Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Sleep Intent Boundaries from Wrist Actigraphy

Stuart Gallina Ottersen

stuart.g.ottersen@ntnu.no

Norwegian University of Science and Technology

Kerstin Bach

Norwegian University of Science and Technology

Simen Berg Saksvik

Norwegian University of Science and Technology

Research Article

Keywords: Sleep Onset, Actigraphy, Wrist accelerometry, Temporal Convolutional Network, Deep Learning, Sleep Restriction Therapy

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11280339/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Predicting Sleep Intent Boundaries from Wrist Actigraphy

Stuart Gallina Ottersen[1*], Kerstin Bach[1] and Simen Berg Saksvik[2]

[1*]Department of Computer Science, Norwegian University of Science and Technology, Trondheim, 7491, Norway.
[2]Department of Mental Health, Norwegian University of Science and Technology, Trondheim, 7491, Norway.

*Corresponding author(s). E-mail(s): stuart.g.ottersen@ntnu.no;
Contributing authors: kerstin.bach@ntnu.no; simen.b.saksvik@ntnu.no;

Abstract

Purpose : Insomnia is common, affecting roughly one in six people, and carries substantial personal and economic cost. A core diagnostic quantity is the delay between the start of sleep attempts and actual sleep onset, which requires knowing the moment a person begins trying to sleep. This moment is typically recorded via sleep diaries, laboratory technicians, or a button-press on a wrist device, all of which are costly or unreliable. We investigated whether wrist-worn accelerometry alone can locate this event automatically, at low cost and with minimal burden on patients.

Methods: We analysed wrist accelerometry from a two-arm, parallel-group randomised controlled trial in Norway of 582 adults with multiple sclerosis and clinically significant insomnia, comparing a digital therapy with a sleep-education control (NorseMS). Both arms were pooled for this secondary analysis. Participants' own button-presses served as reference labels, and concurrently collected sleep diaries served as a second independent reference. We present Lights-Out Time from Unobtrusive Sensing (LOTUS), a framework to predict the nightly button-press event, using a widened probabilistic target rather than a single time point and decoding the predicted distribution to a clock time. Performance was assessed by five-fold cross-validation split by participant, with localisation error in minutes as the primary measure, against a personal-median baseline, an established rule-based sleep-detection algorithm, and a pretrained self-supervised encoder. We conducted ablations and assessed agreement with diaries using Bland–Altman plots.

Results: LOTUS achieved a mean localisation error of 14.0 minutes and a median of 6.7 minutes. Its mean error was lower than all comparators; in terms

of median performance, a self-supervised baseline was comparable but failed on atypical nights, so LOTUS was more robust overall. Agreement between LOTUS and diaries was comparable to that between button-presses and diaries. Down-sampling and a latency analysis indicated LOTUS captures pre-sleep behaviour rather than the button-press movement or sleep onset itself.
Conclusions: Accelerometry alone can locate sleep intent about as closely as existing self-report measures do, potentially recovering nights without a button-press and reducing reliance on diaries and manual marking in insomnia trials and care.

Keywords: Sleep Onset, Actigraphy, Wrist accelerometry, Temporal Convolutional Network, Deep Learning, Sleep Restriction Therapy

1 Introduction

Sleep disturbances are a major public health issue that carries high economic and personal costs. Insomnia has an estimated prevalence of 16.2%, and increases the risk of comorbidities [1–4]. Sleep problems are thought to have a bidirectional relationship with other mental health conditions, meaning sleep treatments can help reduce the onset of clinical disorders [4]. Work on sleep treatments is therefore crucial both to improve sleep health in the general public and to mitigate mental health disorders. One crucial aspect of developing and delivering sleep treatments is the tracking of sleep, and digital technologies that track sleep more accurately and at lower cost can be very valuable.

One important aspect is the ability to track transitions between wake and sleep. A night of sleep can be decomposed into a sequence of events: getting into bed, stopping wakeful activity and attempting to sleep, eventually falling asleep, and finally waking and rising. The interval between the second and third of these, the sleep onset latency (SOL), is the quantitative index of sleep-initiation difficulty, a core symptom of insomnia disorder, and is widely used both in research and diagnostic criteria [5]. Computing it requires a timestamp for the sleep intention event, meaning the moment the person begins trying to sleep.

Work exists on predicting sleep onset and offset from accelerometer or PSG data [6–8], but the sleep intention event is comparatively underexplored despite its diagnostic importance. It is commonly tracked using sleep diaries, lab technicians, or button-press events on actigraphs. Each of these has its own issues: low button-press compliance due to forgetfulness, accuracy issues in sleep diaries caused by recall bias, and the high cost and inaccuracy of lab technicians [9, 10]. Reducing the patient burden and cost of tracking this can therefore be highly useful, both for routine clinical monitoring and for extracting additional information from actigraphy datasets that have already been collected.

Wrist-worn wearable devices offer a route to doing so. They are low cost, unobtrusive, and well suited to long-term remote monitoring outside the clinic. Accelerometry does not directly observe intent, but it does observe its behavioural correlates. We hypothesise that the transition between wake and sleep intent represents a shift between two locally stationary regimes and that this is detectable in raw accelerometer data. This makes wrist-worn wearables a plausible tool for automatic, AI-driven intent detection, at a fraction of the cost and participant burden of the alternatives.

Beyond SOL, automatic intent detection could also support digital and conventional insomnia treatment by enabling objective measurement of adherence to a prescribed sleep window. Sleep restriction therapy, a core component of cognitive behavioural therapy for insomnia (CBT-I), works by assigning the patient a fixed time-in-bed window and progressively widening it as sleep efficiency improves [11]. The therapeutic mechanism depends on the patient actually adhering to the prescribed window, yet adherence is typically assessed through the same self-report diaries whose recall bias motivated this work in the first place, and non-adherence is a known contributor to treatment failure and dropout [12, 13]. Passive, wearable-based adherence monitoring could therefore strengthen the delivery of CBT-I, including app-based digital CBT-I (dCBT-I) programmes.

