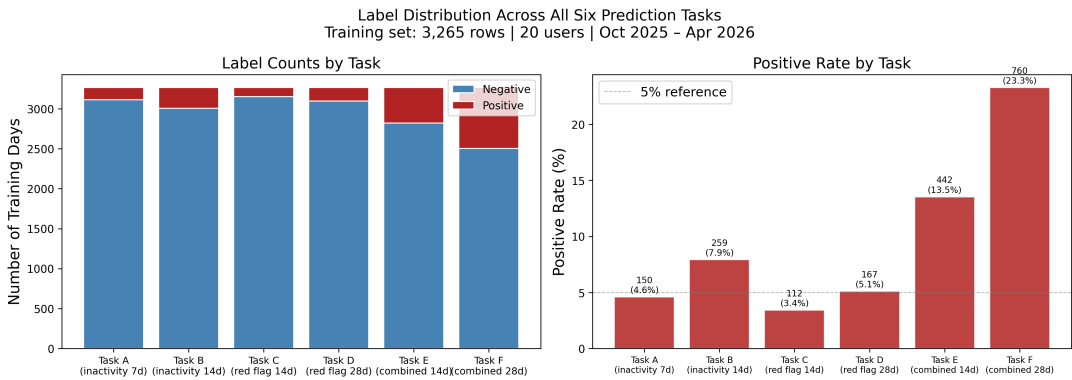


Figure 4.4: Label class distribution across Tasks A through F.

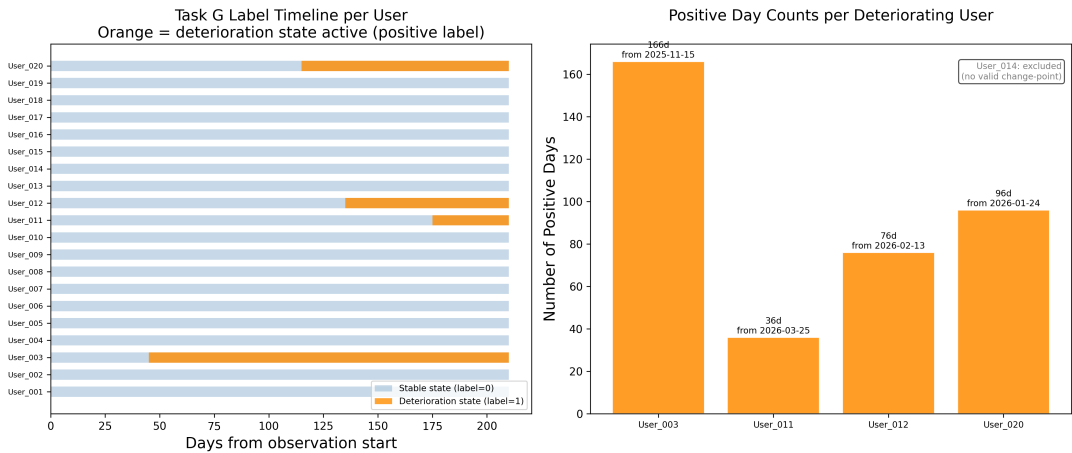


Figure 4.5: Task G state label structure.

4.2.2 Exploratory Task Performance (Tasks A, B, D, E, F)

Logistic Regression, Random Forest, and XGBoost were evaluated under LOUO cross-validation across all six event-window prediction tasks. XGBoost results use the tuned hyperparameter configuration. Task C is the primary RQ2 model (bold); Tasks A, B, D, E, and F are exploratory comparators. The mean PR-AUC and recall at threshold 0.4 are reported in Table 4.6. No-skill baselines equal the positive rate per task: Task A = 0.046, Task B = 0.079, Task C = 0.034, Task D = 0.051, Task E = 0.135, Task F = 0.233.

Table 4.6: LOUO CV results for all three models across all six event-window prediction tasks.

Task	Model	PR-AUC	Recall	No-skill	Ratio
<i>Task A: Automatic inactivity alarm, 7-day horizon</i>					
	LR	0.107 ± 0.083	0.418	0.046	$2.3\times$
	RF	0.093 ± 0.056	0.268	0.046	$2.0\times$
	XGB	0.085 ± 0.036	0.030	0.046	$1.8\times$
<i>Task B: Automatic inactivity alarm, 14-day horizon</i>					
	LR	0.152 ± 0.079	0.485	0.079	$1.9\times$
	RF	0.152 ± 0.070	0.392	0.079	$1.9\times$
	XGB	0.208 ± 0.106	0.128	0.079	$2.6\times$
Task C: Red well-being flag, 14-day horizon (Primary RQ2)					
	LR	0.938 ± 0.033	0.961	0.034	$27.6\times$
	RF	0.981 ± 0.014	0.897	0.034	$28.9\times$
	XGB	0.975 ± 0.020	0.885	0.034	$28.7\times$
<i>Task D: Red well-being flag, 28-day horizon</i>					
	LR	0.754 ± 0.152	0.645	0.051	$14.8\times$
	RF	0.788 ± 0.059	0.641	0.051	$15.5\times$
	XGB	0.783 ± 0.060	0.589	0.051	$15.4\times$
<i>Task E: Combined alarm (inactivity + manual), 14-day horizon</i>					
	LR	0.257 ± 0.171	0.605	0.135	$1.9\times$
	RF	0.229 ± 0.116	0.393	0.135	$1.7\times$
	XGB	0.265 ± 0.112	0.200	0.135	$2.0\times$
<i>Task F: Combined alarm (inactivity + manual), 28-day horizon</i>					
	LR	0.409 ± 0.215	0.639	0.233	$1.8\times$
	RF	0.362 ± 0.151	0.683	0.233	$1.6\times$
	XGB	0.375 ± 0.140	0.340	0.233	$1.6\times$

4.2.3 Label Design Observation

The six-task comparison tells a clear story. Task C stands in a category of its own: all three models achieved PR-AUC values between 0.938 and 0.981, which is roughly 28 times above the no-skill baseline of 0.034.

Tasks A and B, predicting automatic inactivity alarms at 7 and 14-day horizons, produced PR-AUC values of 0.085-0.107 and 0.152-0.208 across the three models. The best result, tuned XGBoost on Task B at 0.208, is 2.6 times above the 0.079 no-skill baseline, a meaningful improvement but nowhere near deployable. Recall at the threshold of 0.4 was near zero for XGBoost on Task A (0.030), confirming that even the tuned model struggles to generate confident predictions for this noisy, short-horizon label.

Task D (red flag, 28 days) reached PR-AUC values of 0.754-0.788 across models – respectable in isolation, but substantially below Task C despite targeting the same event type. The wider horizon is the likely cause, as a 28-day positive window labels days that may still be weeks away from any real behavioural change, introducing noise that narrows the discriminability gap between deteriorating and stable days.

Tasks E and F, combining inactivity and manual alarm events, reached 0.229-0.265 and 0.362-0.409, respectively; above their no-skill baselines of 0.135 and 0.233, but the heterogeneous label construction limits what any model can achieve regardless of architecture.

Logistic Regression achieved a competitive PR-AUC for Task C (0.938), only marginally below Random Forest (0.981) and tuned XGBoost (0.975). It is informative that a linear classifier comes this close to the ensemble models on Task C; the deterioration signal captured by red well-being flags is largely linearly separable, driven by sustained slope features that do not require complex interactions to detect. Across all five exploratory tasks, the model choice made little practical difference. What mattered was the label.

4.3 RQ2 – Event Prediction (Task C)

4.3.1 LOUO Cross-Validation Results

The tuned XGBoost configuration (`max_depth=3`, `learning_rate=0.057`, `n_estimators=356`) was applied to Task C with `scale_pos_weight = 28.15`. The final Task C model was fit on all 3,265 training-eligible rows and serialised as `model.joblib` (314.8KB). Full per-user LOUO results for Task C XGBoost are as reported in Table 4.6 ($\text{PR-AUC} = 0.975 \pm 0.020$, $\text{recall} = 0.885$).

4.3.2 Threshold Analysis

Threshold analysis was conducted on LOUO out-of-fold predictions for Task C XGBoost, computing precision, recall, and F1 at decision thresholds from 0.05 to 0.95. Results at the selected operating threshold of 0.4 are presented in Table 4.7.

Table 4.7: Task C XGBoost threshold analysis at threshold = 0.4. Metrics computed on pooled LOUO out-of-fold predictions across all five evaluable users.

Threshold	Recall	Precision	F1
0.40	0.893	0.775	≈ 0.83

The recall target of ≥ 0.70 was met at threshold 0.4 with a substantial margin. F1 remained stable from threshold 0.40 to approximately 0.85, indicating a wide operating region of acceptable precision-recall balance.

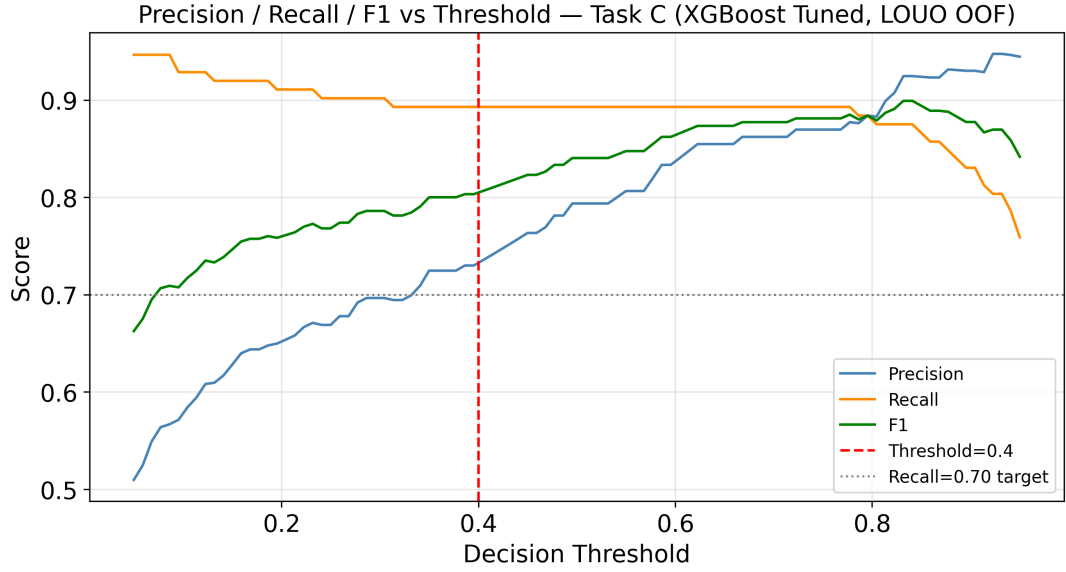


Figure 4.6: Precision, recall and F1 vs decision threshold for the Task C XGBoost model (LOUO out-of-fold predictions).

4.3.3 SHAP Feature Importance

SHAP analysis applied to the final Task C XGBoost model identified 14-day activity slope features as the primary predictive drivers, as they dominate by a factor of more than two over all others. The two highest-ranked features were 14-day linear slope coefficients for daytime and mean activity, each exceeding the third-ranked feature by more than a factor of two. Beeswarm analysis confirmed clinically consistent directional effects for all six features; declining slopes produced large positive SHAP contributions increasing predicted risk, and below-baseline activity and weakening circadian amplitude similarly increased risk. No feature exhibited a counterintuitive directional effect.

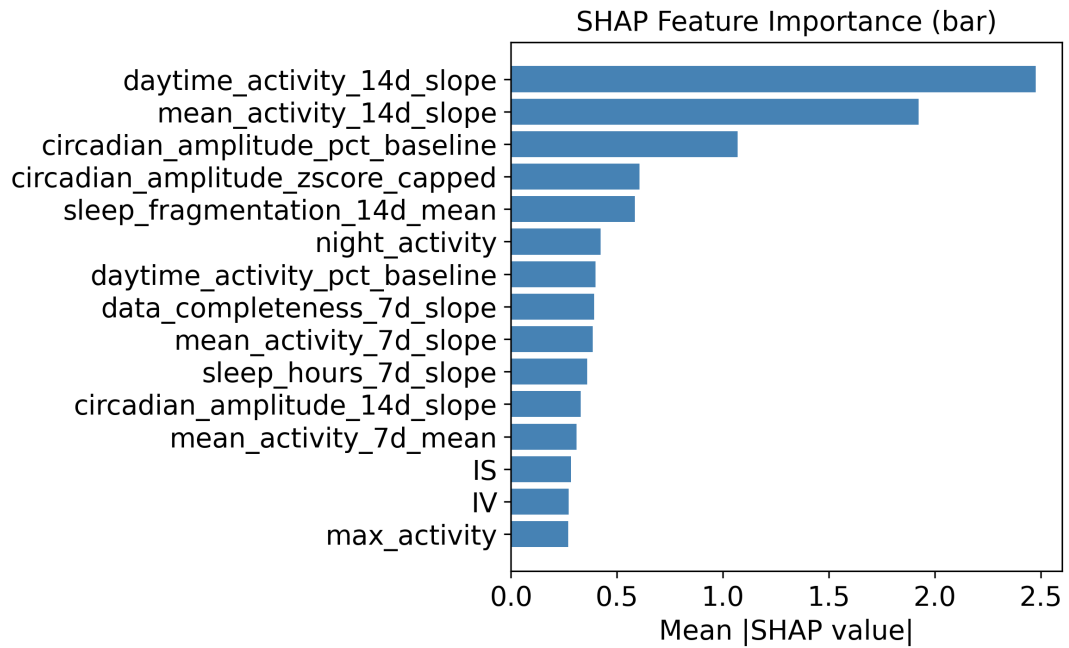


Figure 4.7: Mean absolute SHAP values for the top 15 Task C features.

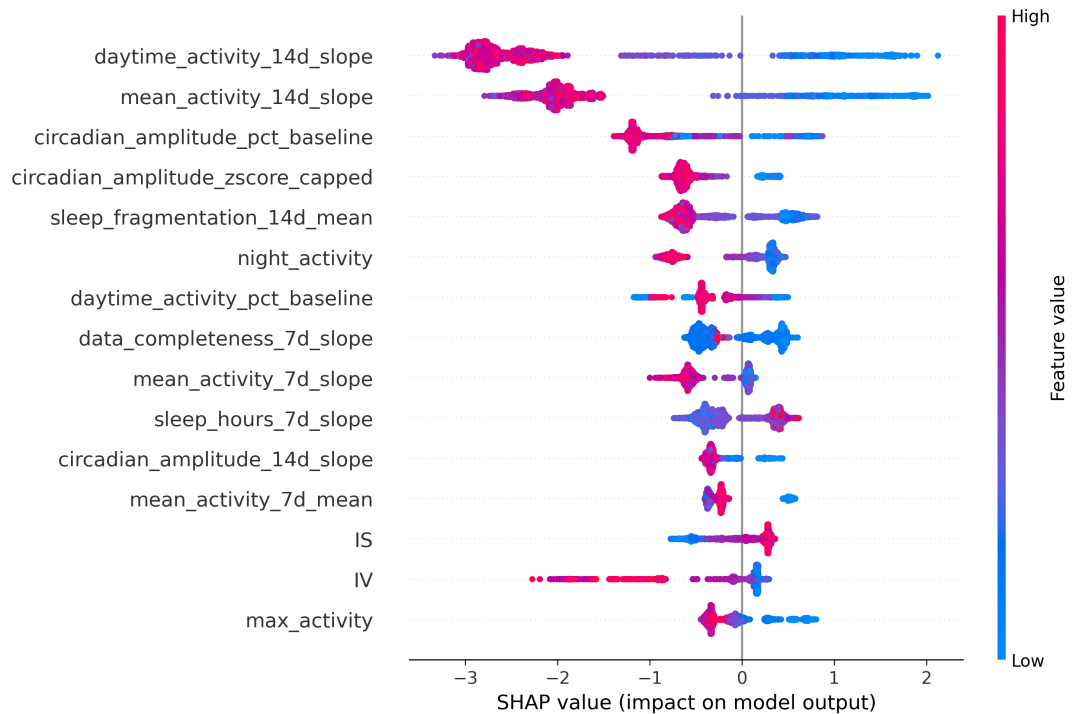


Figure 4.8: SHAP beeswarm plot for Task C.

Per-user SHAP analysis confirmed that `daytime_activity_14d_slope` ranked

first for Users 003, 011, 012, and 020, while `mean_activity_14d_slope` ranked first for User_014.

Both features, `circadian_amplitude_pct_baseline` and `mean_activity_pct_baseline` appeared in the top five for all five deteriorating users, indicating a generalisable deterioration signature across users with distinct trajectories and onset dates.

4.3.4 Risk Score Trajectories

Daily Task C risk scores were computed for all 20 users using the full dataset model. Risk scores rose from near-zero baseline values to high values (frequently 1.0) surrounding each deteriorating user’s yellow flag date, while remaining below 0.30 for all three confirmed alarm-free stable users throughout the full observation period. The task C risk trajectory for User_003 is shown in Figure 4.9. Risk crosses the threshold of 0.4 on 26 January 2026, exactly 14 days before the first red flag on 9 February 2026. Grey band: 30-day baseline warm-up period. For User_011 it’s visible in Figure 4.10. Risk scores fluctuate and then rise abruptly in late March 2026, crossing the threshold on 29 March 2026, exactly 14 days before the red flag on 12 April 2026. The flat early-period score reflects the stable phase of this dynamic-onset user. The Task C risk trajectories for User_012 are shown in Figure 4.11. Risk crosses the threshold on 21 March 2026, exactly 14 days before the red flag on 4 April 2026. For User_014, the model crosses the threshold on 23 February 2026, exactly 14 days before the red flag on 9 March 2026, Figure 4.12. For User_020 in Figure 4.13, the risk fluctuates till mid-March and then rises exactly 14 days before the red flag on 28 March 2026.

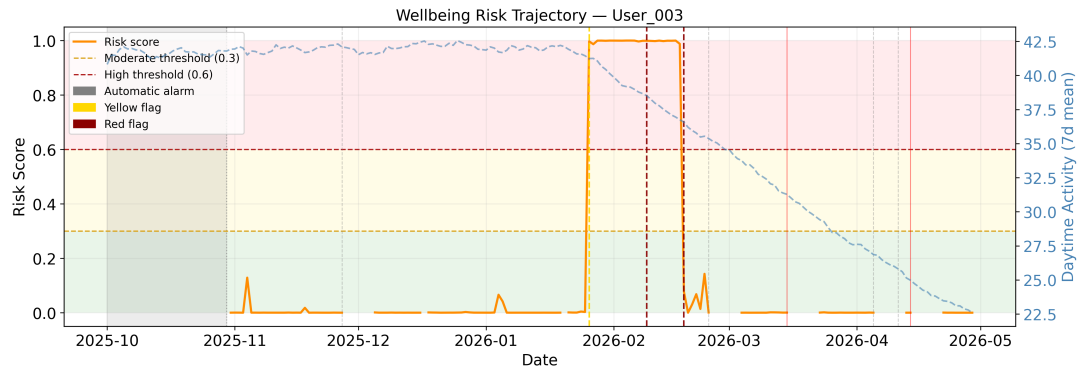


Figure 4.9: Task C risk trajectory for User_003.

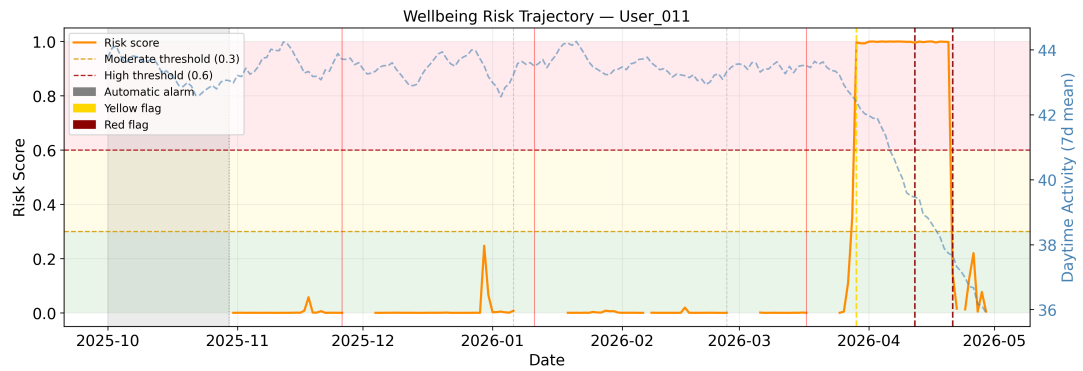


Figure 4.10: Task C risk trajectory for User_011.

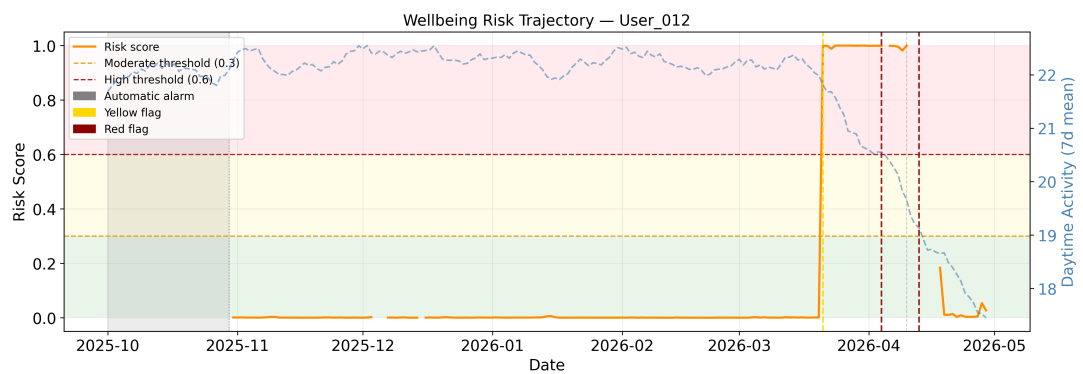


Figure 4.11: Task C risk trajectory for User_012.

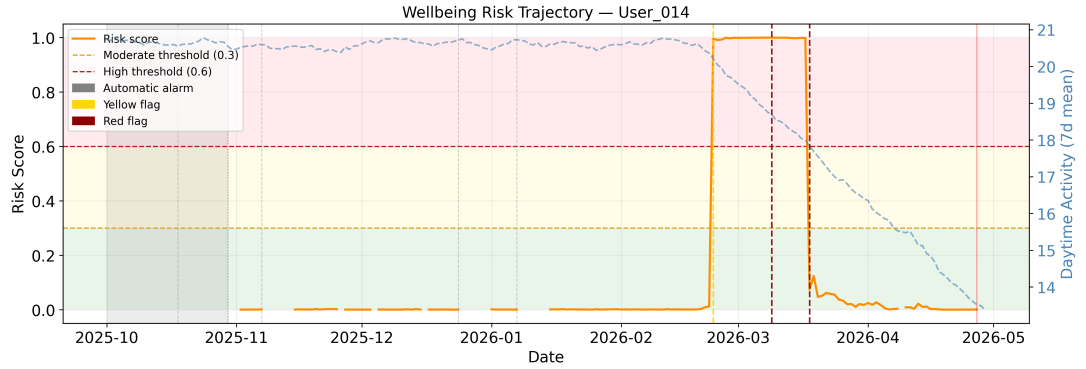


Figure 4.12: Task C risk trajectory for User_014.

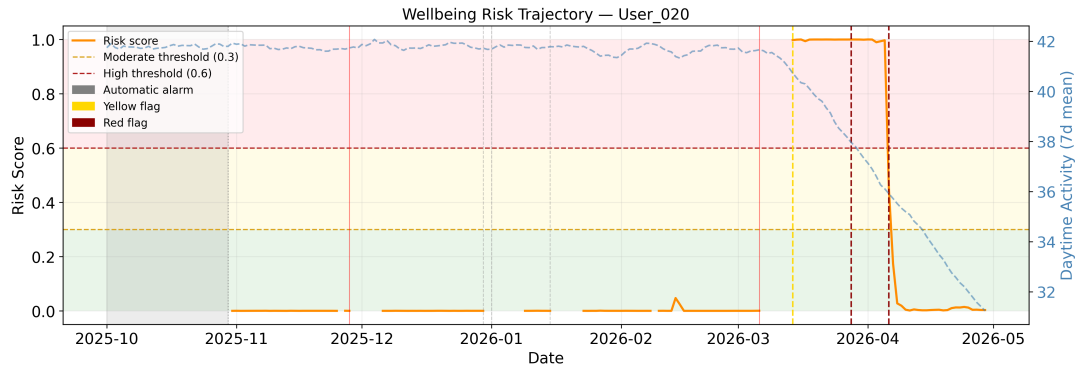


Figure 4.13: Task C risk trajectory for User_020.

4.3.5 Lead Time Validation

Lead time analysis was conducted using three complementary methods, all restricted to post-baseline dates ($\text{day_number} > 30$). Method (c) is the primary clinically actionable result. Methods (a) and (b) are reported as contextual findings that motivate the Task G state detection approach. Results are summarised in Table 4.8.

Table 4.8: Task C lead time analysis. All methods were restricted to post-baseline dates ($\text{day_number} > 30$).

Method	Model	Reference event	Mean lead	Median lead
(a)	Task C	Yellow flag ($t=0.4$)	0.0 days	0.0 days
(b)	Task C	Yellow flag ($t=0.2$)	17.8 days	0.0 days
(c)	Task C	First red flag ($t=0.4$)	14.0 days	14.0 days

Method (a): at threshold 0.4, Task C first crossed the alert boundary on the same day as the Vivago yellow flag for all five deteriorating users (mean = 0.0 days, SD = 0.0). Method (b): at threshold 0.2, a sub-threshold signal was identified for User_011 beginning 89 days before the yellow flag; the remaining four users showed zero lead time. Method (c): at threshold 0.4 versus the first red flag, Task C provided exactly 14 days of lead time for all five deteriorating users without exception (mean = 14.0 days, SD = 0.0). Per-user results are presented in Table 4.9.

Table 4.9: Task C vs the first red flag: per-user lead times at threshold 0.4.

User	First red flag	First alert	Lead (days)
User_003	2026-02-09	2026-01-26	14
User_011	2026-04-12	2026-03-29	14
User_012	2026-04-04	2026-03-21	14
User_014	2026-03-09	2026-02-23	14
User_020	2026-03-28	2026-03-14	14
Mean			14.0
SD			0.0

4.3.6 Negative Control Validation

All three confirmed alarm-free stable users (User_004, User_016, User_018 had zero alarm events of any type across the full observation period) maintained Task C risk scores below 0.30 throughout the full observation period. All three maintained scores below 0.30 for 100% of monitoring days across the full observation period.

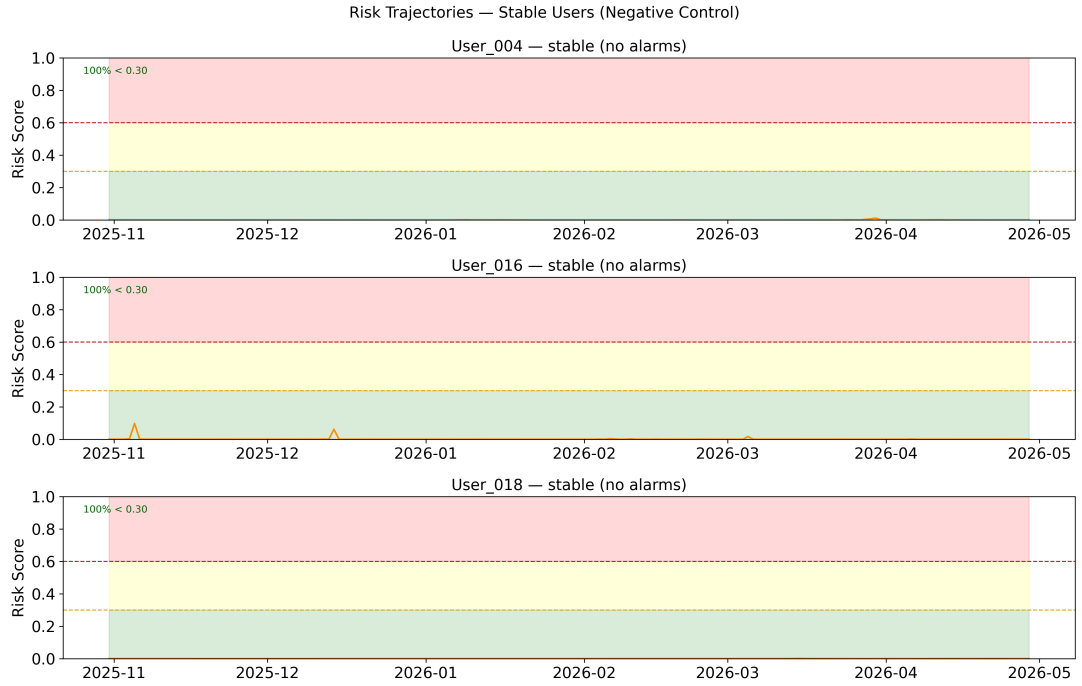


Figure 4.14: Task C risk trajectories for the three confirmed alarm-free stable users (negative control).

4.4 RQ1 – Behavioural State Detection (Task G)

4.4.1 Change-Point Detection Results

Change-points were detected using the PELT algorithm applied to the smoothed daytime activity series (`daytime_activity_7d_mean`). Valid downward change-points were identified for four of five deteriorating users. The detection procedure is described in Section 3.7.2; the resulting dates and lead times before yellow well-being flags are presented in Table 4.10.

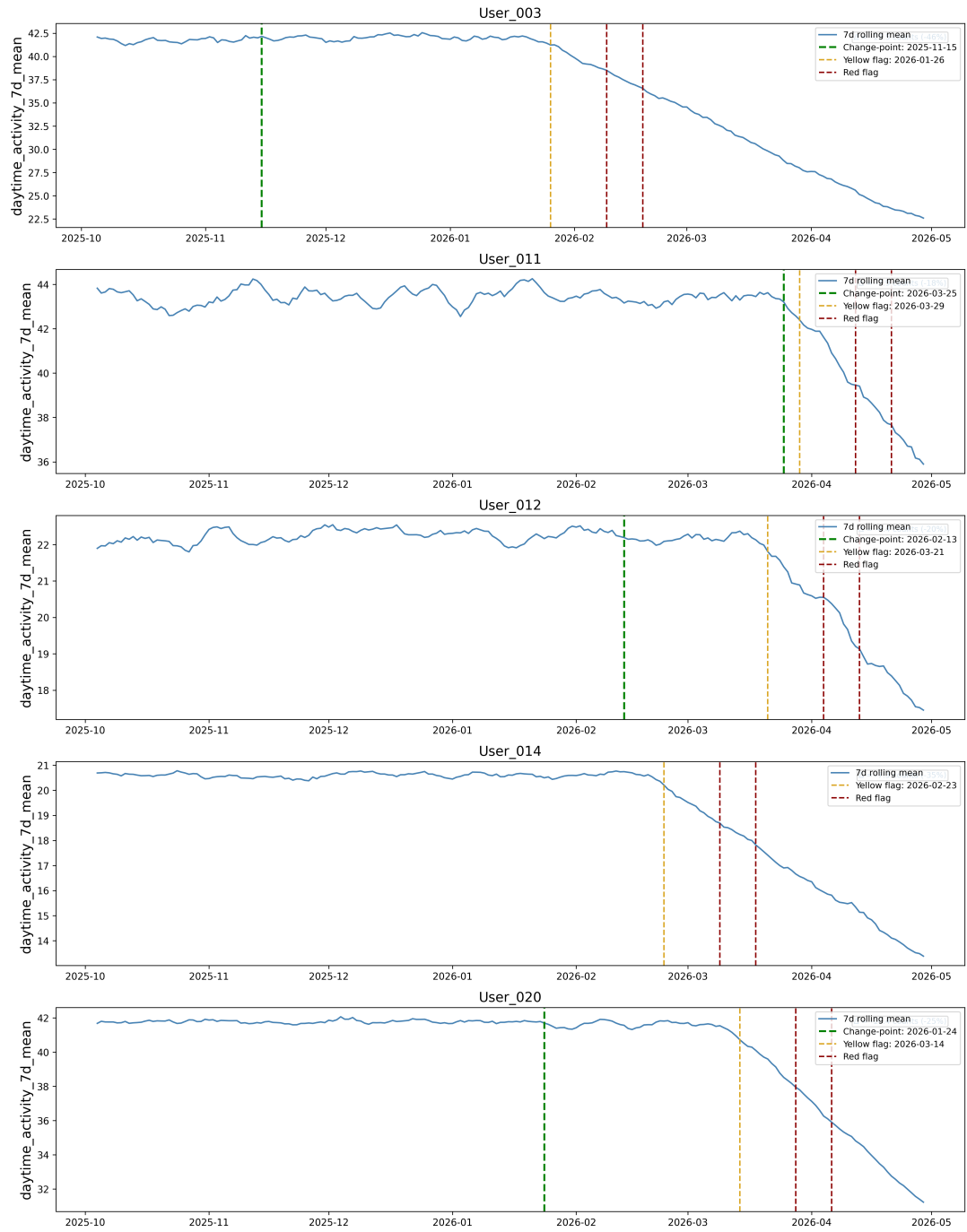


Figure 4.15: Change-point detection validation for all five deteriorating users.

Table 4.10: Detected change-point dates and lead times before yellow well-being flags.

User	Change-point date	Yellow flag date	Lead (days)
User_003	2025-11-15 (manual override)	2026-01-26	72
User_011	2026-03-25	2026-03-29	4
User_012	2026-02-13	2026-03-21	36
User_014	None detected	2026-02-23	N/A
User_020	2026-01-24	2026-03-14	49

4.4.2 LOUO Cross-Validation Results

Task G XGBoost achieved a mean PR-AUC of 0.829 (± 0.140) under LOUO cross-validation across the four users with valid change-point dates. The no-skill baseline was 0.098. Per-user results are presented in Table 4.11.

Table 4.11: Task G XGBoost LOUO CV results.

User	PR-AUC	ROC-AUC	Recall (t=0.2)	n positive
User_003	0.952	0.642	0.512	129/144
User_011	0.948	0.975	1.000	35/146
User_012	0.698	0.630	0.015	68/170
User_020	0.718	0.684	0.000	88/149
Mean	0.829 $\pm$ 0.140	—	0.382 $\pm$ 0.476	—

All four evaluated users exceeded the no-skill baseline. Users 003 and 011 achieved high PR-AUC values (0.952 and 0.948). Users 012 and 020 showed lower PR-AUC values (0.698 and 0.718) and near-zero recall, reflecting a LOUO calibration constraint caused by label length imbalance rather than a fundamental model failure.

4.4.3 Threshold Analysis

A threshold sweep across the range 0.05 to 0.60 was conducted on Task G LOUO out-of-fold predictions (Figure 4.16). The OOF probability distributions for positive and negative class rows overlap substantially for Users 012 and 020, indicating that the discriminative limitation for these users is not addressable through threshold

adjustment. Recall does not reach 0.50 at any threshold in the evaluated range (0.05–0.60), confirming that low recall reflects overlapping probability distributions rather than suboptimal threshold selection. PR-AUC, which evaluates ranking quality across all thresholds simultaneously, is therefore reported as the primary and sole summary metric for Task G throughout.