To test this hypothesis, we introduce Lights-Out Time from Unobtrusive Sensing (LOTUS), a machine learning framework that predicts the timing of the nightly button-press event from the wearable accelerometer signal alone. We train on multiple sclerosis trial data, using a total of 9606 days of accelerometer data from 513 participants who were asked to press a button on the device when trying to sleep and again on waking. LOTUS achieves a mean error of 14.00 minutes and a median of 6.72 minutes. Because the labels themselves are imperfect, we additionally evaluate LOTUS against concurrently collected sleep diary entries, allowing us to characterise the disagreement between the two reference measures rather than treating either as ground truth.

Our contributions are: **(1)**, we formalise sleep intent boundary detection as a prediction task over raw wearable accelerometer data, a problem that has not previously been posed in this form despite its role in SOL computation and insomnia diagnosis. **(2)**, we introduce Lights-Out Time from Unobtrusive Sensing (LOTUS), and report its performance on a cohort of 513 participants, establishing a first baseline for future work. **(3)**, we conduct a series of ablation studies on LOTUS. **(4)**, we conduct a three-way comparison between LOTUS, button-press labels, and sleep diary entries, quantifying the extent to which the two conventional reference measures diverge from one another and from LOTUS. **(5)**, we outline the application of automatic intent detection to SOL measurement and to adherence monitoring in CBT-I and the digital alternative dCBT-I.

The remainder of this paper is organised as follows. Section 2 reviews related work on sleep event prediction. Section 3 describes the proposed framework and data

preprocessing. The experimental setup is presented in Section 4, followed by the results in Section 5. Section 6 briefly introduces some application areas for the framework. We discuss the results in Section 7 and their limitations in Section 8, and Section 9 concludes the paper.

2 Related Work

Event detection in time series is commonly organised by the type of event sought, the degree of supervision available, and the strategy used to produce the detection [14]. Given this definition our work addresses punctual change point detection, a single instant separating two regimes, here wakefulness and sleep intent. Classical change-point detection addresses exactly this event type, but does so unsupervised and theory-driven, locating boundaries by optimising a segmentation cost or by maintaining a posterior over the time since the last change [15, 16]. An accelerometer signal contains many regime changes per day, and nothing in such a criterion identifies which of them corresponds to intent, so a supervised, data-driven detector is required. Among supervised strategies, the classification approach labels every timestep and recovers events by thresholding the resulting score sequence, as in dense activity labelling from wearables [17], typically using temporal convolutional networks or patch-based transformers as the backbone [18, 19], a per-frame objective is a poor fit when exactly one event is expected per window. The regression strategy instead predicts the event time directly, but a scalar output must commit to a single value even where several candidate moments are plausible. We therefore adopt an intermediate formulation borrowed from keypoint detection, in which the point label is replaced by a Gaussian target over positions and a continuous estimate is recovered by taking an expectation over the predicted distribution [20].

With an increasing number of large-scale population-based actigraphy datasets, such as UK-Biobank, HUNT, and NHANES, there is growing interest in applying machine learning models to actigraphy datasets [21–23]. A wide variety of tasks have been explored in sleep, such as sleep/wake classification, sleep stage classification, sleep parameter estimation or even psychiatric screenings [7, 24–26]. Works such as [6] attempt to predict sleep period time using accelerometer data by training against polysomnography and sleep diary data. Similarly, [27] estimates time in bed using accelerometer data. More traditional algorithms, such as [28, 29], have also been used for sleep window prediction, with more recent development showing interest in self-supervised models for sleep/wake prediction [30, 31].

To our knowledge, no machine learning work exists for estimating sleep intent from actigraphy, but some wearables have features that estimate sleep duration. Estimates of lights-off time are usually based on questionnaire data, marked by others in lab experiments or an event marker press [32–34].

3 Methods

We treat the onset of sleep intent as a boundary between two locally stationary activity regimes, and cast its detection as predicting a single event within a 24-hour window of wrist accelerometry. The participant's nightly button-press serves as a proxy label for this event. Below, we describe the framework architecture (LOTUS), how its output is decoded into a predicted clock time, and the training task.

3.1 Lights-Out Time from Unobtrusive Sensing

The LOTUS architecture is shown in Figure 1. A window of accelerometry is split into patches and each embedded by a shared convolutional encoder (CNN). The patch sequence receives a positional encoding (PE) and is passed on to a temporal backbone and a prediction head, whose output is decoded to a clock time and compared with the button-press label. The highlighted components (positional encoding, backbone, head, and decoder) each have two variants that we compare in the ablation studies (Section 4.4.2); the remaining components are fixed. We use LOTUS to refer to the framework and *LOTUS model* to refer to an instance of the framework.

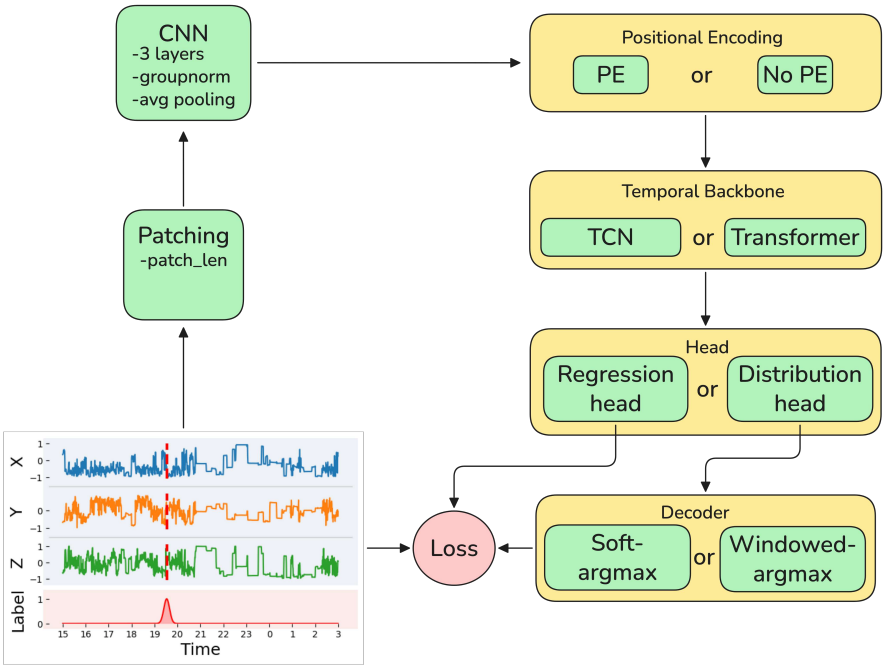


Fig. 1 Showing tri-axial data and the structure of LOTUS.

LOTUS operates on tri-axial actigraphy split into non-overlapping patches of length *patch_len*, a searched hyperparameter (see Table **??** in the appendix). Because

a 24-hour window at 10 Hz contains on the order of 864,000 samples, patching is what makes the sequence tractable for the temporal backbone.

We hypothesise that wake and start of sleep intent constitute two locally stationary regimes and that the intent event is the boundary between them; capturing local structure within each patch should therefore be sufficient for the task. Each patch is independently embedded by a CNN followed by global average pooling, yielding one embedding per patch.

We then add a PE to the sequence. Weight sharing in the convolutional stack makes the transition detector position-independent, but it also means the stack carries only relative position information. Since every window is a fixed 24-hour segment, the PE supplies absolute clock time, which LOTUS can condition on.

The patched sequence is fed into a temporal backbone. We test two architectures, the Temporal Convolutional Network (TCN) and the transformer. The TCN was chosen because the wake–sleep transition is largely independent of its timing and because we use locality importance as an inductive bias. The transformer is included for its ability to capture global structure. The prediction is performed retroactively, so the TCN is not autoregressive. This is crucial, as the reduced activity following lights-off is as informative as the activity preceding it.

The LOTUS framework is tested using two types of tasks, a distribution prediction task and a regression task. This calls for two different types of heads.

The distribution task is tested with two different types of decoders. The task of these decoders is to turn the predicted distribution into a hard prediction to compute a more understandable metric. The implementation is made available on GitHub[1].

3.2 Decoder of soft predictions

When using a soft target variable, LOTUS does not predict a single time point but a distribution over the M patches of the input window, i.e. a soft prediction. While this is the quantity optimised during training, it is difficult to interpret: a cross-entropy value provides no information about how far a prediction is from the true event. As a more intuitive measure, we therefore report the average time difference between the predicted and the actual button-press event, expressed in minutes. This requires collapsing the soft prediction into a hard one, i.e. the exact moment at which the *LOTUS model* believes the button-press occurred. The distribution decoders below both produce an index for a patch representing the point in time the button-press event was recorded. To evaluate, we map the estimate back to a sample position relative to the start of the recording, using the centre of the patch,

$$\hat{y} = \hat{\jmath} \cdot patch_len + patch_len/2, \tag{1}$$

and report the localisation error in minutes as $|\hat{y} - y|/(60 f_s)$, where f_s is the sampling rate in Hz.

[1] github.com/stuartgo/lotus

Expected patch (soft-argmax).

The first approach interprets p as a distribution over patch indices and takes its expectation,

$$\hat{j}_{\text{soft}} = \mathbb{E}_{j\sim p}[\, j \,] = \sum_{j=0}^{M-1} j\, p_j. \tag{2}$$

Because $\hat{j}_{\text{soft}}$ is real-valued rather than an integer index, the estimate is not restricted to patch centres and can represent event times at a finer resolution than the patch length *patch_len*. Its drawback is that it is the centre of mass of the entire distribution: if p is multi-modal, for example when there are two plausible candidate moments, the expectation falls between the modes and may land in a region the model itself considers unlikely.

Windowed argmax (local centre of mass).

The second approach starts from the mode of the distribution,

$$k^* = \arg\max_k\, p_k, \tag{3}$$

which by construction lies in a high-probability region and is therefore insensitive to secondary peaks. Taking k^* directly, however, discards all sub-patch information and incurs a quantisation error of up to *patch_len*$/2$ samples. We therefore refine the mode by a centre-of-mass computation restricted to a local window of $\pm w$ patches around the peak. With $\mathcal{W} = \{\, k : |k - k^*| \le w \,\}$,

$$\hat{j}_{\text{arg}} = \frac{\sum_{k\in\mathcal{W}} k\, p_k}{\sum_{k\in\mathcal{W}} p_k}. \tag{4}$$

Probability mass outside $\mathcal{W}$ is ignored, so the estimator combines the robustness of the mode with the sub-patch precision of the expectation.

4 Experimental setup

In this section, we present the tasks used during training and the data used for the experiments.

4.1 Data

The data used is from NorseMS, a two-arm, parallel-group randomised controlled trial conducted in Norway with 582 participants [35]. Participants were recruited through the Norwegian MS Registry, which covers more than 90% of people diagnosed with MS in Norway, and were eligible if they were at least 18 years old and reported clinically significant insomnia. Eligible participants were randomised 1:1 to either fully automated digital cognitive behavioural therapy for insomnia (dCBT-I) or a static web page with patient education about sleep, which served as the control condition. The trial is conducted entirely online, from screening through follow-up, and the primary endpoint is the between-group difference in self-reported insomnia severity nine

weeks after randomisation, with further assessments at 6 and 12 months. Wrist-worn actigraphy was collected throughout the intervention phase as a secondary, objective measure of sleep and daytime activity.

This actigraphy forms the basis of the present work, and participants from both the control and intervention groups are included. Each participant wore a GENEActiv actigraph mounted on the wrist for 90 days, covering the intervention period and including the 10 days before and after it, recording tri-axial accelerometry at 10 Hz. Participants were instructed to press the event button on the device when going to sleep and upon waking, giving self-reported sleep and wake times alongside the raw signal. In total, 26 208 nights with no missing accelerometer data were collected, of which 33.1% contained no button-press during the night window (18:00–06:00). Sleep diaries were collected for the 10 days before and after the intervention, and additionally throughout treatment for the treatment group. The written instruction given to participants was: *Press the button every night when you go to sleep and every morning when you get up.* A follow-up phone call clarified that the button should be pressed when intending to sleep rather than when getting into bed. Similarly the diary question was: *At what time did you try to sleep?*

4.2 Data representation

We represent each night as a 24-hour window of tri-axial actigraphy sampled at 10 Hz. Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathbb{R}^{N\times 3}$ denote such a window, where $\mathbf{x}_n \in \mathbb{R}^3$ is the reading at sample index n. Each window is associated with a single label $y \in \{0, \ldots, N-1\}$, the index at which the participant pressed the button on the device to indicate they were beginning to try to fall asleep. M represents the number of patches, given by $N/patch_len$

Our solution is a deep learning model f_θ that recovers y from X alone. We test two formulations of this prediction task. In the first, f_θ regresses the patch index directly, $\hat{j} = f_\theta(X) \in \mathbb{R}$. Some nights contain multiple plausible sleep bouts, so the posterior over intent time has several modes, and a scalar head is forced to land between candidates rather than committing to one.