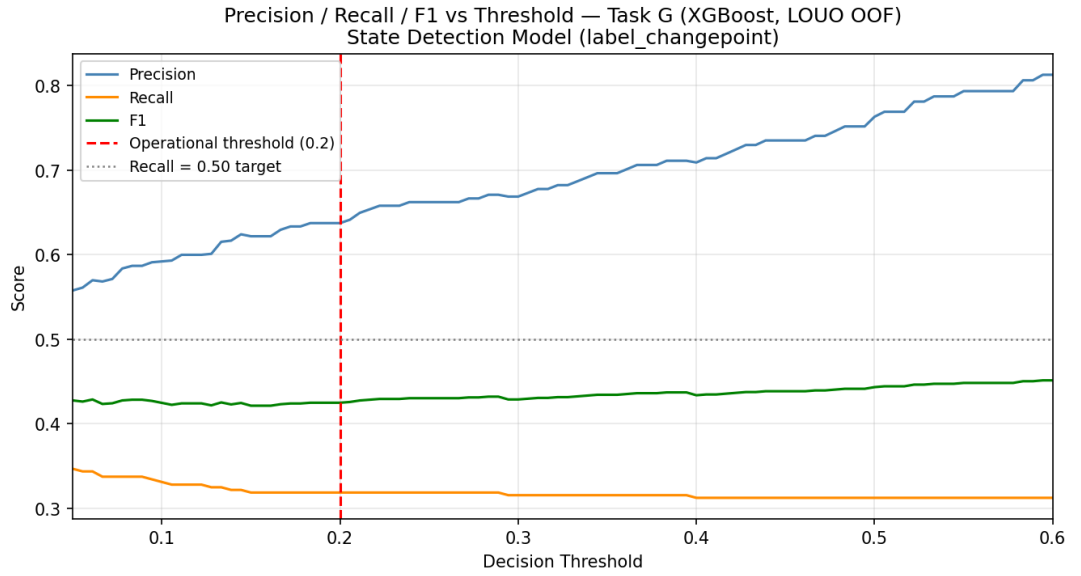


Figure 4.16: Precision, recall and F1 vs decision threshold for the Task G XGBoost model (LOUO out-of-fold predictions).

4.4.4 SHAP Feature Importance

SHAP analysis applied to the final Task G XGBoost model revealed that 14-day mean (level) features dominate in contrast to the slope-dominated profile of Task C.

The top four features were all 14-day rolling mean values, with circadian amplitude leading by a substantial margin. Per-user SHAP analysis confirmed that `circadian_amplitude_14d_mean` appeared in the top five for Users 003, 012, 014, and 020, while `daytime_activity_14d_slope` appeared for Users 003, 011, 012, and 020.

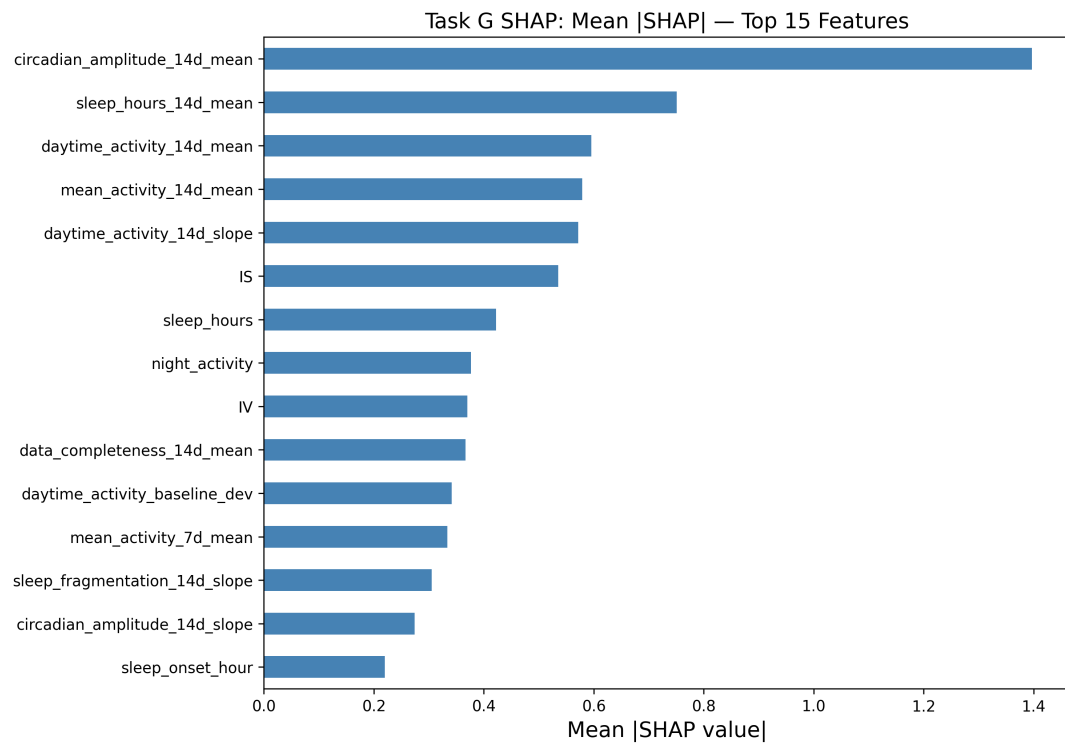


Figure 4.17: Mean absolute SHAP values for the top 15 Task G features.

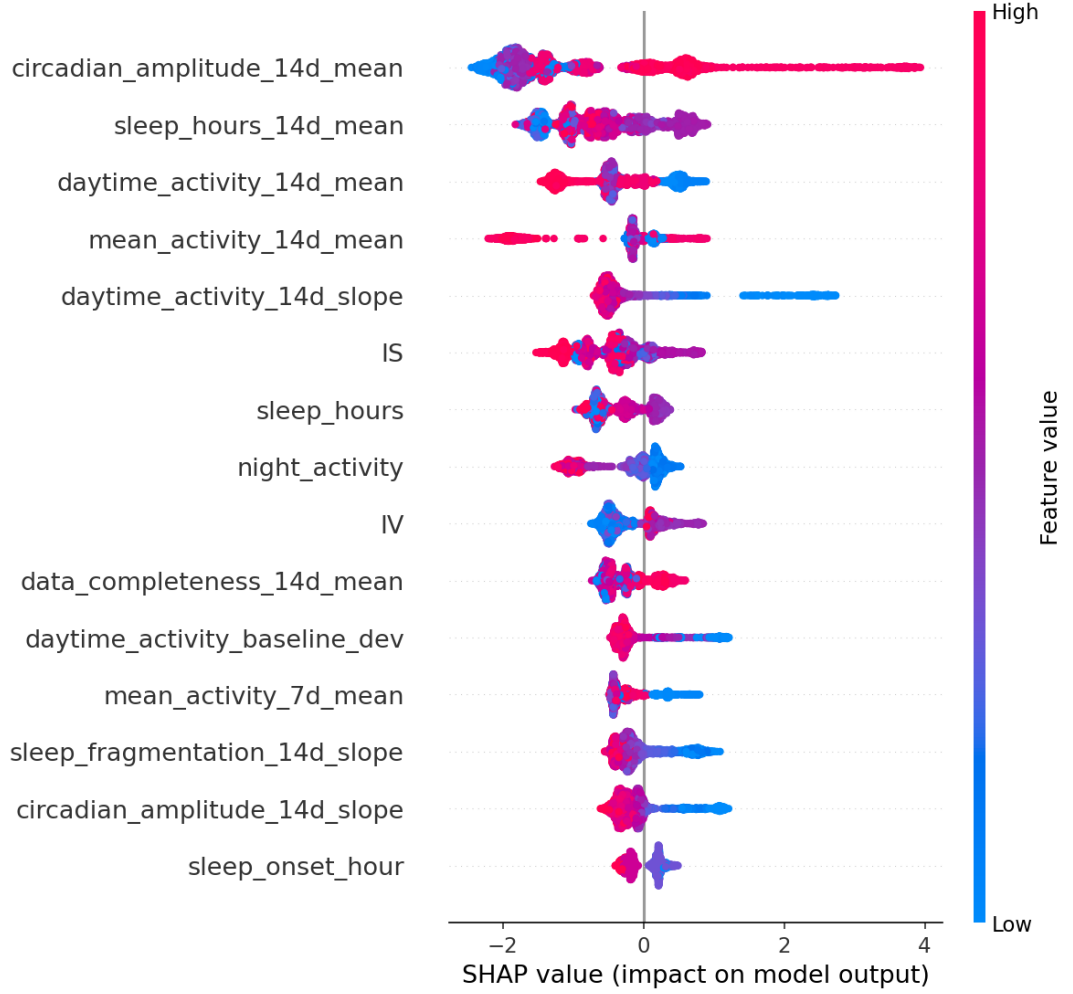


Figure 4.18: SHAP beeswarm for Task G.

4.4.5 Lead Time Analysis

Post-baseline lead time analysis (restricted to `day_number > 30`) produced the results in Table 4.12. The column `lead_days_before_cp` is positive when the model's first alert preceded the algorithmically detected change-point date.

The mean lead time before yellow flags was 56.0 days (median 55.5 days). The model fired before the algorithmically detected change-point for two of four users (User_003 by 3 days, User_020 by 60 days) and concurrently with the change-point for Users 011 and 012.

Table 4.12: Task G lead time results (post-baseline dates only).

User	First alert	Yellow flag	Lead vs yellow	Lead before CP
User_003	2025-11-12	2026-01-26	75 days	3 days
User_011	2026-03-25	2026-03-29	4 days	0 days
User_012	2026-02-13	2026-03-21	36 days	0 days
User_020	2025-11-25	2026-03-14	109 days	60 days
User_014	None	2026-02-23	N/A	N/A
Mean	—	—	56.0 days	15.8 days
Median	—	—	55.5 days	1.5 days

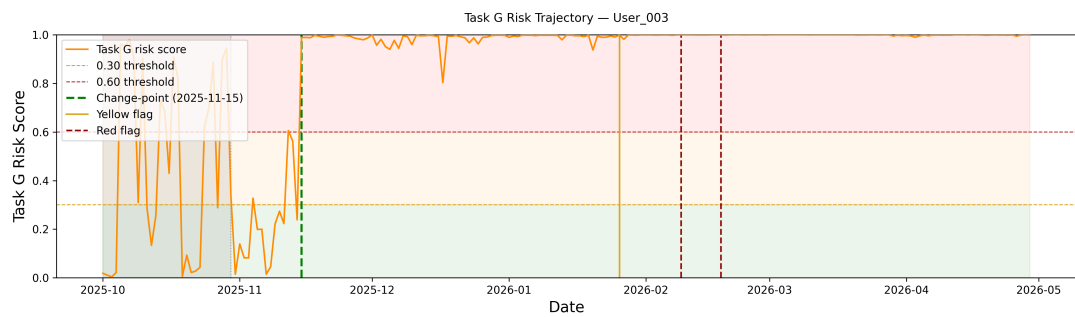


Figure 4.19: Task G risk trajectory for User_003.

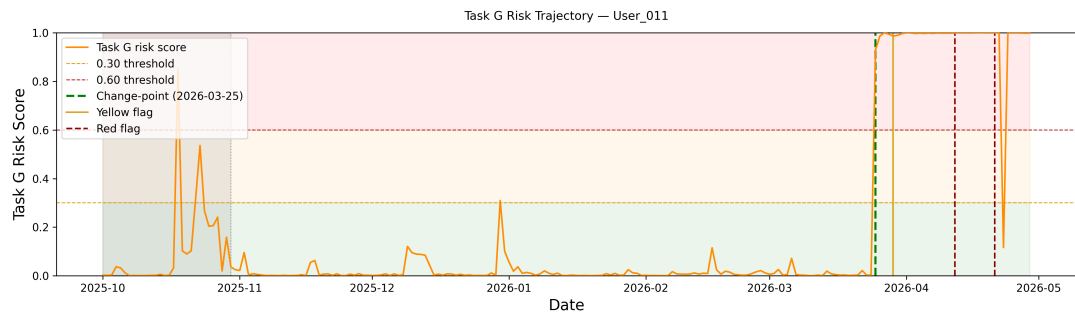


Figure 4.20: Task G risk trajectory for User_011.

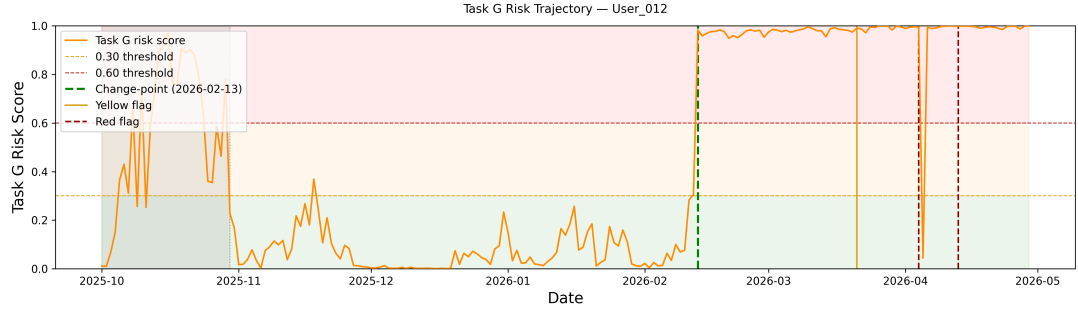


Figure 4.21: Task G risk trajectory for User_012.

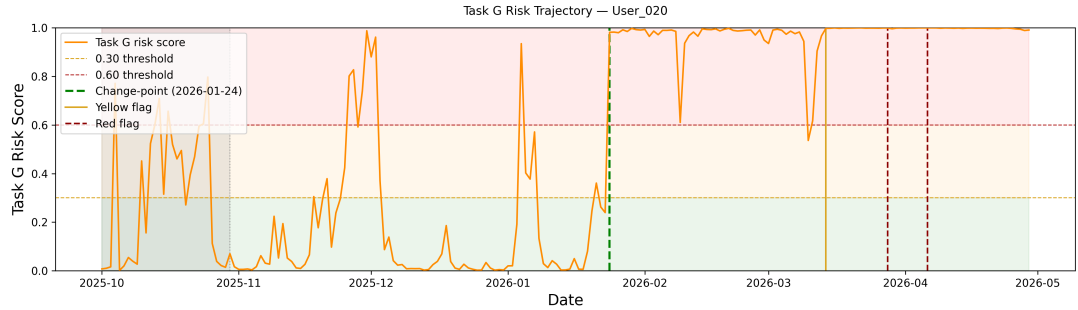


Figure 4.22: Task G risk trajectory for User_020.

4.4.6 Post-Baseline Validation

Risk scores during days 31–60 (the first valid month after the baseline warm-up period) were compared between deteriorating and stable users using the Mann-Whitney U test. The results are presented in Table 4.13.

Table 4.13: Task G post-baseline validation: risk score comparison (days 31–60 and 31–90).

Window	Deteriorating mean	Stable mean	p-value
Days 31–60	0.230	0.002	0.029
Days 31–90	0.259	0.002	0.029

Individual risk scores during days 31–60 are reported in Table 4.14. User_003 showed an already-elevated risk (0.603) during this early post-baseline window; User_011, consistent with its acute late-onset deterioration pattern, showed a score

of 0.011, indistinguishable from stable users during this period.

Table 4.14: Individual Task G risk scores during days 31–60 (post-baseline).

User	Mean score (days 31–60)	Group
User_003	0.603	Deteriorating
User_011	0.011	Deteriorating
User_012	0.106	Deteriorating
User_020	0.201	Deteriorating
User_004	0.005	Stable
User_016	0.001	Stable
User_018	0.000	Stable

Two distinct risk patterns were observed among deteriorating users. Users 003, 012, and 020 exhibited elevated risk scores from November 2025 onwards, months before any system-generated alert, consistent with a chronic risk profile driven by persistently low circadian amplitude and reduced sleep duration. User_011 showed risk scores near zero throughout the early and middle observation period, rising sharply to 1.0 only at the onset of the acute decline in March 2026.

This chapter presented the empirical results across all seven prediction tasks, including performance metrics, lead-time analysis, and user-level evaluation. Detailed interpretation is provided in Chapter 5.

5 Discussion

5.1 Label Design

The most consequential methodological finding of this study is not a model performance metric but rather a structural insight about how prediction targets are defined. Tasks A through F all frame the learning problem as event-window prediction; a given day is labelled positive if a system event will occur within the next N days. This framing is operationally intuitive, but it creates a training signal that systematically suppresses early detection: days that precede an event by more than the window length are labelled negative even when the user’s behavioural profile is already deteriorating. The model cannot learn to fire early on these days because the training data explicitly tells it they are negative examples.

Task G removes this constraint entirely. Every day from the algorithmically detected change-point onwards is labelled positive, regardless of how far away any system alert might be. The model therefore trains on the full deterioration profile, weeks or months of data, rather than just the narrow pre-event window. The result is not subtle: mean lead time before yellow well-being flags jumped from 0 days under Task C to 56 days under Task G. This implies that for future system development, the direct and practical approach is state-based labels over event-window labels.

The complementarity of the two label designs maps naturally onto two distinct clinical questions. Task C is suited to answering "When is this resident likely to

need intervention?" a question about event timing. Task G is suited to answering "Is this resident currently at elevated chronic risk?" a question about current status. Both questions have clinical value, and neither is redundant given the other.

5.2 Task C Performance

The Task C XGBoost model achieved $\text{PR-AUC} = 0.975$ under leave-one-user-out cross-validation, with $\text{recall} = 0.893$ and $\text{precision} = 0.775$ at the operating threshold. This performance level, particularly for a 3.4% positive class under the most stringent available cross-validation protocol, substantially exceeds what the no-skill baseline (0.034) would suggest is feasible from this data. The result is consistent with and contextualises the Vivago system's own detection mechanism: the red well-being flag targets exactly the same behavioural signals, which is the sustained low daytime activity and weakened circadian rhythm that the SHAP analysis identified as the primary predictive features (14-day slopes in daytime and mean activity, and circadian amplitude relative to personal baseline).