The second formulation, therefore, addresses this by predicting a distribution over patches, $p = f_\theta(X) \in \Delta^{M-1}$, and replacing the point label with a soft target built from a radial basis function kernel centred on the label's patch, $j_y = \lfloor y/\texttt{patch_len} \rfloor$:

$$\mathbf{y}_{\text{soft}} = \text{softmax}\left(-\frac{(\mathbf{m} - j_y)^2}{2\sigma^2}\right), \qquad \mathbf{m} = (0, \ldots, M-1). \tag{5}$$

The width σ is specified in minutes as $\sigma_{\min}$ and converted to patches by

$$\sigma = \sigma_{\min} \cdot 60 \cdot f_s \cdot patch_len^{-1}. \tag{6}$$

The model is trained against this target with a cross-entropy loss:

$$\mathcal{L} = -\sum_{m=0}^{M-1} y_m^{\text{soft}} \log p_m. \tag{7}$$

Figure 2 shows a single button-press event alongside the resulting soft target. Widening the label in this way increases the number of patches contributing to the loss for each training example, producing a denser training signal and a smoother loss landscape than the sparse, high-variance signal from a single-point label.

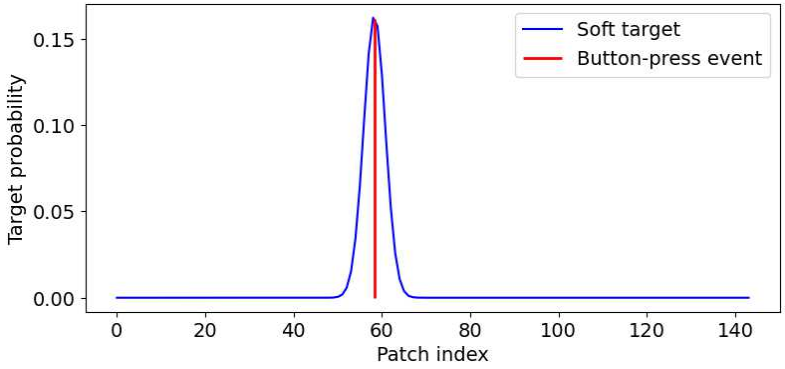


Fig. 2 A button-press label and the soft target derived from it: The single button-press event (red) is replaced by a Gaussian soft target (blue) centred on the same patch, so more patches contribute to the loss.

4.3 Data preprocessing

Each participant usually has two or three recordings from different actigraphs, since the battery life of a single device does not cover the full 90-day protocol. These recordings are not concatenated. Recordings are read at the native sampling rate of 10 Hz and split into individual 24-hour segments running from 15:00 to 15:00, starting at the first 15:00 boundary in each recording. Segments are defined as fixed blocks of 864,000 consecutive samples; any trailing partial segment is discarded. Only the triaxial acceleration (x, y, z) is retained. Pooled 1 Hz, 1-minute and 5-minute versions are obtained by averaging non-overlapping windows of the 10 Hz signal, and are used where noted.

4.3.1 Non-wear and missing data

Non-wear is detected with the GGIR (commonly used R package for actigraphy analysis) acceleration-only algorithm as implemented in `wristpy`, which operates on the native-rate signal and labels 90-second windows. A 24-hour segment is discarded if *any* of its samples is flagged as non-wear or has a missing value in any axis. This is deliberately conservative: LOTUS is trained only on days that are complete and worn throughout.

4.3.2 Button-press selection

The most challenging part of the preprocessing is selecting button-presses. Patients are instructed to press when they try to sleep and when they wake up. There are, therefore, cases in which multiple sleep bouts can be included within a single 24-hour segment.

The challenge is therefore to select the "correct" button-presses for training. We developed a rule base that determines which button-presses should be treated as intent-to-sleep events. Presses falling within 2 minutes of a preceding one are removed. Each remaining button-press is assigned to a morning class if it occurs between 03:00 and 14:00, and to a night class if it occurs between 18:00 and 06:00; presses outside these windows are discarded. Button-presses can therefore be analysed as both morning and night presses. All pairings of a night button-press with a subsequent morning button-press whose separation lies between 2 and 14 hours are then evaluated.

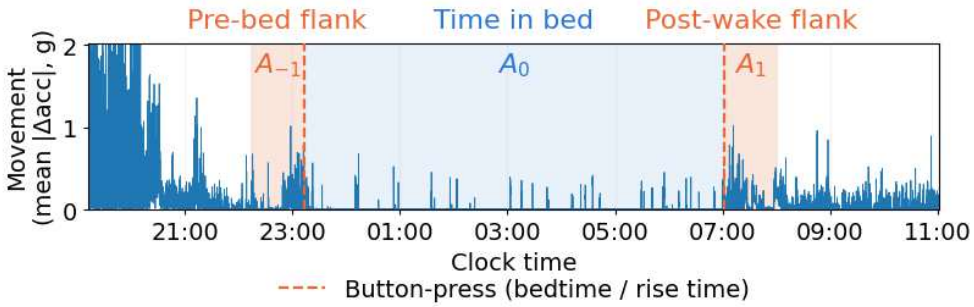


Fig. 3 Activity-contrast windows used to select the nightly button-press pair. Mean absolute acceleration in the time-in-bed interval (A_0) is compared with the flanking pre-bed and post-wake windows (A_{-1}, A_1).

To avoid the selection of accidental button-press events we aim to find a time slot bounded by considerably more active hours. For each candidate pair, we maximise the contrast in 1-minute binned mean absolute first difference in movement between the interval spanned by the two presses and the 60-minute windows flanking each press, as shown in Figure 3, with 5-minute guard bands excluded on either side of each press. Activity in both flanking windows must exceed that of the enclosed interval by a factor of at least $\alpha = 1.5$. We formalise the process in Equations (8) and (9). For the analysis only the lights-off button-press is of interest, but due to the activity window based filtering only nights with both night and morning button-presses are kept.

For each 24-hour segment d, let $\bar{a}_t$ be the mean of $\frac{1}{3}\left(|\Delta a_n^x| + |\Delta a_n^y| + |\Delta a_n^z|\right)$ over the samples in 1-minute bin t, and $median(A) = \text{median}_{t \in A}\, \bar{a}_t$. For a candidate pair $c = (t_b, t_w)$ with $2\,\text{h} \leq t_w - t_b \leq 14\,\text{h}$, write $A_0 = [t_b + g,\, t_w - g]$ for the interval

between presses and $A_{-1}, A_1 = [t_b - g - \ell,\, t_b - g],\, [t_w + g,\, t_w + g + \ell]$ for the flanking windows, with guard $g = 5\,\text{min}$ and flank length $\ell = 60\,\text{min}$.

$$C_s(c) = \frac{\text{median}\big(A_s(c)\big) + \varepsilon_d}{\text{median}\big(A_0(c)\big) + \varepsilon_d}, \qquad s \in \{-1, +1\}, \qquad \varepsilon_d = 0.05 \cdot \text{median}_{t \in d}\, \bar{a}_t \tag{8}$$

$$c_d^* = \underset{c\,:\,\min_s C_s(c) \geq \alpha}{\arg\max} \sum_{s \in \{-1,+1\}} \log C_s(c), \qquad \alpha = 1.5 \tag{9}$$

4.3.3 Training target and data volume

We analyse the distribution of the selected nights from the NorseMS dataset and compare it with the nights lost to the button-press selection logic. Nights for which no valid pair could be selected are unlabeled and excluded from training. Starting from 26,208 nights, 9,606 remain after preprocessing. This is a retention rate of about 37%, which is quite low. Looking at the distribution of kept nights across treatment days, it is quite even, as shown in Figure 4.

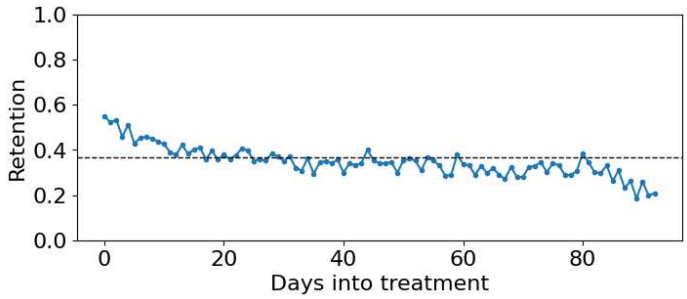


Fig. 4 Fraction of nights retained after preprocessing (by day). Proportion of recorded nights passing button-press selection against days since first measurement; dashed line: overall mean retention ($\approx$ 0.37).

The filtering process also removed 58 of the 571 participants with usable actigraphy (approximately 10%). For 54 of these 58 participants, removal occurred because they had fewer than five nights with both a morning and an evening button-press, indicating that participant loss was driven primarily by low adherence to the button-press protocol. At the night level, 16602 nights were removed in total. Of these, 11,366 (68.5%) lacked either a morning or an evening button-press, while the remaining 5,236 (31.5%) were excluded by the other filtering rules.

4.3.4 Splits and normalisation

The data are split into 5 folds at the participant level with a fixed random seed, so that all recordings and all nights from a given participant fall in the same fold.

For each fold, one held-out fold is used for testing, the next for validation, and the remaining three for training. Per-axis means and standard deviations are computed on the training rows of that fold only and applied unchanged to the validation and test folds, so no normalisation statistics leak across the split.

4.3.5 Hyperparameter optimisation

Hyperparameter optimisation is performed on all trained models. We run 30 trials with a Tree-structured Parzen estimator, using median pruning, which stops trials whose median performance at the current step is worse than that of other trials [36]. A detailed overview of the hyperparameters used can be found in the Appendix.

4.4 Experiments

We perform three sets of experiments, the first tests three baselines, we then introduce and test LOTUS with a series of ablation studies, and finally, we test the behaviour of LOTUS in different contexts.

Each experiment is run with 5-fold cross-participant splits. Hyperparameters are selected on the validation split of a single fold and then applied unchanged to all five, and all reported results are test-fold results from the best checkpoint of each fold. The experiments are run on 1Hz data due to computational cost.

4.4.1 Baselines

We compare LOTUS against three baselines: a personal median (naive baseline), the HDCZA sleep-onset algorithm [6], and harnet10 [37], a self-supervised model pretrained on actigraphy data.

Personal median

Bedtimes are habitual, so a participant's own history is a strong predictor. For each night, we take the median button-press time of all other nights from the same participant and use it as the prediction.

Sleep period onset (HDCZA)

The established actigraphy approach to locating the night without a sleep diary is the heuristic distribution of change in z-angle [6]. Rather than measuring how much a participant moves, it measures how much the wrist changes orientation: the angle of the device relative to horizontal is computed over five-second epochs, smoothed with a five-minute rolling median, and differenced. Epochs whose change falls below a threshold are marked inactive, runs of at least thirty minutes are retained, and runs separated by less than sixty minutes are merged. The longest resulting block is the sleep period time (SPT) window, and we take its start as the predicted bedtime. We use the implementation in `wristpy` [38], which uses the current GGIR default of a fixed 0.2° threshold rather than the adaptive percentile rule in the original publication.

Pretrained self-supervised encoder (harnet10)

LOTUS is trained from scratch on a single cohort, which raises the question of whether a representation learned from population-scale accelerometry would serve better. We therefore evaluate the self-supervised encoder of [37], a residual network pretrained on approximately 700,000 person-days of UK Biobank wrist accelerometry data, as a frozen feature extractor. Each night is divided into ten-second windows, each window is mapped to a 1024-dimensional embedding, and a small dilated convolutional head is trained on the resulting sequence to predict the button-press. The head is supervised with the same soft-label objective and decoded with the same rule as LOTUS.

Two properties of this baseline should be borne in mind when reading its results. The encoder was pretrained at 30 Hz, whereas our recordings are sampled at 10 Hz; windows are therefore interpolated to the pretraining rate, which introduces no additional information and constitutes a genuine domain shift. The model is trained for 30 epochs.