The concurrent detection finding (Method a: lead time = 0 days for all five deteriorating users) is a validation of the signal structure rather than a limitation of the model. The Task C model fires at the same behavioural threshold that triggers the Vivago yellow flag because both systems are responding to the same underlying phenomenon. What Task C adds is a quantified probability rather than a binary flag and SHAP-based feature attribution that identifies the specific behavioural drivers of each prediction. The 14-day lead before red flags (Method c) represents the clinically actionable window: an alert generated when the yellow flag fires gives the care team two weeks before the most severe confirmed deterioration event.

The zero-false-alarm result for the three confirmed stable users across seven months of continuous monitoring is particularly noteworthy. Alert fatigue is a documented barrier to clinical AI adoption, as systems that generate excessive false

positives lose credibility among care staff [11]. The model’s ability to distinguish the chronically low but stable activity of a habitually sedentary resident (User_004, User_016, User_018) from the declining trajectory of a genuinely deteriorating user reflects the effectiveness of per-user normalisation as the model reasons about each individual relative to their own established pattern, not against a population norm.

This finding aligns with the broader literature on personalised versus pooled modelling in elderly care populations. Chen et al. [7] highlighted that population-level norms poorly represent individual baselines in heterogeneous cohorts and that within-subject normalisation is a methodological prerequisite for meaningful deterioration signal extraction. The statistical validation in this study $p = 0.997$ for raw absolute values versus $p = 0.033$ for normalised values at the 14-day window provides direct support for this claim.

The near-identical PR-AUC values achieved by Logistic Regression (0.938), Random Forest (0.981), and XGBoost (0.975) for Task C suggest that it is linearly separable: a simple linear classifier already achieves high discrimination, suggesting that the underlying behavioural signature of red flag deterioration is a relatively clean, sustained pattern rather than a complex non-linear interaction. This is consistent with the SHAP finding that the two slope features account for the overwhelming majority of predictive weight, with all other features contributing comparatively little.

The SHAP feature profile confirms why Task C works as well as it does. The two highest-ranked features, `daytime_activity_14d_slope` and `mean_activity_14d_slope` each exceed the third-ranked feature by more than a factor of two in mean absolute SHAP value. Declining slopes produced large positive SHAP contributions, increasing predicted risk; below-baseline activity and weakening circadian amplitude similarly increased risk. No feature exhibited a counterintuitive directional effect. Task C’s use of formal red wellbeing flags as labels, events representing confirmed,

sustained deterioration rather than acute system-triggered alarms, provides a clean training signal that supports reliable directional attribution. When the goal is a clinically interpretable model, label quality matters as much as label quantity.

5.3 Task G Performance

The Task G model achieved a mean PR-AUC = 0.829 across four users with valid change-points. All four exceeded the no-skill baseline of 0.098 by a substantial margin, and the post-baseline validation confirmed statistically significant separation between deteriorating and stable users in days 31–60 of the observation period (Mann-Whitney $p = 0.029$). Two distinct detection patterns were observed: a chronic risk profile for Users 003, 012, and 020, with elevated scores from early in the post-baseline period, driven by persistently low circadian amplitude and sleep hours, and a dynamic detection pattern for User_011, where risk remained near zero until the acute March 2026 decline and then spiked immediately.

The November 2025 clustering of first alert dates for Users 003, 012, and 020 raised an obvious question: was this early detection, or was the model doing baseline risk stratification? The post-baseline validation answered it directly. In days 31–60, the first full month of valid post-baseline data, deteriorating users already scored a mean of 0.230 against 0.002 for stable users ($p = 0.029$). The model was recognising, from the earliest weeks of meaningful monitoring, that these users’ behavioural profiles, including their chronic circadian amplitude, their sleep hours, and their baseline activity levels, already matched the deterioration pattern. Whether their decline predated the observation period is unknowable from this dataset, but the signal was real and the outcomes confirmed it.

This distinction between baseline risk stratification and dynamic change detection has important practical implications. Task G, as implemented, is most accurately characterised as a *chronic risk stratification* model: it identifies which users,

based on their ongoing behavioural profile, belong to the pattern class associated with confirmed long-term deterioration. For Users 003, 012, and 020, this classification was detectable from their first post-baseline weeks. For User_011, whose deterioration was acute and late-onset, the model correctly produced low scores until the change occurred. The single model thus captured both detection modes, a strength in that it generalises across deterioration phenotypes rather than being calibrated only to one.

The label length imbalance that produced near-zero recall for users 012 and 020 is a genuine limitation rather than an artefact. User_003 contributed 129 positive training days in the LOUO fold; User_012 contributed only 68. Under leave-one-user-out, the model is calibrated by the majority of positive examples it has seen, like those from longer deterioration windows, and produces probabilities below the operating threshold for the shorter-window users even when the directional discrimination is correct (PR-AUC = 0.698 and 0.718, respectively). A lower operating threshold of 0.2 is used in the proof-of-concept application to partially address this, but prospective calibration on a larger cohort with more diverse positive window lengths is the correct long-term solution.

The threshold analysis conducted on Task G LOUO out-of-fold predictions confirmed this diagnosis empirically. No threshold in the range 0.05 to 0.60 produced recall above 0.50 across the cohort, and per-user recall for users 012 and 020 remained near zero at all thresholds. The positive and negative class probability distributions overlap substantially under LOUO for these users, indicating that the performance ceiling is set by the generalisation constraint, the model trained predominantly on User_003's 129-day deterioration window does not learn representations that transfer reliably to shorter deterioration patterns, rather than by threshold placement. This means the limitation cannot be addressed by adjusting the deployment threshold and can only be resolved by increasing the diversity and quantity of positive

training users. PR-AUC is reported as the primary evaluation metric throughout because threshold-based recall metrics for Task G reflect this calibration constraint rather than true discriminative ability.

The SHAP profiles for Task C and Task G are complementary in a clinically meaningful way. Task C is dominated by 14-day activity slopes; the direction and rate of recent change in daytime and mean activity are the primary signals of an imminent critical event. Task G is dominated by 14-day mean levels of circadian amplitude and sleep duration; the chronic level of the day-night activity contrast and the amount of sleep are the primary signals of an active deterioration state. These two feature sets answer different questions: "Is this person declining recently?" versus "Does this person's behavioural profile resemble that of someone in a deterioration phase?" Their complementarity confirms that the two models target genuinely distinct aspects of deterioration.

The consistency of SHAP rankings across all five deteriorating users, evaluated independently under leave-one-user-out, strengthens confidence that the models have learned generalisable patterns rather than user-specific characteristics. `daytime_activity_14d_slope` ranked first for four of five users under Task C; `circadian_amplitude_14d_mean` appeared in the top five for four of five users under Task G. The clinical interpretation is direct: a declining daytime activity trend over two weeks is the strongest near-term predictor of an imminent critical event, while sustained weakening of the circadian rhythm is the strongest chronic indicator of a deterioration state. Both findings align with established gerontological knowledge [5], [24].

5.4 Comparison with Existing Literature

The performance levels achieved in this study are consistent with, and in some respects exceed, those reported in comparable wearable-based deterioration prediction literature. Shao et al. [15] reported AUC values of up to 0.82 for XGBoost applied

to multimodal clinical data for six-month fall risk prediction in nursing home residents. Baig et al. [16] achieved $AUC = 0.78$ for three-month fall risk prediction using retrospective time-series data in long-term care facilities. The $PR-AUC = 0.975$ achieved here for Task C is not directly comparable because of different metrics, different prediction targets, and different data, but it confirms that wrist-worn activity data from a commercial care-monitoring device contains sufficient signal for high-confidence prediction of confirmed deterioration events when appropriate normalisation and label design are applied.

The label design contribution is, to the knowledge of the author, underexplored in the published literature. Most wearable-based prediction studies adopt fixed forward-window labels without examining the effect of window design on temporal detection behaviour. The empirical demonstration in this study that label design determines lead time, not feature engineering or modelling approach, provides a transferable methodological insight that extends beyond the specific dataset and clinical context used here.

5.5 Industry Use

The proof-of-concept pipeline and dashboard developed in this study address a concrete operational need for Oiva Health. The Vivago system, as currently deployed, generates well-being alerts that are binary and retrospective: a flag fires when a threshold has been crossed, without a probability estimate, without feature attribution, and without advance warning beyond the flag itself.

What the Vivago system currently provides is a binary flag: either a threshold was crossed, or it was not. What the pipeline developed here provides is different in kind. The Task C model outputs a daily probability that a confirmed deterioration event will occur within 14 days, updated as new data arrive and actionable well before the flag fires. Fourteen days is enough time to schedule a home visit, review

medications, or escalate monitoring. Task G adds a parallel dimension: not when something will happen, but whether this user's current behavioural profile looks like someone who is deteriorating. A consistently elevated Task G score over several weeks is a different kind of signal, slower, more chronic, and more relevant to resource planning than to emergency response. And the SHAP layer converts both scores from opaque numbers into named behavioural features: a care coordinator can see exactly which aspects of a user's activity and sleep pattern are driving the risk assessment and can make a clinical judgement about whether that explanation makes sense. For example, "Daytime activity has been declining over the past 14 days; circadian amplitude is below this user's personal baseline." This specificity enables a care professional to assess the clinical plausibility of the alert, communicate the reason to colleagues, and document a rationale for any resulting care decision. This is fundamentally different from an opaque algorithmic score and directly addresses the clinical acceptance barrier identified by O'Connor et al. [11] as a primary constraint on AI adoption in care settings.

The proof-of-concept dashboard demonstrates that these capabilities can be packaged in an accessible interface requiring no specialist data science expertise to operate. The Streamlit application loads pre-trained model artefacts and operates entirely on daily aggregated features derived from the same data streams the Vivago system already collects, requiring no additional sensing infrastructure.

Oiva Health's digital home care platform can provide a standardised framework for health data interoperability. The pipeline developed in this study operates on daily aggregated features that can be derived from any continuous activity monitoring data stream; the specific Vivago data format is not a prerequisite for deployment. Adaptation to a FHIR-native data pipeline would require mapping of the Vivago activity and state fields to appropriate FHIR Observation resources and implementation of the daily aggregation and feature engineering steps within a compatible

processing layer. The pre-trained XGBoost models themselves require only the 49-feature input vector, which could be constructed from any compatible wearable data source.

The negative control result, zero false alerts for three confirmed stable users across seven months, provides preliminary evidence that the Task C model does not generate the kind of persistent false-positive background noise that renders clinical AI systems operationally unviable. A deployed system would serve a much larger and more diverse population, and false-positive behaviour in users with complex comorbidity profiles, irregular device wear, or atypical activity patterns cannot be predicted from this cohort.

Clinical governance requirements for a system of this type in a Finnish care setting would include prospective validation on a representative population sample, documentation of model performance characteristics stratified by relevant clinical subgroups, a defined process for handling conflicting model outputs and clinical judgement, and ongoing monitoring for model drift as the user population or care context evolves. The pipeline described in this thesis should be understood as a foundation for this process, not as a production-ready system.

5.6 Limitations

Cohort size: Twenty users is not a large cohort, and five confirmed deteriorating users is even less. For Task G, only four were evaluable, meaning the mean PR-AUC of 0.829 rests on four independent test folds. The standard deviation (SD) of 0.140 is not model instability; it is what four users with different deterioration trajectories and different positive window lengths produce when evaluated independently. The confidence intervals are wide, and the findings cannot reasonably be claimed to generalise beyond this cohort without further validation.

The label length imbalance compounds this. Positive windows ranged from 35

to 129 days across Task G users, and the near-zero recall for Users 012 and 020 follows directly from this asymmetry. The operational threshold of 0.2 for Task G was not cross-validated in the same way as the Task C threshold of 0.4, and its sensitivity-specificity trade-off in a broader population remains untested.

Observation period and temporal scope: Seven months is a short window for elderly care monitoring. Seasonal variation in activity, changes in care staffing, and health trajectories that unfold over years are all absent from this dataset. Whether a model trained on a single autumn-to-spring observation would transfer to other time periods or longer-term deterioration patterns is an open question.

Manual change-point override: User_003’s change-point required a manual override because the PELT algorithm does not reliably detect gradual linear declines; it identifies structural breaks. The override date was chosen based on visual inspection of the smoothed activity trajectory and applied consistently, but it remains a human judgement. User_014’s exclusion from Task G entirely, due to the absence of a valid automated change-point, further reduced the positive training pool and may have affected model calibration for users with complex alarm histories.

Single-device context: All data came from one device over seven months. Vivago activity counts are not directly comparable to outputs from other accelerometer brands, and the validated cut points used here are manufacturer-specific. Whether the deterioration signatures identified in this cohort would appear in the same form in a different care model or country is unknown.

The pipeline also relies entirely on activity and sleep features. Physiological signals like heart rate variability, blood oxygen saturation, and skin temperature were not available and are known to provide complementary and sometimes earlier warning of deterioration [51]. What was built here captures one dimension of a

multimodal picture.

5.7 Future Work

Larger and more diverse cohorts: Expanding to 200 or more users across multiple care sites is the most immediate priority. A larger cohort would narrow the confidence intervals around LOUO estimates, enable prospective threshold calibration for Task G, and allow subgroup analysis by cognitive status, care setting, or comorbidity profile that a 20-user study cannot support.

Multimodal sensing: Heart rate variability, skin temperature, and blood oxygen saturation are plausible additions that photoplethysmography-capable wrist devices could provide without requiring users to wear additional hardware. Multimodal fusion with physiological indicators has been shown to improve risk prediction accuracy in comparable clinical contexts [51], [52].

Prospective clinical validation: Every result in this thesis is retrospective. A prospective pilot study, deploying the dashboard in an operational setting and measuring how care staff respond to its alerts, is the necessary next step. Key outcomes would include the false-positive rate under real conditions, the rate at which model alerts led to preventive interventions, and whether care staff find the SHAP explanations credible and actionable.

FHIR integration and scalability: Connecting the pipeline to Oiva Health’s FHIR-compliant infrastructure would standardise data ingestion and allow risk scores to be produced as FHIR Observation resources integrated directly into care workflow interfaces. It would also enable multisite deployment without site-specific data transformation. Concept drift monitoring and model update schedules would need to be addressed as part of this process.

Personalised modelling: The current pipeline applies a single global model universally. User-specific fine-tuning is a natural extension, particularly for users whose activity profile falls outside the range of the training cohort. Continual learning, where the model updates incrementally as new confirmed events accumulate in the deployed population, represents a longer-term direction worth investigating.

6 Conclusion

This thesis investigated whether machine learning applied to longitudinal wrist-worn activity sensor data can support proactive well-being risk detection in elderly care, addressing a gap that commercial monitoring systems leave unresolved: the identification of deterioration risk before formal system thresholds are crossed and reactive alerts are generated. Working in collaboration with Oiva Health and using pseudonymised real-world data from 20 users monitored continuously for approximately seven months via the Vivago wrist-worn activity device, a complete proof-of-concept machine learning pipeline was developed, evaluated, and implemented as a deployable dashboard addressing two complementary clinical questions.

The first research question asked whether a model can detect whether a user is currently in a behavioural deterioration state and how early. RQ1 was answered affirmatively: the Task G state classification model, trained on change-point-derived labels that mark the onset of a confirmed behavioural deterioration state, achieved a mean PR-AUC of 0.829 under leave-one-user-out cross-validation and identified deteriorating users a mean of 56.0 days before formal yellow well-being flags, with statistically significant separation from confirmed stable users in the first post-baseline monitoring month (Mann-Whitney $p = 0.029$). SHAP analysis identified 14-day mean levels of circadian amplitude and sleep duration as the dominant predictive features, consistent with established gerontological knowledge about circadian rhythm disruption as an early indicator of functional decline.

The second research question asked whether a model can predict the occurrence of a critical adverse event within a defined horizon. RQ2 was also answered affirmatively: the Task C event prediction model, which was targeting red well-being flags within 14 days, achieved a PR-AUC of 0.975 at the operating threshold, providing exactly 14 days of advance warning before confirmed red flag events for all five deteriorating users with zero false alerts for stable users across the full observation period. Here, SHAP analysis revealed a complementary feature profile: 14-day downward slopes in daytime and mean activity dominated, reflecting a recent trend direction rather than chronic level.

The most transferable finding from this work is not a performance number. It is a methodological observation: label design determines how early a model detects. Event-window labels suppress early signal by construction because days before the window are negative by definition, even when the user is already deteriorating. State-based labels do not have this constraint. The 56-day versus 0-day lead time difference between Task G and Task C is not a difference in models or features. It is a difference in what was asked of the data. This insight is applicable to any domain where the distinction between detecting a current state and predicting an upcoming event matters clinically.

The pipeline runs on real monitoring data from a real care setting, subject to real data quality issues like irregular device wear, sparse clinical labels, and a heterogeneous cohort that does not behave like a textbook dataset. The Streamlit dashboard shows that deploying both models simultaneously is achievable in a practical interface: a care coordinator can see at a glance which users are at chronic risk and which face imminent adverse events, with SHAP explanations that translate the model's output into the specific behavioural changes driving each score. That is what clinical AI adoption actually requires; not just accuracy but interpretability.

For Oiva Health and digital home care providers more broadly, the findings sug-

gest that the wrist-worn activity data already collected through routine monitoring carries substantially more predictive content than current threshold-based alert systems exploit. A machine learning layer operating on the same data streams, without additional sensing infrastructure, can provide quantified risk probabilities, personalised behavioural context, and advanced warning that extends from days to months depending on the clinical question being asked. The natural next step is a prospective pilot study in which care staff interact with the system in an operational setting, generating the evidence base required for clinical validation and regulatory consideration.

The limitations of this work are real and should not be understated. A cohort of 20 users from a single care site over seven months cannot establish generalised clinical validity; the statistical power of leave-one-user-out evaluation is inherently constrained at this scale; and the absence of prospective validation means that deployed false-positive behaviour in more heterogeneous populations remains uncharacterised. They define the boundaries within which the results should be interpreted and the conditions under which further development is warranted.