Offset correction

The HDCZA model estimates the onset of sleep rather than the moment a participant decides to sleep. The predictions, therefore, reflect a systematic shift from the button-press, which is a property of the target definition, not of detection quality. To make the comparison fair, we therefore add a shift to compensate.

For each fold, $offset$ is the median of the signed differences between label and prediction over the training folds, which is the constant minimising mean absolute error. This is added to the predictions. All differences are computed circularly, so times either side of midnight are compared along the shorter arc.

4.4.2 Ablation studies

We show a comprehensive set of ablation studies and present them in Table 1. LOTUS is tested and then each component is removed or replaced to test the effect it has on performance. The components we test are: positional encoding vs no positional encoding. The use of a transformer encoder vs a TCN. We additionally vary the *LOTUS model's* head by testing it as a regression task or as one requiring the model to reconstruct a distribution. Finally, we test the decoder on the distribution produced.

Table 1 Overview of ablation studies. The first row gives the LOTUS configuration; each ablation study changes one component, and "–||–" denotes the LOTUS setting. PE is Positional Encoding.

Study	PE	Temporal model	Head	Decoder
LOTUS	Yes	TCN	Distribution	Soft-argmax
A1: PE	No	–||–	–||–	–||–
A2: Temporal model	–||–	Transformer	–||–	–||–
A3: Head	–||–	–||–	Regression	None
A4: Decoder	–||–	–||–	–||–	Windowed-argmax

4.4.3 Model behaviour analyses

To further understand LOTUS's behaviour we run additional experiments to test if the *LOTUS model* behaves as expected. The button-press itself is an event in the accelerometer data so we run experiments testing LOTUS in reduced resolution data. The accelerometer data gathered in the trial is collected alongside sleep diaries, we run an experiment testing the agreement of LOTUS vs. button-press vs. diary data. We continue this by testing LOTUS behaviour against Sleep Onset Latency from the diaries

5 Results

In this section, we present the main results from our experiments and ablation studies. We also discuss the potential impact of these results on sleep research.

5.1 Baselines

To compare LOTUS's performance we compare against three baselines: a naive personal median of each participant's button-press times, the rule-based sleep detection algorithm HDCZA, and the pretrained deep network harnet10. Performance is measured as the absolute error, in minutes, between the predicted and recorded button-presses. We see considerable variation across the three in Table 2.

Table 2 Performance of baselines on predicting button-presses.

Model	Mean error [min] ↓	Median error [min] ↓
Personal median	41.32	30.82
HDCZA	27.17	7.22
harnet10	72.47 ± 5.73	5.84 ± 0.88

The personal median works as a floor as a naive baseline, but is still able to beat harnet10 on the mean score. However harnet10 has considerable better median performance. The poor mean performance of harnet indicates overfitting to the majority of samples and struggling with the more difficult samples. The same behaviour is seen for HDCZA, but HDCZA has a lower difference between mean and median while scoring in the same range as harnet10 for the median. Only std. is given for harnet10 as it's a statistical approach while the two others are rule based.

We present the results from the ablation studies introduced in Table 1 in Table 3. Overall LOTUS performs the best and also outcompetes the baselines by a considerable margin on the mean, while the median error is comparable, with harnet10 slightly lower. This indicates LOTUS is able to perform considerably better on the more difficult samples, reducing the skew of the error distribution. The baselines on the other hand can only capture sleep intent times close to the sleep onset, which is the case in the majority of nights. The more difficult-to-predict nights are not

predicted nearly as well, skewing the mean.

Table 3 Ablation study.

	Mean	Median error [min] ↓
LOTUS	$\mathbf{14.00 \pm 0.74}$	$\mathbf{6.72 \pm 0.43}$
A1: PE	15.13 ± 1.72	7.59 ± 0.56
A2: Temporal model	14.17 ± 0.89	6.77 ± 0.50
A3: Head	18.29 ± 1.92	10.42 ± 0.51
A4: Decoder	14.88 ± 1.55	7.00 ± 0.59

Position Encoding

LOTUS without PE performs slightly worse in both score and stability compared to with PE. This indicates that the PE helps the TCN capture and place events accurately.

Temporal backbone

We compare two temporal structures, a TCN and a transformer encoder. The difference between the transformer and TCN, 0.17 minutes on average, is similar, so the two backbones perform comparably on this task. This suggests that the information needed to locate the transition is largely local, and that a model with a global receptive field gains little beyond it.

Head

We test regression vs. distribution prediction and see that regression performs considerably worse. This could be due to the weaker label signal from a single distance between the prediction and the target. The distribution prediction task receives a distribution which provides a loss that is more informative.

Decoder

Soft-argmax decoding gives a slightly lower mean error than windowed argmax and a noticeably smaller spread across folds. We therefore treat soft-argmax as the more stable of the two rather than as clearly more accurate. One plausible reason for the reduced variance is the within-patch resolution of soft-argmax compared to windowed argmax. Windowed argmax can only place a prediction at patch granularity, with some correction from the windowing process. The lower performance of windowed argmax might also indicate that most LOTUS-predicted distributions are unimodal, since windowed argmax is primarily expected to help when the predicted distribution has several competing modes. Soft-argmax is used for the rest of the experiments.

5.2 Model behaviour analyses

To assess its generality, we examine how LOTUS behaves. Figure 5 shows the error distribution pooled over the test set in the NorseMS dataset. It is strongly right-skewed. Inspection of the largest errors suggests that many of them arise from nights with unusual sleep times, where the button-press selection logic is more likely to pick the wrong button-press as the nighttime label.

The values differ from Table 3 because this is averaged over all individuals while the table averages over the folds.
This is further shown in Figure 6 where the error increases at the extremes, due to

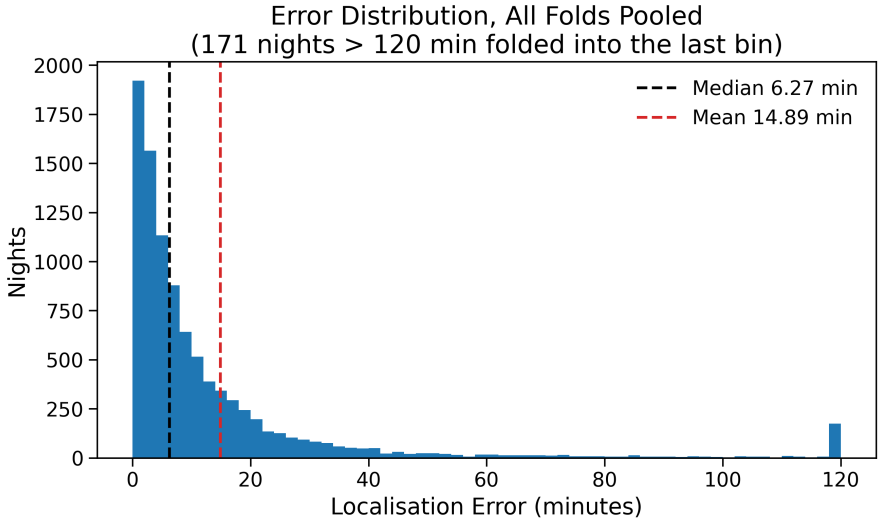


Fig. 5 Distribution of absolute prediction error, pooled over the test set from all folds.

considerably fewer datapoints. This result highlights the *LOTUS model's* strength while revealing its limits. The predictions are better for the more typical bedtimes, reflecting the greater data abundance, while showing the model's limits at the extremes.

A particular source of this skew is nights with multiple bouts of sleep, which are not uncommon among MS patients. Figure 7 shows such a case: the recording contains a series of button-presses consistent with several sleep bouts, and LOTUS predicts one of them closely, but not the one that the button-press filtering selected as the reference. Since the *LOTUS model* emits a single transition per night, it has no way to express this ambiguity, and the resulting disagreement is counted in full as error even where the prediction is plausible.

Because the reference labels are button-presses, and a button-press is itself a movement recorded by the accelerometer, there is a risk that the *LOTUS model* learns to detect the press rather than the behaviour surrounding it. The press is a brief,

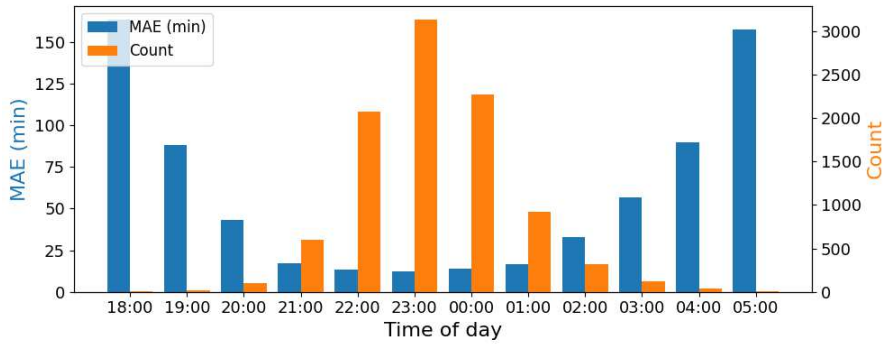


Fig. 6 Grouped barchart showing error distribution and number of samples by time of day.

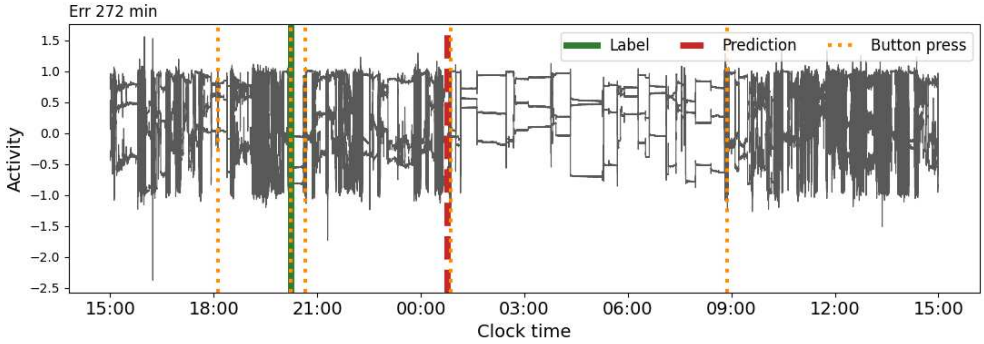


Fig. 7 Showing movement data and button-presses for participant with multiple bouts of sleep. Dotted lines represent button-presses, green line shows the button-press selected as label. Red dashed lines is the prediction of LOTUS.

high-frequency event, so reducing the sampling rate should remove its signature while leaving the slower changes in activity that precede sleep intact. If performance survives downsampling, the model cannot rely on the press alone.

Figure 8 shows the error across sampling rates, where lower resolutions are obtained by patched averaging of the original 10 Hz signal. Performance decreases slightly on average as the frequency decreases, but surprisingly seems to improve in the 1Hz compared to 10Hz. Performance remains good and, on average, better than the baselines, even with the 5-minute samples. We take this as evidence that LOTUS is capturing the behavioural context of sleep onset rather than the button-press artefact. This could also speak for the generalisability of LOTUS as its performance is stable across frequencies.

As a secondary check on the *LOTUS model's* behaviour, we examine the spread of the predicted distribution over time. Figure 9 shows the error by quintile of the standard deviation of that distribution: error increases monotonically with spread. The relationship holds at the level of quintiles, it is not reliable enough to attach a

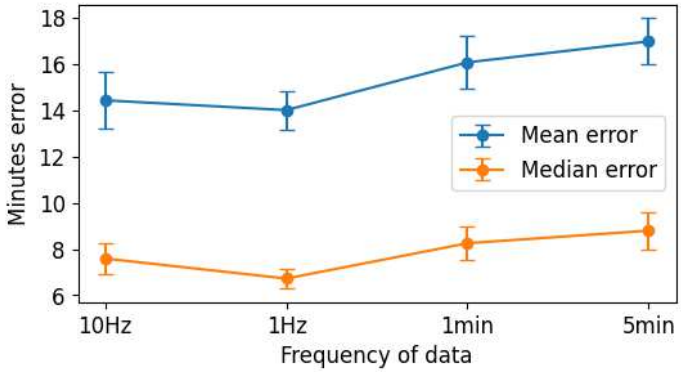


Fig. 8 Prediction error across data sampling rates (10Hz, 1Hz, 1min, 5min).

meaningful confidence value to any individual prediction, and we do not use it that way.