As elderly populations grow and the demand on home care and residential care services intensifies, the capacity to identify who is deteriorating, how severely, and how soon becomes an increasingly critical component of effective care delivery. This thesis demonstrates that the data to support such identification already exists within deployed commercial monitoring infrastructure and that the analytical methods to extract actionable risk signals from that data are available and feasible. Realising the clinical and operational value of these capabilities requires continued collaboration between health technology research, care system providers, and the care professionals whose judgement any algorithmic system must ultimately support and augment rather than replace.

References

- [1] N. Salari, N. Darvishi, M. Ahmadipanah, S. Shohaimi, and M. Mohammadi, “Global prevalence of falls in the older adults: A comprehensive systematic review and meta-analysis”, *Journal of orthopaedic surgery and research*, vol. 17, no. 1, p. 334, 2022.
- [2] World Health Organization, *Falls*, <https://www.who.int/news-room/fact-sheets/detail/falls>, Accessed: 2026-02-10, 2021.
- [3] S. Giovannini et al., “Falls among older adults: Screening, identification, rehabilitation, and management”, *Applied Sciences*, vol. 12, no. 15, p. 7934, 2022.
- [4] R. Vaishya and A. Vaish, “Falls in older adults are serious”, *Indian journal of orthopaedics*, vol. 54, no. 1, pp. 69–74, 2020.
- [5] Q. Xu, X. Ou, and J. Li, “The risk of falls among the aging population: A systematic review and meta-analysis”, *Frontiers in public health*, vol. 10, p. 902599, 2022.
- [6] S. Saunders et al., “Risk factors for falls in community-dwelling older adults: An umbrella review”, *medRxiv*, pp. 2025–04, 2025.
- [7] M. Chen et al., “A systematic review of wearable sensor-based technologies for fall risk assessment in older adults”, *Sensors*, vol. 22, no. 18, p. 6752, 2022.

-
- [8] G. L. Bate et al., “The role of wearable sensors to monitor physical activity and sleep patterns in older adult inpatients: A structured review”, *Sensors*, vol. 23, no. 10, p. 4881, 2023.
 - [9] S. K. Gharghan and H. A. Hashim, “A comprehensive review of elderly fall detection using wireless communication and artificial intelligence techniques”, *Measurement*, vol. 226, p. 114 186, 2024.
 - [10] Y. Li et al., “A decade of progress in wearable sensors for fall detection (2015–2024): A network-based visualization review”, *Sensors*, vol. 25, no. 7, p. 2205, 2025.
 - [11] S. O’Connor et al., “Artificial intelligence for falls management in older adult care”, *Journal of Nursing Management*, vol. 31, no. 5, pp. 1234–1245, 2022.
 - [12] MDPI Sensors Authors, “Machine learning applied to edge computing and wearable devices for healthcare: Systematic mapping of the literature”, *Sensors*, vol. 24, no. 19, p. 6322, 2024.
 - [13] R. Rajagopalan, I. Litvan, and T.-P. Jung, “Fall prediction and prevention systems: Recent trends, challenges, and future research directions”, *Sensors*, vol. 17, no. 11, p. 2509, 2017.
 - [14] A. Capodici et al., “A scoping review of machine learning models to predict risk of falls in elders, without using sensor data”, *Diagnostic and Prognostic Research*, vol. 9, no. 1, p. 11, 2025.
 - [15] L. Shao et al., “Development and external validation of a machine learning–based fall prediction model for nursing home residents: A prospective cohort study”, *Journal of the American Medical Directors Association*, vol. 25, no. 9, p. 105 169, 2024.
 - [16] M. M. Baig et al., “Predicting falls in long-term care facilities: Machine learning study”, *JMIR Aging*, vol. 5, no. 2, e35373, 2022.

-
- [17] Y. Chen et al., “Global, regional, and national burden of falls among older adults: Findings from the global burden of disease study 2021 and projections to 2040”, *npj Aging*, vol. 11, no. 1, p. 85, 2025.
- [18] Y. Liu, J. Qiu, R. Huang, Y. Zhou, and Z. Huang, “Global, regional, and national burden of falls from 1990 to 2021: A comprehensive analysis with a focus on falls attributable to reduced bone mass and projections to 2045”, *Frontiers in Public Health*, vol. 13, p. 1695026, 2025.
- [19] S. Z. Fadem, “Consequences of falls”, in *Understanding and Preventing Falls: A Guide to Reducing Your Risks*, Springer, 2023, pp. 61–69.
- [20] D. A. Jehu et al., “Risk factors for recurrent falls in older adults: A systematic review with meta-analysis”, *Maturitas*, vol. 144, pp. 23–28, 2021.
- [21] A. Kristoffersson and M. Lindén, “A systematic review of wearable sensors for monitoring physical activity”, *Sensors*, vol. 22, no. 2, p. 573, 2022.
- [22] V. Birrer, M. Elgendi, O. Lambercy, and C. Menon, “Evaluating reliability in wearable devices for sleep staging”, *NPJ Digital Medicine*, vol. 7, no. 1, p. 74, 2024.
- [23] N. Deguchi et al., “In-home validation of wrist-and waist-worn devices against portable electroencephalography for sleep assessment in older adults”, *medRxiv*, pp. 2025–10, 2025.
- [24] A. Sasaki-Otomaru, K. Saito, K. Yamasue, O. Tochikubo, and Y. Kanoya, “Relationship of sleep and activity, assessed via a wristwatch-type pulsometer with an accelerometer, with health status in community-dwelling older adults: A preliminary study”, *Plos one*, vol. 20, no. 3, e0317524, 2025.
- [25] P. Gorce and J. Jacquier-Bret, “Fall detection in elderly people: A systematic review of ambient assisted living and smart home-related technology performance”, *Sensors*, vol. 25, no. 21, p. 6540, 2025.

-
- [26] K. Abdul Jabbar, R. Mc Ardle, S. Lord, N. Kerse, S. Del Din, and R. Teh, “Physical activity in community-dwelling older adults: Which real-world accelerometry measures are robust? a systematic review”, *Sensors*, vol. 23, no. 17, p. 7615, 2023.
- [27] M. Giurgiu et al., “Assessment of 24-hour physical behaviour in adults via wearables: A systematic review of validation studies under laboratory conditions”, *International Journal of Behavioral Nutrition and Physical Activity*, vol. 20, no. 1, p. 68, 2023.
- [28] V. T. Van Hees et al., “Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: An evaluation on four continents”, *Journal of applied physiology*, vol. 117, no. 7, pp. 738–744, 2014.
- [29] J. Vanhelst et al., “Validation of the vivago wrist-worn accelerometer in the assessment of physical activity”, *BMC public health*, vol. 12, no. 1, p. 690, 2012.
- [30] X. Wang, J. Ellul, and G. Azzopardi, “Elderly fall detection systems: A literature survey”, *Frontiers in Robotics and AI*, vol. 7, p. 71, 2020.
- [31] R. O. Zafar and F. Zafar, “Real-time activity and fall detection using transformer-based deep learning models for elderly care applications”, *BMJ Health & Care Informatics*, vol. 32, no. 1, e101439, 2025.
- [32] S. Li, “Fall detection with wrist-worn watch by observations in statistics of acceleration”, *IEEE Access*, vol. 11, pp. 19 567–19 578, 2023.
- [33] A. Lee, H. Lee, and S.-H. Lee, “Digital healthcare approaches for fall detection and prediction in older adults: A systematic review of evidence from hospital and long-term care settings”, *Medicina*, vol. 61, no. 11, p. 1926, 2025.

- [34] T. B. Aderinola et al., “Watch your step: A cost-sensitive framework for accelerometer-based fall detection in real-world streaming scenarios”, *arXiv preprint arXiv:2509.11789*, 2025.
- [35] M. C. Olson and T. E. Lockhart, “Predicting fall risk through automatic wearable monitoring: A systematic review”, *International journal of prognostics and health management*, vol. 12, no. 4, 2021.
- [36] S. Jahandideh et al., “Using machine learning models to predict falls in hospitalised adults”, *International Journal of Medical Informatics*, vol. 187, p. 105436, 2024.
- [37] V. Roshdibenam, G. J. Jogerst, N. R. Butler, and S. Baek, “Machine learning prediction of fall risk in older adults using timed up and go test kinematics”, *Sensors*, vol. 21, no. 10, p. 3481, 2021.
- [38] F.-C. Lin, P.-H. Chen, C.-H. Yang, C.-C. Yang, H.-Y. Chuang, and C.-H. Hung, “Ai-driven fall prediction across generations: Integrating deep learning and machine learning for young, middle-aged, and older adults”, *Gerontology*, vol. 72, no. 1, pp. 3–18, 2026.
- [39] Y. Guo et al., “Explainable ai reveals temporal risk pathways in fall prediction: Extracting clinical insights from multi-horizon machine learning models”, *GeroScience*, 2026.
- [40] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions”, *Advances in neural information processing systems*, vol. 30, 2017.
- [41] A. Lakhdari, J. Li, A. Abusafia, and A. Bouguettaya, “Privacy-aware iot fall detection services for aging in place”, in *2025 IEEE International Conference on Web Services (ICWS)*, IEEE, 2025, pp. 324–334.
- [42] E. GDPR, “General data protection regulation (gdpr)”, *Cit. on*, p. 4, 2018.

- [43] Y. Xu et al., “A systematic review of wearable sensor-based technologies for fall risk assessment in older adults”, *Sensors*, vol. 22, no. 18, p. 6752, 2022.
- [44] I. Gani, I. Litchfield, D. Shukla, G. Delanerolle, N. Cockburn, and A. Pathmanathan, “Understanding “alert fatigue” in primary care: Qualitative systematic review of general practitioners attitudes and experiences of clinical alerts, prompts, and reminders”, *Journal of medical Internet research*, vol. 27, e62763, 2025.
- [45] S. Alamgeer, Y. Souissi, and A. Ngu, “Ai-generated fall data: Assessing llms and diffusion model for wearable fall detection”, *Sensors*, vol. 25, no. 16, p. 5144, 2025.
- [46] E. Lamminmäki, A. Saarinen, J. Lötjönen, M. Partinen, and I. Korhonen, “Differences in light sleep and deep sleep measured with ist vivago® wristcare”, in *Ifmbe proceedings*, vol. 11, 2005, p. 1.
- [47] C. Truong, L. Oudre, and N. Vayatis, “Selective review of offline change point detection methods”, *Signal processing*, vol. 167, p. 107 299, 2020.
- [48] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system”, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [49] W. Hong et al., “A comparison of xgboost, random forest, and nomograph for the prediction of disease severity in patients with covid-19 pneumonia: Implications of cytokine and immune cell profile”, *Frontiers in Cellular and Infection Microbiology*, vol. 12, p. 819 267, 2022.
- [50] L. Wang, X. Wang, A. Chen, X. Jin, and H. Che, “Prediction of type 2 diabetes risk and its effect evaluation based on the xgboost model”, in *Healthcare*, MDPI, vol. 8, 2020, p. 247.

-
- [51] A. González Castro et al., “Accurate fall risk prediction in older adults: Integrating sensor and clinical data through machine learning”, *Journal of Geriatric Physical Therapy*, 2026.
- [52] K. Nakajima et al., “Machine learning prediction for postdischarge falls in older adults: A multicenter prospective study”, *Journal of the American Medical Directors Association*, 2024.

Appendix A Supplementary Model Performance Figures

This appendix presents precision-recall curves for all five exploratory prediction tasks (A, B, D, E, and F) across Logistic Regression, Random Forest, and XGBoost, complementing the primary results reported in Chapter 4. These tasks were evaluated to contextualise the selection of Task C as the primary RQ2 model. Confusion matrices are omitted as they reflect in-sample performance only and are not meaningful for assessing generalisability under leave-one-user-out cross-validation.

Task A: Automatic Inactivity Alarm within 7 Days

Task A predicted whether an automatic inactivity alarm would occur within 7 days. All three models showed PR-AUC values only marginally above the no-skill baseline, reflecting the short horizon and the heterogeneous etiology of automatic inactivity alarms. XGBoost achieved $\text{PR-AUC} = 0.085 (\pm 0.036)$, the lowest among the three models for this task, with recall near zero at threshold 0.4 (0.030).

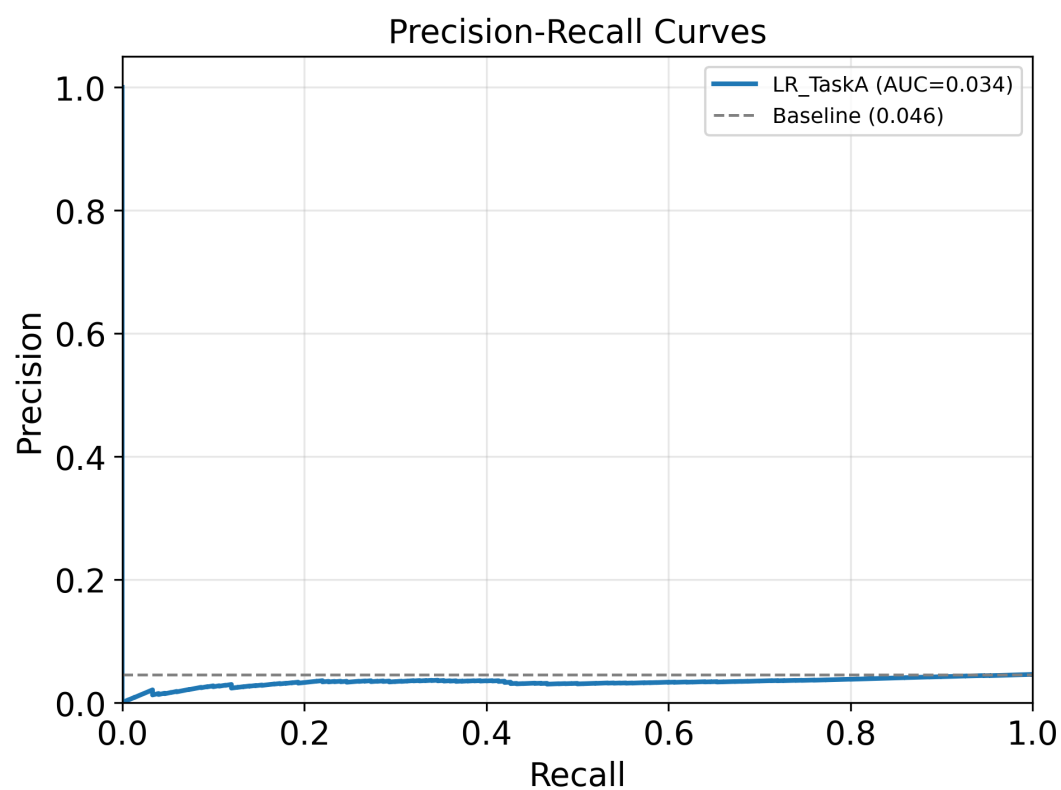


Figure A.1: Precision-recall curve for Logistic Regression on Task A.

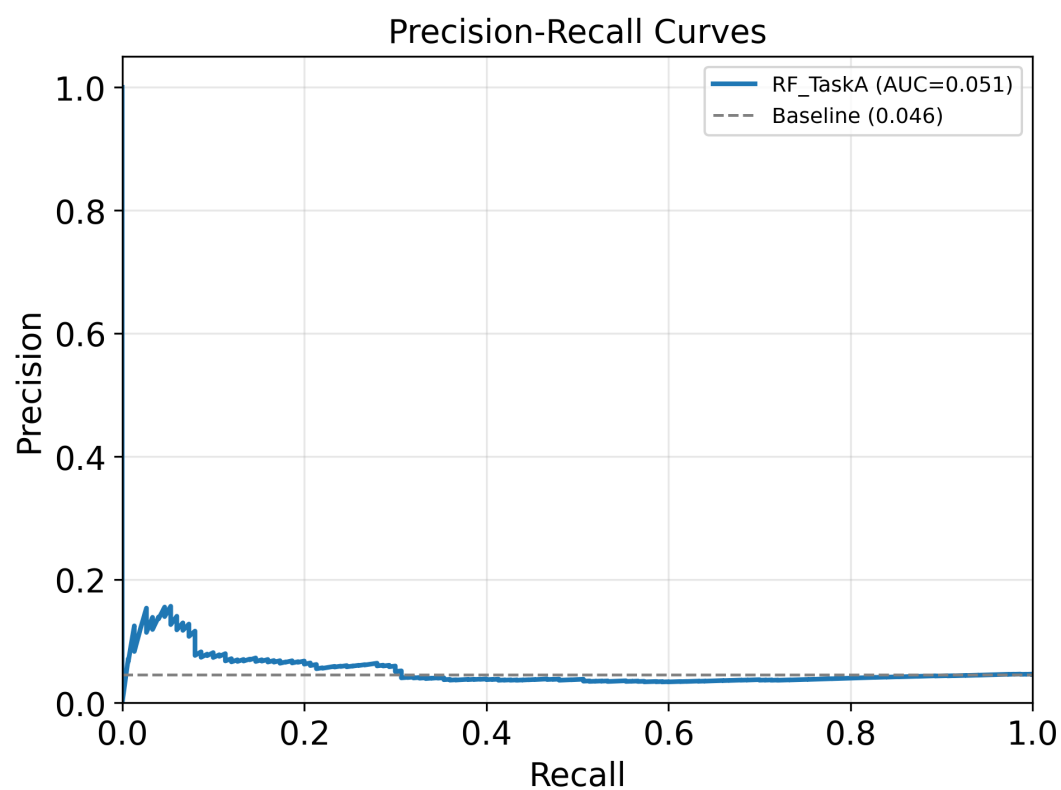


Figure A.2: Precision-recall curve for Random Forest on Task A.

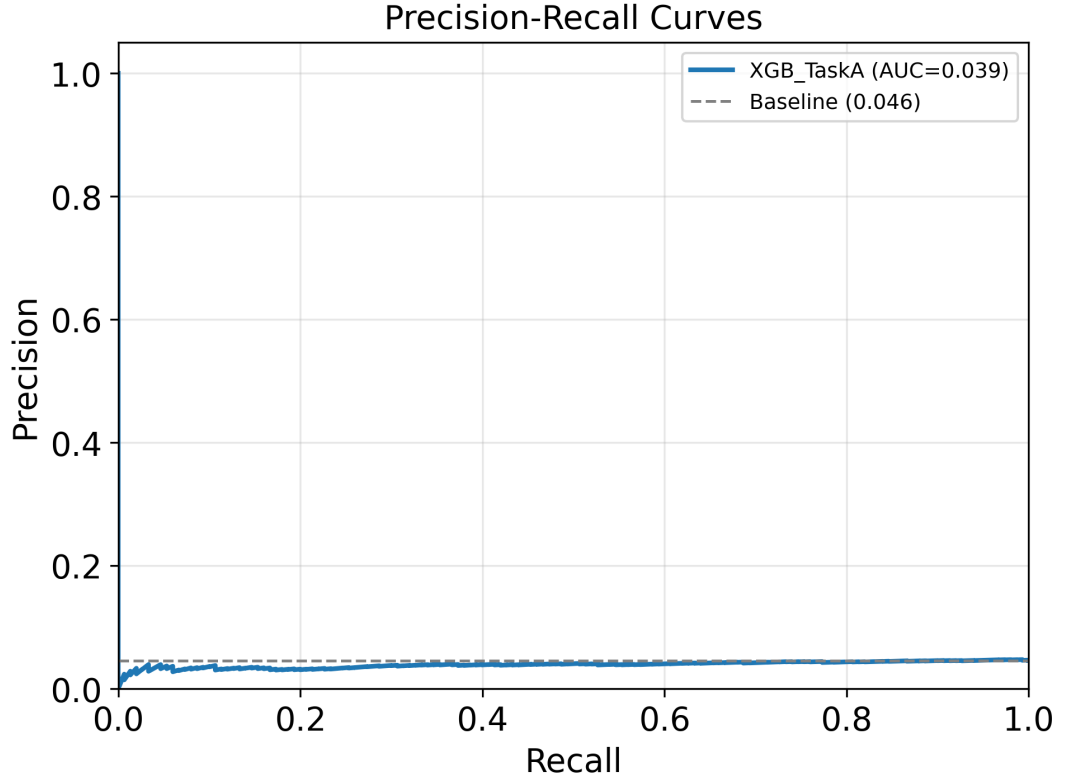


Figure A.3: Precision-recall curve for XGBoost (tuned) on Task A.

Task B: Automatic Inactivity Alarm within 14 Days

Task B extended the inactivity alarm horizon from 7 to 14 days. Performance improved modestly across all three models relative to Task A, with XGBoost achieving $\text{PR-AUC} = 0.208 (\pm 0.106)$, the highest of the three models for this task and approximately 2.6-fold above the no-skill baseline of 0.079. Task B was used as the hyperparameter tuning target for XGBoost due to its larger positive class relative to Task C, providing a more stable optimisation landscape.

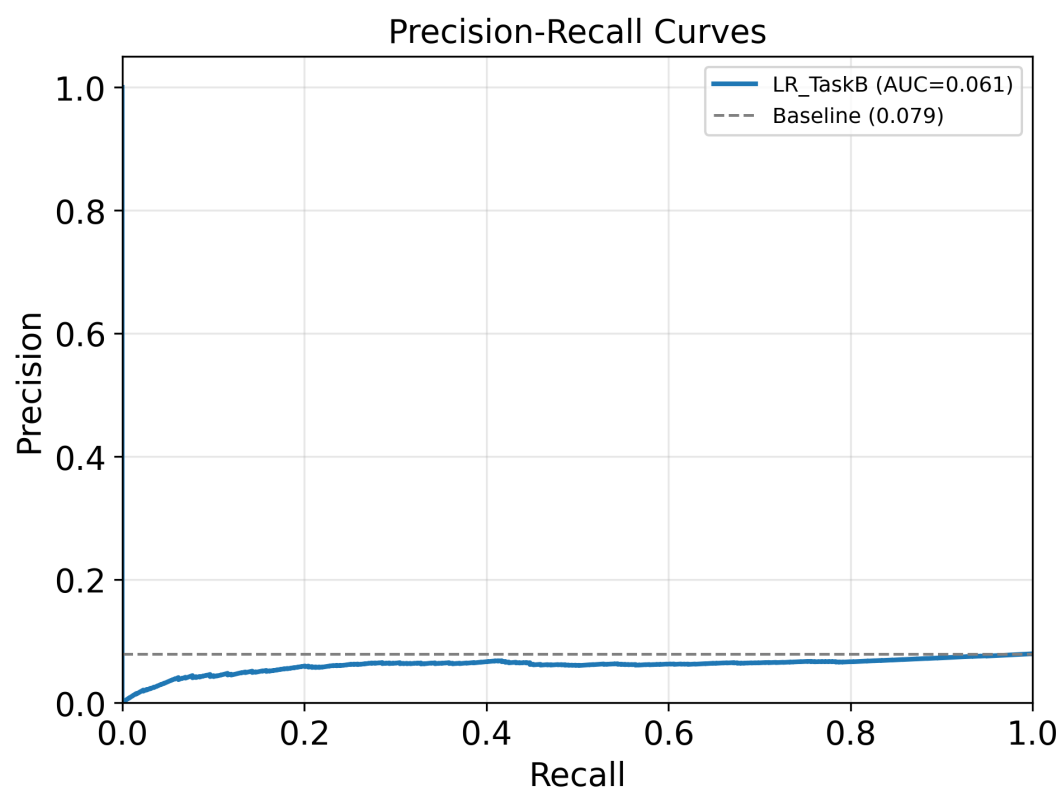


Figure A.4: Precision-recall curve for Logistic Regression on Task B.

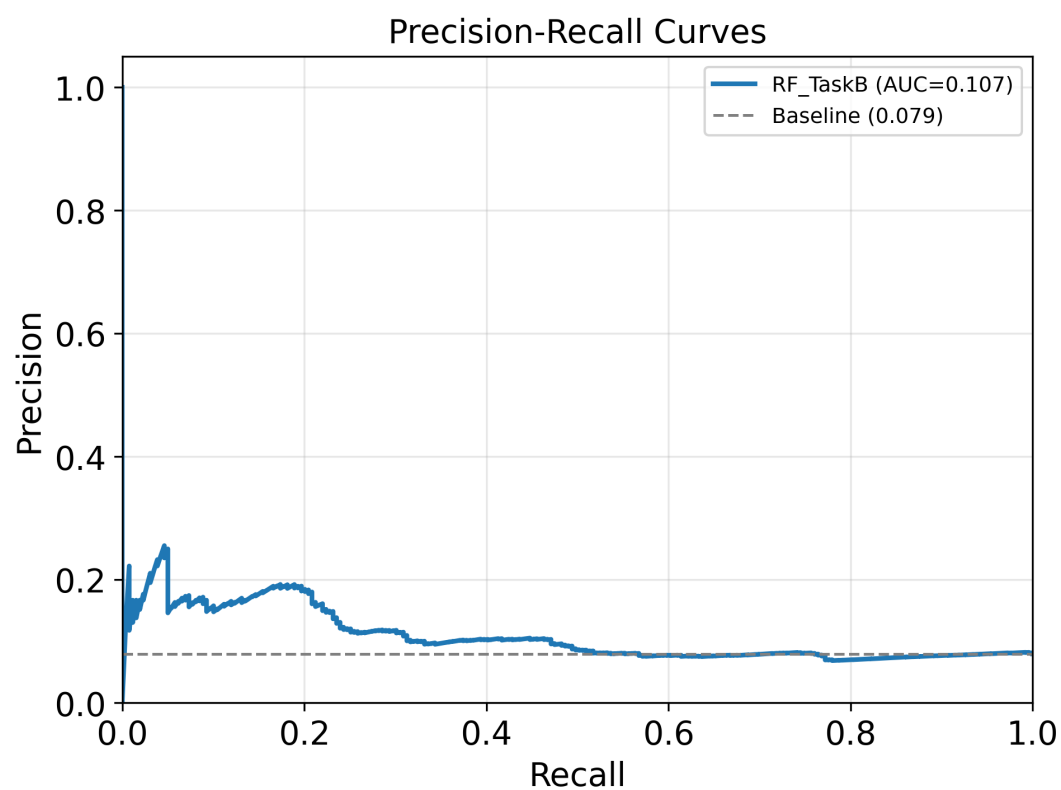


Figure A.5: Precision-recall curve for Random Forest on Task B.

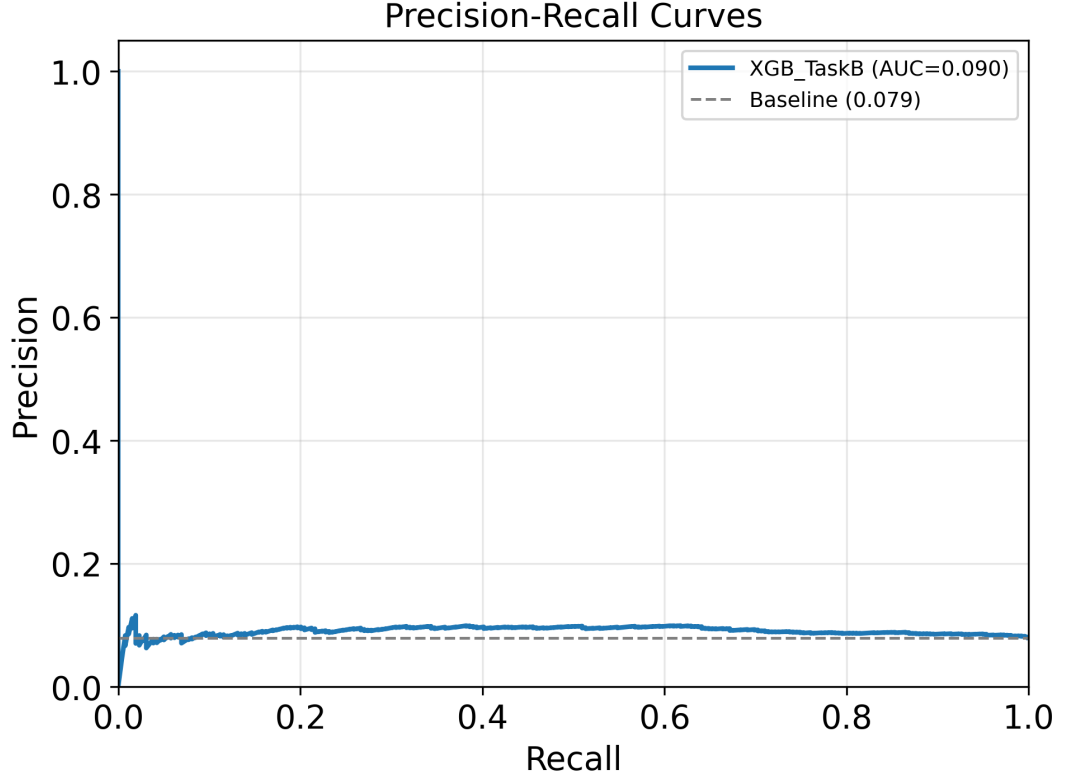


Figure A.6: Precision-recall curve for XGBoost (tuned) on Task B.

Task D: Red Wellbeing Flag within 28 Days

Task D extended the red flag prediction horizon from 14 to 28 days. XGBoost achieved $\text{PR-AUC} = 0.646 (\pm 0.077)$ compared to 0.975 for Task C, indicating that the additional two-week window introduced label noise that reduced model discriminability despite a higher positive rate (5.1% versus 3.4%).

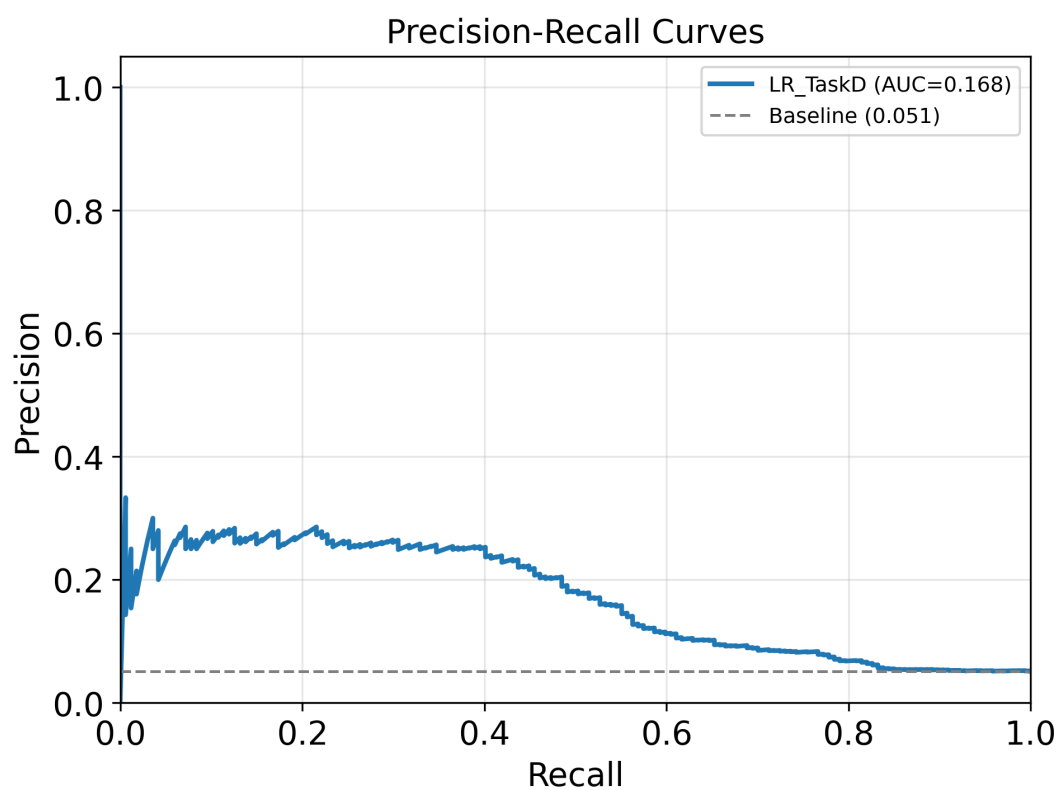


Figure A.7: Precision-recall curve for Logistic Regression on Task D.

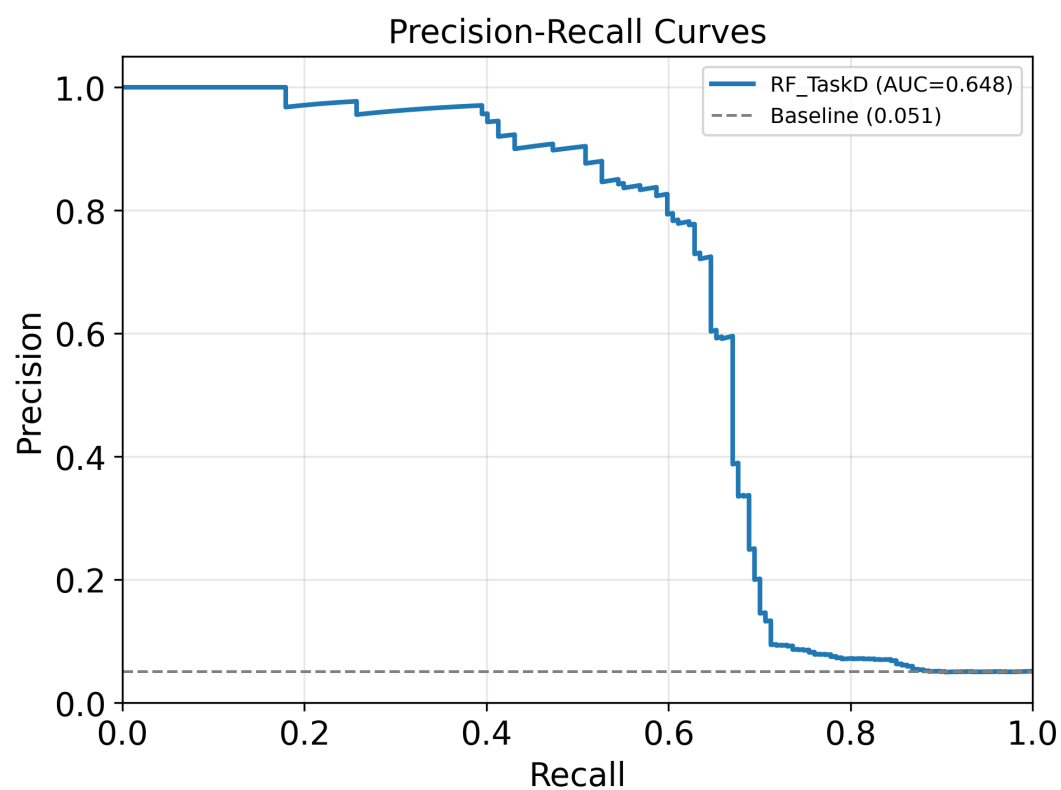


Figure A.8: Precision-recall curve for Random Forest on Task D.

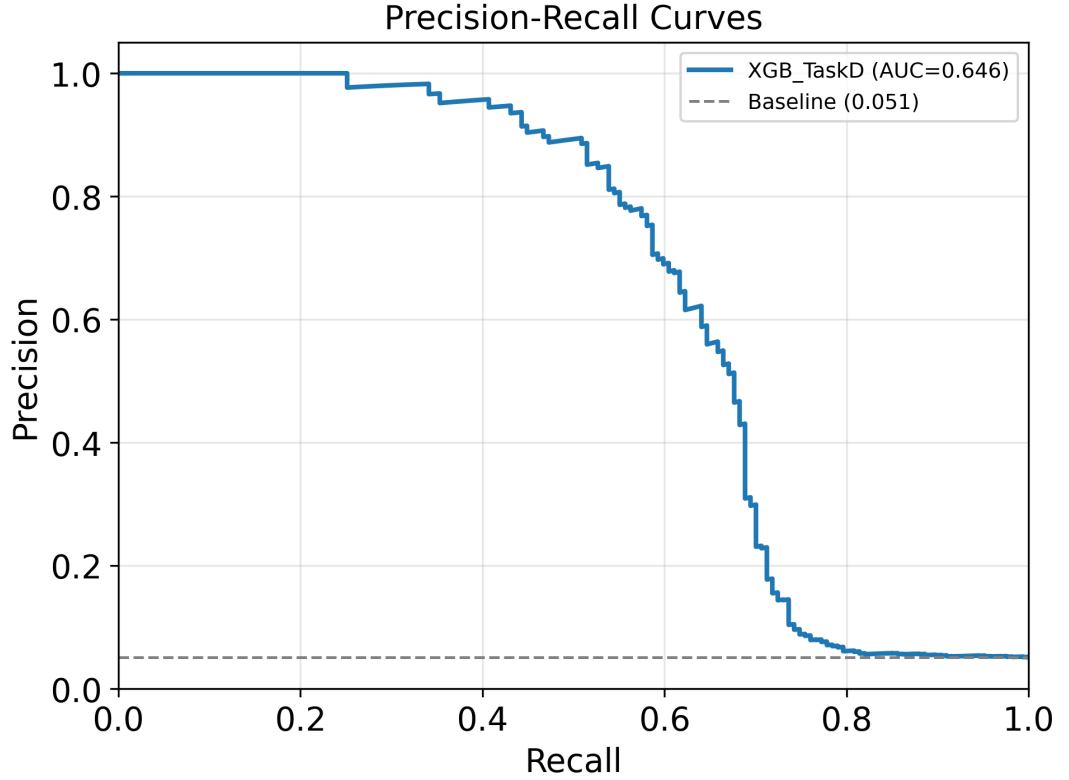


Figure A.9: Precision-recall curve for XGBoost (tuned) on Task D.

Task E: Combined Alarm within 14 Days

Task E incorporated manual alarm button presses alongside automatic inactivity alarms as a proxy for falls and acute distress episodes. XGBoost achieved PR-AUC = 0.181 (± 0.112), reflecting the heterogeneous etiology of combined alarm events and the resulting noisy training signal.

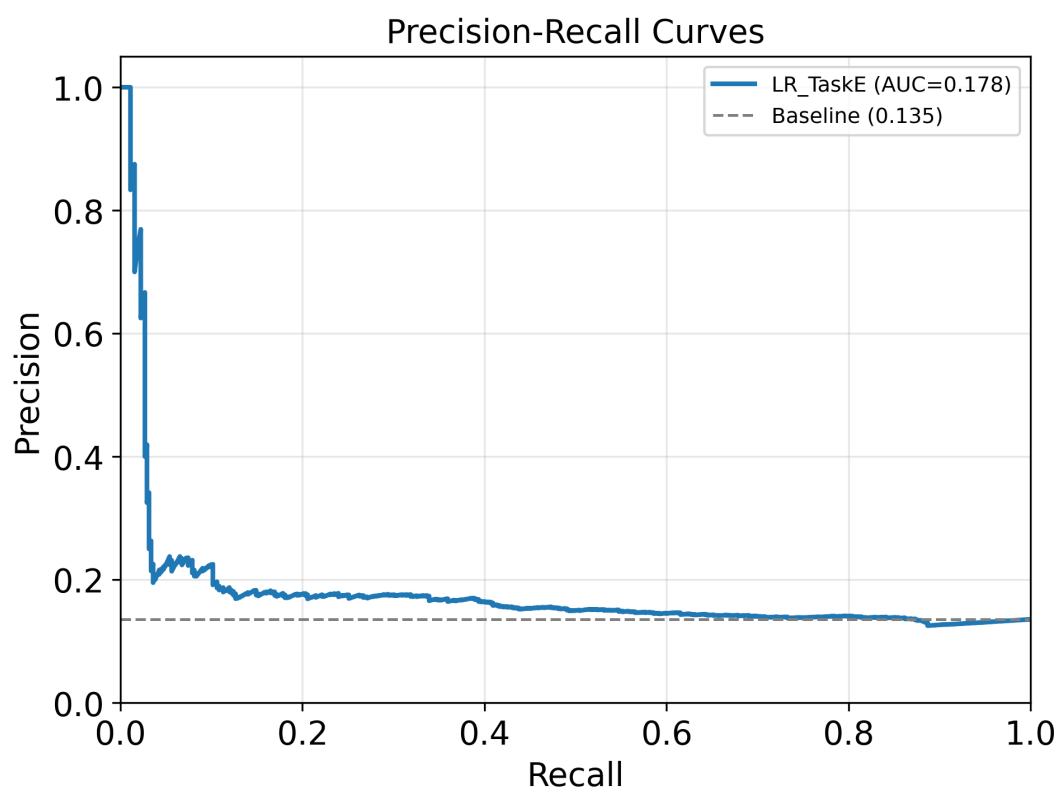


Figure A.10: Precision-recall curve for Logistic Regression on Task E.

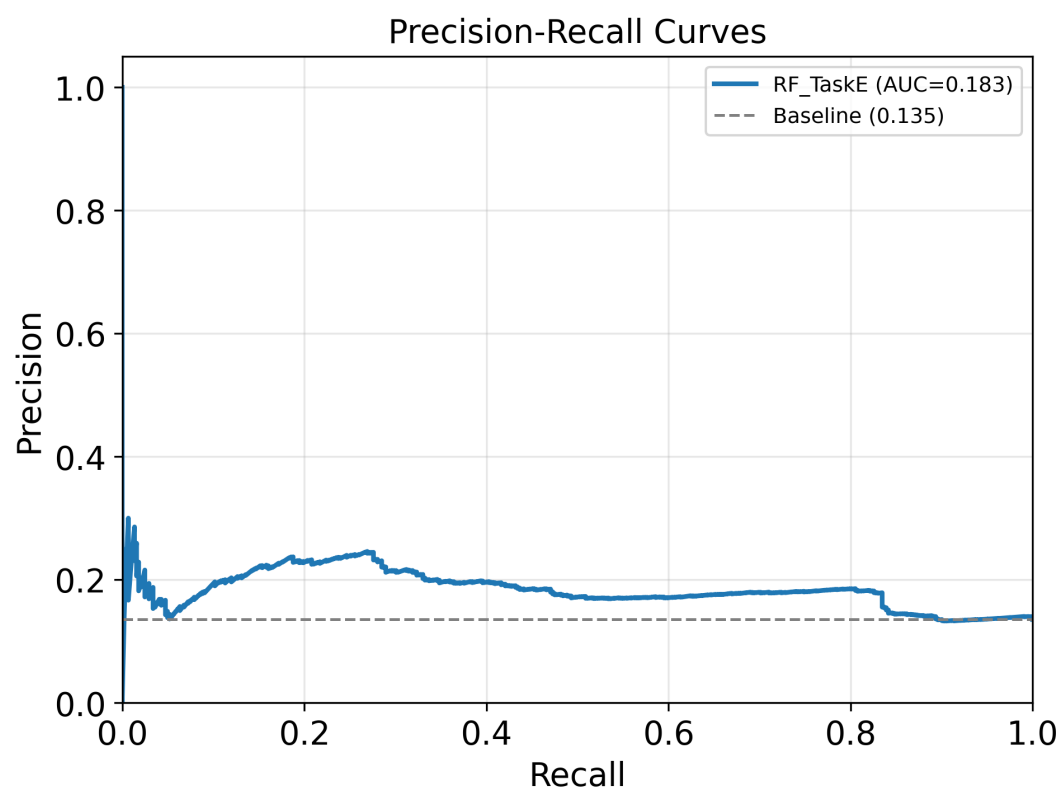


Figure A.11: Precision-recall curve for Random Forest on Task E.

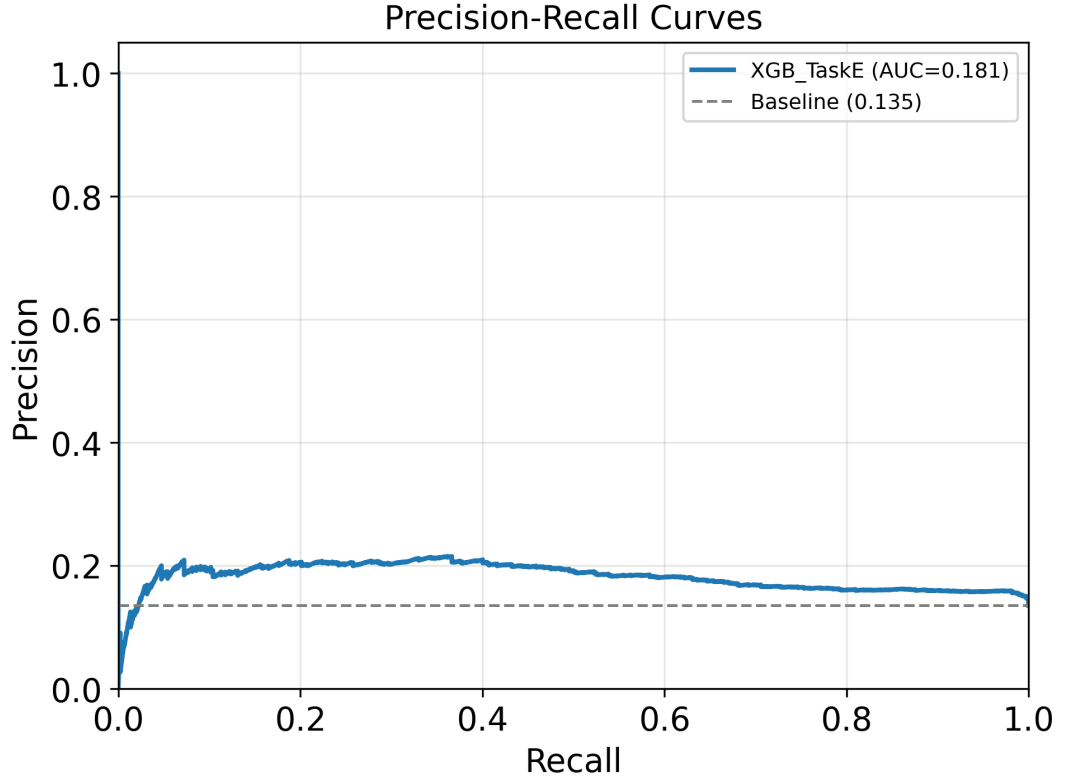


Figure A.12: Precision-recall curve for XGBoost (tuned) on Task E.

Task F: Combined Alarm within 28 Days

Task F extended the combined alarm horizon to 28 days, resulting in a positive rate of 23.3%, the highest across all six event-window tasks. XGBoost achieved PR-AUC = 0.291 (± 0.140). The elevated positive rate reduced the `scale_pos_weight` parameter to approximately 3.3, substantially altering the class-weighting regime relative to tasks C and D.

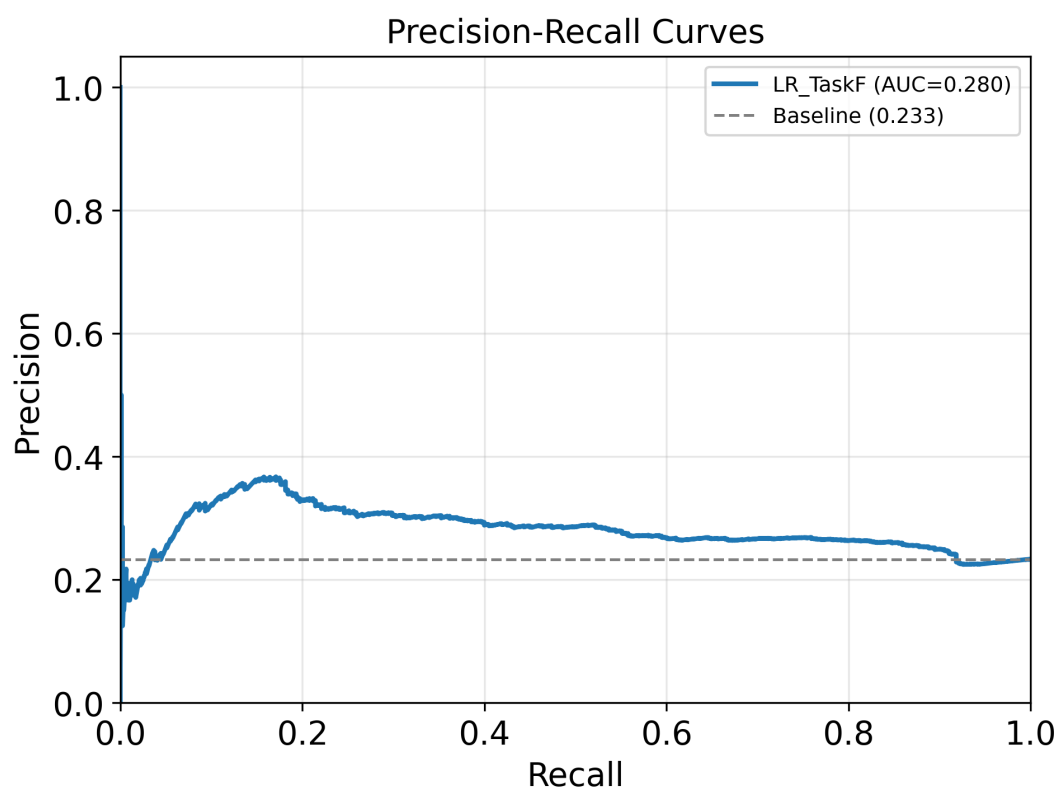


Figure A.13: Precision-recall curve for Logistic Regression on Task F.

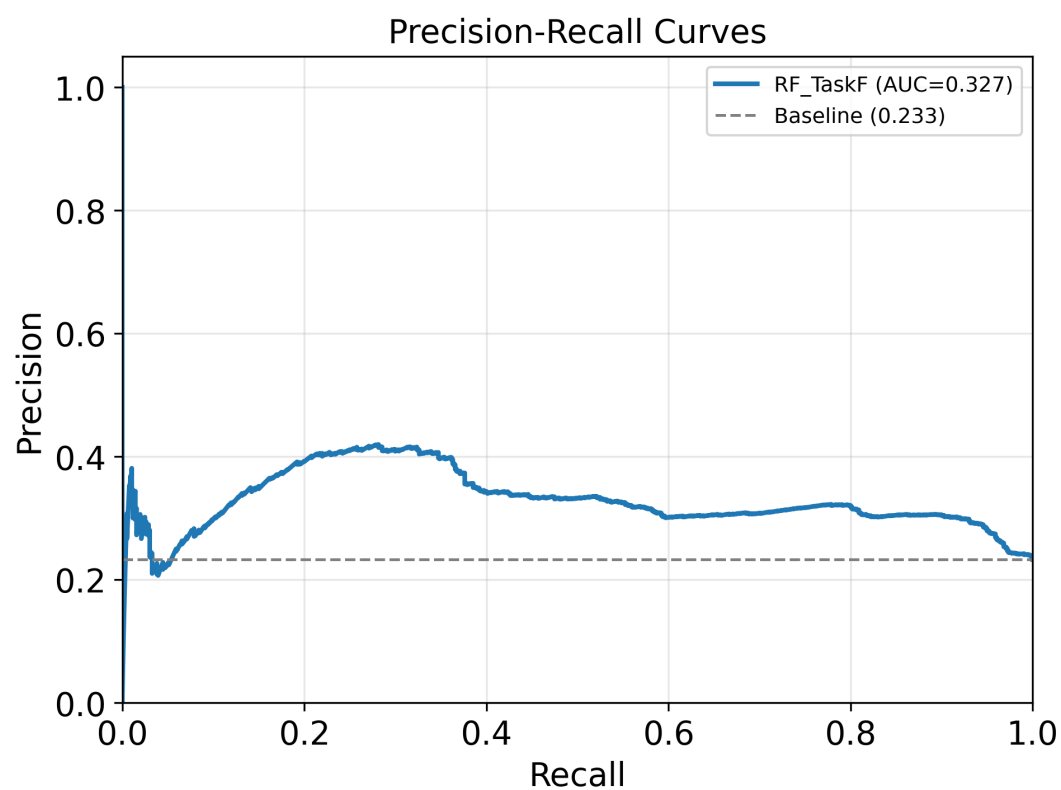


Figure A.14: Precision-recall curve for Random Forest on Task F.

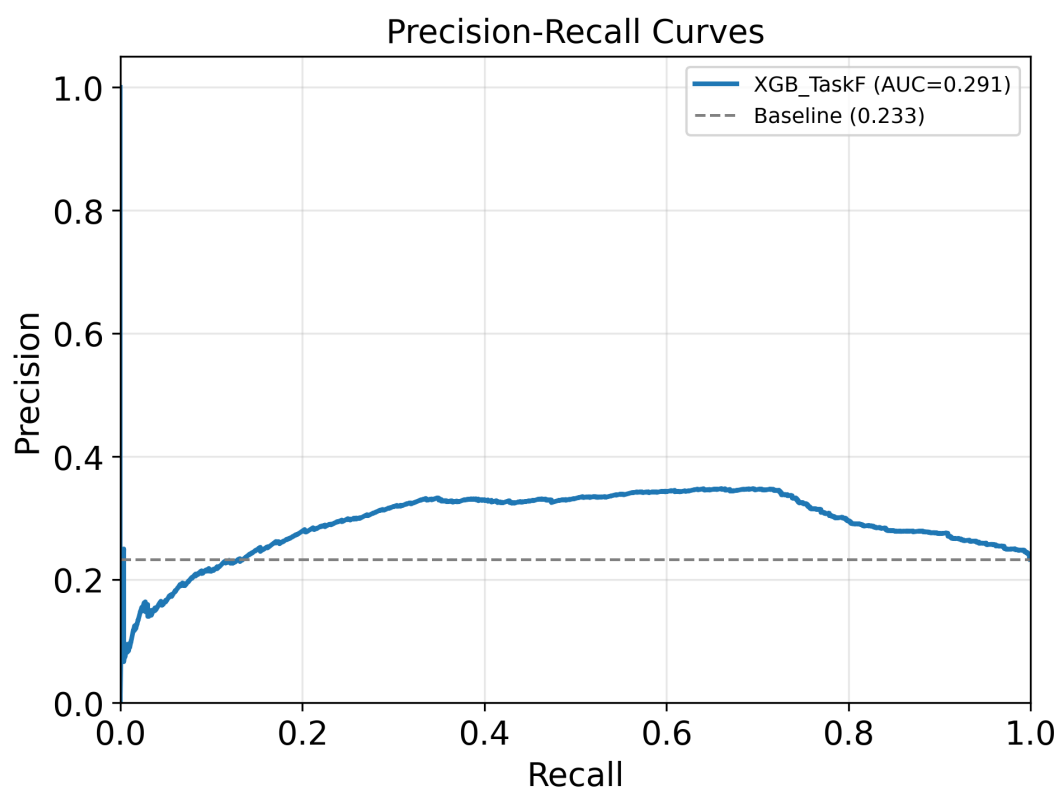


Figure A.15: Precision-recall curve for XGBoost (tuned) on Task F.

Appendix B Full Model Input Feature List

The following table lists all 49 features comprising the model input matrix $\mathbf{X}$ used across all prediction tasks (Tasks A through G). Features were derived from minute-level Vivago activity records through the preprocessing and feature engineering pipeline described in Chapter 3. The four device quality columns excluded from model input (`off_wrist_minutes`, `gap_minutes`, `wear_minutes`, `raw_data_completeness`) are not listed here.

Table B.1: Complete list of 49 model input features.

Feature name	Type	Description
mean_activity	Level	Mean activity count, awake minutes
max_activity	Level	Maximum activity count, any awake minute
active_minutes	Level	Awake minutes with activity > 15 counts
daytime_activity	Level	Mean activity 08:00–19:00, awake state
night_activity	Level	Mean activity 22:00–05:00, all states

Table B.1 continued

Feature name	Type	Description
circadian_amplitude	Level	Peak minus trough hourly activity
peak_activity_hour	Level	Hour of maximum mean activity
hourly_cv	Level	CV of hourly mean activity across waking day
sleep_hours	Level	Total hours in sleep state
sleep_onset_hour	Level	Hour at which sleep state first began
sleep_fragmentation	Level	State transitions per night (sleep to awake)
IS	Level	Interdaily Stability (7-day rolling window)
IV	Level	Intradaily Variability (7-day rolling window)
mean_activity_zscore_capped	Z-score	Within-person z-score, clipped to $[-5, 5]$
daytime_activity_zscore_capped	Z-score	Within-person z-score, clipped to $[-5, 5]$
sleep_hours_zscore_capped	Z-score	Within-person z-score, clipped to $[-5, 5]$
circadian_amplitude_zscore_capped	Z-score	Within-person z-score, clipped to $[-5, 5]$
mean_activity_pct_baseline	Ratio	Daily value / personal baseline mean
daytime_activity_pct_baseline	Ratio	Daily value / personal baseline mean

Table B.1 continued

Feature name	Type	Description
sleep_hours_pct_baseline	Ratio	Daily value / personal baseline mean
circadian_amplitude_pct_baseline	Ratio	Daily value / personal baseline mean
mean_activity_baseline_dev	Deviation	Daily value minus personal baseline mean
daytime_activity_baseline_dev	Deviation	Daily value minus personal baseline mean
sleep_hours_baseline_dev	Deviation	Daily value minus personal baseline mean
circadian_amplitude_baseline_dev	Deviation	Daily value minus personal baseline mean
mean_activity_7d_mean	Rolling	7-day rolling mean
daytime_activity_7d_mean	Rolling	7-day rolling mean
sleep_hours_7d_mean	Rolling	7-day rolling mean
circadian_amplitude_7d_mean	Rolling	7-day rolling mean
sleep_fragmentation_7d_mean	Rolling	7-day rolling mean
data_completeness_7d_mean	Rolling	7-day rolling mean wear completeness
mean_activity_14d_mean	Rolling	14-day rolling mean
daytime_activity_14d_mean	Rolling	14-day rolling mean
sleep_hours_14d_mean	Rolling	14-day rolling mean
circadian_amplitude_14d_mean	Rolling	14-day rolling mean
sleep_fragmentation_14d_mean	Rolling	14-day rolling mean
data_completeness_14d_mean	Rolling	14-day rolling mean wear completeness
mean_activity_7d_slope	Slope	Linear slope over 7-day window

Table B.1 continued

Feature name	Type	Description
daytime_activity_7d_slope	Slope	Linear slope over 7-day window
sleep_hours_7d_slope	Slope	Linear slope over 7-day window
circadian_amplitude_7d_slope	Slope	Linear slope over 7-day window
sleep_fragmentation_7d_slope	Slope	Linear slope over 7-day window
mean_activity_14d_slope	Slope	Linear slope over 14-day window
daytime_activity_14d_slope	Slope	Linear slope over 14-day window
sleep_hours_14d_slope	Slope	Linear slope over 14-day window
circadian_amplitude_14d_slope	Slope	Linear slope over 14-day window
sleep_fragmentation_14d_slope	Slope	Linear slope over 14-day window
peak_activity_hour_7d_mean	Rolling	7-day rolling mean of peak activity hour
peak_activity_hour_14d_mean	Rolling	14-day rolling mean of peak activity hour

Appendix C Pipeline

Reproducibility Notes

This appendix documents the software environment, data file structure, notebook execution order, and model artefact locations necessary to reproduce the results reported in this thesis.

Software Environment

All modelling experiments were executed on Kaggle Notebooks (Python 3.12, GPU disabled, 20 GB RAM). The key library versions used are listed in Table C.1.

Table C.1: Key software library versions used in the pipeline.

Library	Version
Python	3.12
pandas	2.x
numpy	1.x
scikit-learn	1.x
xgboost	2.x
shap	0.44.x
ruptures	1.1.x
matplotlib	3.x
seaborn	0.13.x
streamlit	1.x
joblib	1.x

Data Files

Three source data files were provided by Oiva Health and stored in a private Kaggle dataset ([rimshakaokab/vivago-data](#)). These files are not publicly available due to GDPR pseudonymisation requirements and the terms of the data sharing agreement between the University of Turku and Oiva Health. The row counts below reflect the raw unfiltered files prior to the preprocessing steps described in Section 3.4.1.

- `activity_data.csv`: 6,076,800 rows; columns: `user_id`, `timestamp`, `activity_level`, `state`
- `alarm_events.csv`: 128 rows (raw, pre-filtering); 59 clinically relevant alarm records retained after excluding battery and device status events; columns: `user_id`, `timestamp`, `alarm_code`, `severity`, `alarm_type`
- `wellbeing_flags.csv`: 15 rows; columns: `user_id`, `timestamp`, `flag_type`, `reason`

Notebook Execution Order

The pipeline notebooks must be executed in the following order. Each notebook reads outputs from preceding notebooks via a shared Kaggle input dataset to which intermediate files were uploaded after each run.

Table C.2: Pipeline notebook execution order and primary outputs.

NB	File	Primary outputs
01	01_preprocessing.ipynb	<code>alarms_clean.csv</code> , <code>flags_clean.csv</code>
02	02_eda.ipynb	EDA figures (F2–F6), window comparison (F4)
03	03_baseline-profiling.ipynb	<code>daily_normalized.csv</code> , F7, F8
04	04_feature-engineering.ipynb	<code>daily_features.csv</code> , F9
05	05_label-generation.ipynb	<code>X_train.csv</code> , <code>y_train_*.csv</code> (6 files), F10
06	06_logistic-regression.ipynb	LR LOUO results, PR curves Tasks A–F
07	07_random-forest.ipynb	RF LOUO results, PR curves Tasks A–F
08	08_xgboost.ipynb	XGBoost LOUO results, SHAP figures, <code>model.joblib</code>
09	09_trajectory_validation.ipynb	F16–F21, lead time CSVs
10	10_changepoint_experiment.ipynb	F22–F23, Task G results, <code>model_taskG.joblib</code>

Model Artefacts

The following serialised artefacts are required to run the proof-of-concept dashboard:

- `model.joblib`: XGBoost tuned classifier, Task C (`label_redflag_14d`), trained on the full 3,265-row dataset; 314.8 KB
- `model_taskG.joblib`: XGBoost classifier, Task G (`label_changepoint`), trained on the full dataset; 378 KB
- `feature_cols.joblib`: ordered list of 49 feature column names used as model input

The proof-of-concept application is launched from the project root directory with:

```
streamlit run poc_app/app.py
```

Appendix D Proof-of-Concept

Dashboard Interface

This appendix presents the proof-of-concept Streamlit dashboard developed as a deliverable of this thesis, demonstrating the operational integration of the Task C (RQ2: event prediction) and Task G (RQ1: state detection) models in a care coordinator interface. The dashboard was designed to require no specialist data science knowledge to operate and to present both research question outputs simultaneously with plain-language explanations.

Dashboard Structure

The dashboard comprises five interface components presented sequentially in the main page body, with a configuration sidebar on the left.

1. **Population overview panel (expandable):** An expandable table at the top of the main page listing all 20 monitored users ranked by current Task C event risk score, with the Task G state detection score shown alongside each entry. Three summary metrics display the cohort-level count of high-risk, monitoring, and stable users. This enables population-level triage without navigating individual records.
2. **Individual risk assessment panel (RQ1 and RQ2):** Two side-by-side

colour-coded status displays showing the current Task G state detection score (RQ1: Is this user currently deteriorating?) and the current Task C event risk score (RQ2: Is a critical event likely within 14 days?). Status levels are: Stable (score < 0.30 for Task C; < 0.20 for Task G), Monitoring (0.30–0.60 for Task C; 0.20–0.40 for Task G), and Deteriorating/High Risk (≥ 0.60 for Task C; ≥ 0.40 for Task G). Each panel displays a plain-language interpretation of the current score.

3. **Risk score timeline:** A dual-line chart displaying the Task C event risk score (orange, solid) and Task G state detection score (blue, dashed) over a user-selectable date range, with colour-coded background bands at 0.30 and 0.60. The default view covers the most recent 60 days.
4. **SHAP feature explanation panel:** A horizontal bar chart showing the top 10 SHAP feature contributions driving the current Task C risk score, computed for the most recent day in the view window. Red bars indicate features increasing predicted risk; blue bars indicate features decreasing it.
5. **Risk score history table:** A date-sorted table displaying Task C and Task G risk scores alongside raw feature values (daytime activity, sleep hours, data completeness) for the selected date range, supporting detailed inspection of individual monitoring days.

Sidebar controls include user selection (with current status labels), a "Jump to deterioration event" button for users with confirmed red flags (shifting the view window to the period surrounding the event), date range controls for the timeline and history table, and threshold reference guidelines.

Dashboard Screenshots

Figure D.1 shows the dashboard for a confirmed stable user (User_016), with both the RQ1 state detection panel and the RQ2 event prediction panel displaying green stable status. Risk scores remain well below the monitoring thresholds across the full observation period, consistent with the zero false-alert result reported in Section 4.3.6.

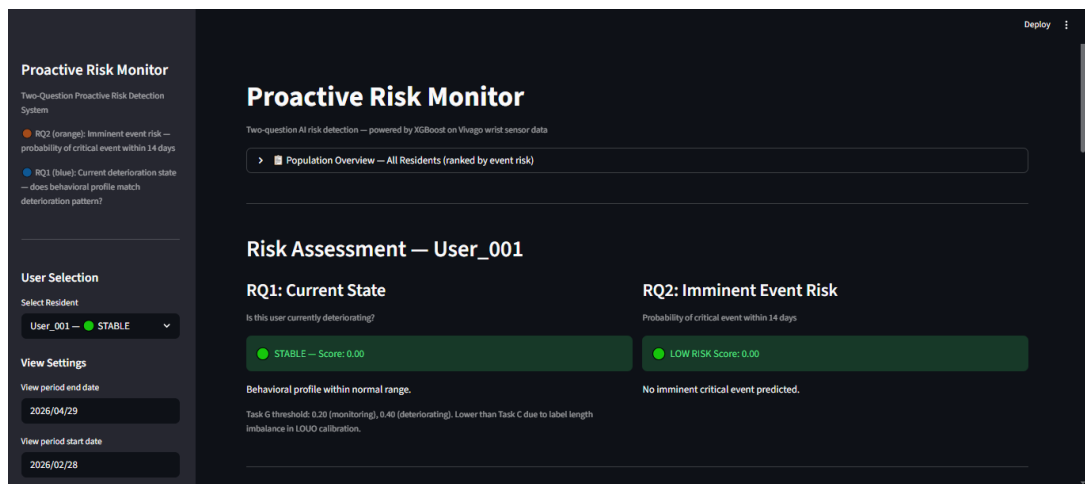


Figure D.1: Proof-of-concept dashboard showing the stable user view (User_016).

Figure D.2 shows the dashboard for User_003 at the time of the confirmed red wellbeing flag on 9 February 2026. The Task C event risk score has crossed the 0.60 threshold, triggering a high-risk alert in the RQ2 panel. The Task G state detection score reflects the chronic deterioration state that was detectable from November 2025 onward.

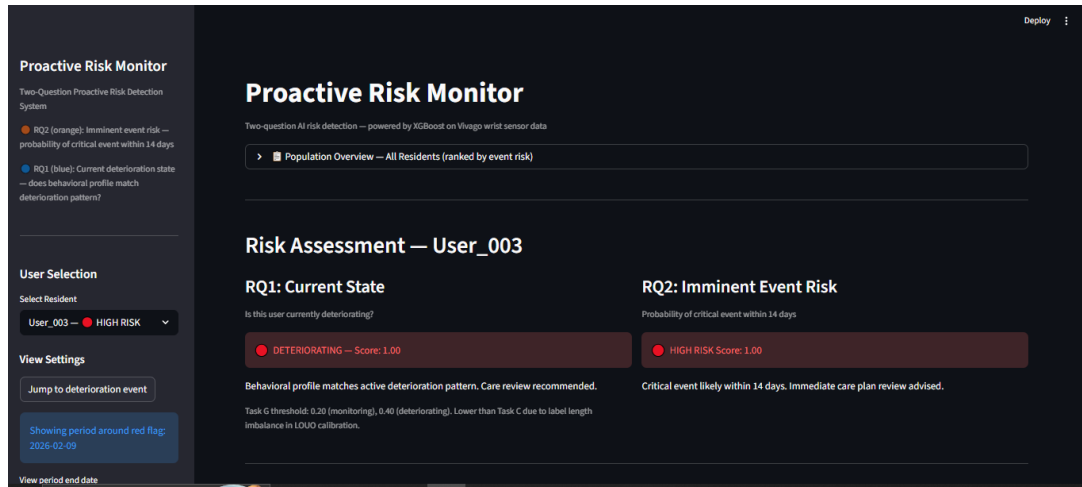


Figure D.2: Proof-of-concept dashboard showing the deteriorating user view (User_003).

Figure D.3 shows the SHAP feature explanation panel for User_003 at the alert date of 26 January 2026. The two dominant features driving the elevated risk score are the 14-day downward slopes in daytime activity and mean activity, consistent with the global SHAP analysis reported in Section 4.3.3. Red bars indicate features increasing the predicted risk score; blue bars indicate features decreasing it.

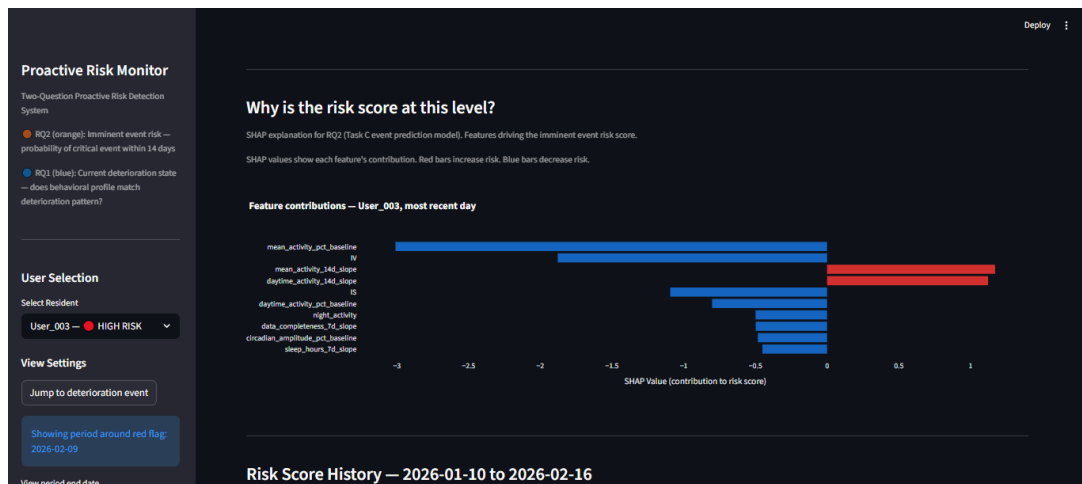


Figure D.3: Proof-of-concept dashboard showing the SHAP feature explanation panel for User_003.

Deployment Notes

The dashboard was tested locally against the pseudonymised Vivago dataset. It requires the three model artefacts listed in Appendix C and access to the daily features data file. No external API calls or network access is required at inference time. The application was not deployed to a public server during the thesis period, consistent with the GDPR data handling requirements applicable to the pseudonymised dataset.