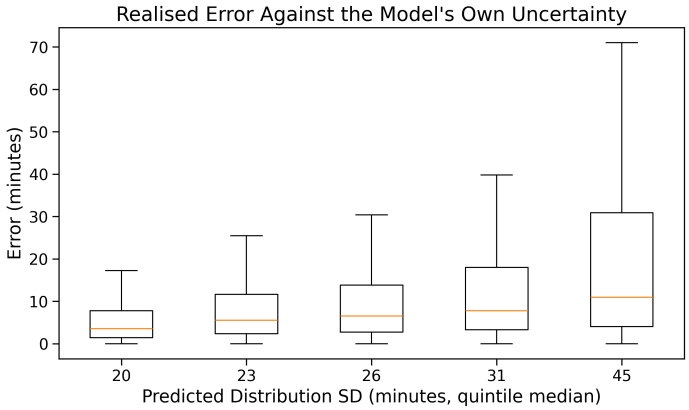


Fig. 9 Error by quintile of the standard deviation of the predicted distribution.

6 Application

In this section we discuss the applications of LOTUS by analysing how its behaviour relates to sleep diaries.

6.1 Comparison with sleep diaries

The button-press is a proxy rather than a ground truth, so we additionally compare against self-reported sleep diaries. Diary coverage is considerably sparser than actigraphy coverage, but the comparison indicates how the *LOTUS model* behaves relative to an independent report of the same event. Figure 10 shows Bland–Altman plots for

each pair of measures. This plot shows the difference between the two methods by plotting them against their mean. A tall plot, therefore, means high disagreement. The middle dashed line represents the bias, in other words, the mean difference. The two other dashed lines represent the limits of agreement, which are 1.96 standard deviations from the mean.

The informative comparisons here are between LOTUS and the diary on the one hand, and between the button-press and the diary on the other, since these are the two comparisons in which the diary serves as an independent reference. Both the average error and the spread are similar for LOTUS vs Diary and Button-press vs Diary, indicating that LOTUS agrees with the diary about as well as the button-press does.

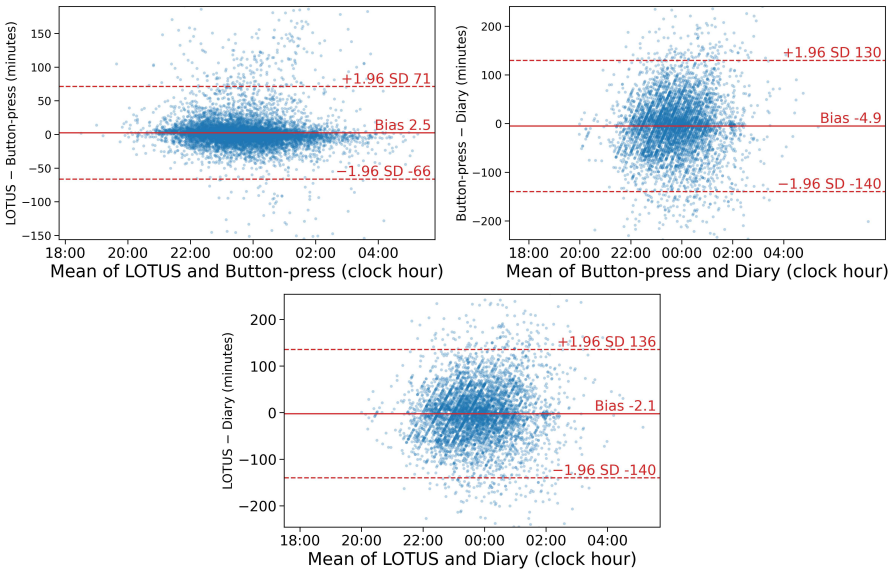


Fig. 10 Bland–Altman plots for button-press vs. *LOTUS model*, button-press vs. diary and *LOTUS model* vs. diary.

Sleep Onset Latency

To show that LOTUS actually predicts sleep intent rather than predicting sleep onset, we do a secondary analysis. We perform an ordinary least squares regression of the sleep-intent error between the diary and the *LOTUS model* predictions and compare it with the sleep-onset latency reported in the diary. This is shown in Figure 11. The two dashed lines represent perfect correlation between LOTUS and sleep intent in the horisontal line and between LOTUS and sleep onset in the diagonal line. The idea is that if the sleep-intent prediction increases with SOL, then the *LOTUS model* is actually predicting sleep onset. From the black ordinary least squares (OLS) line we see that LOTUS follows sleep-intent closer than the sleep onset and has captured behaviour that reflects actual sleep intent.

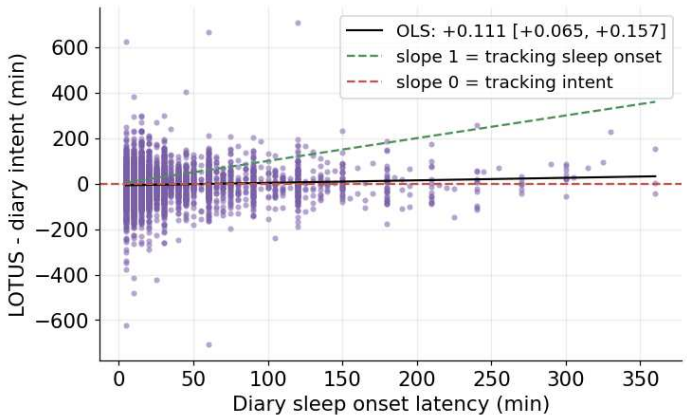


Fig. 11 Showing ordinary least squares between error in sleep intent between diary and LOTUS vs SOL from the diary. The dashed lines show the scenario if the *LOTUS model* predicted sleep onset perfectly and sleep intent perfectly.

7 Discussion

LOTUS locates the sleep-intent event to within a median of 6.7 minutes of the participant's own button press, and its agreement with sleep diaries is comparable to the agreement between button presses and diaries. Because no external ground truth exists, the appropriate reading is not that the model is "accurate" in absolute terms but that it stands in the same relation to self-report as the existing button-press measure does, while requiring no action from the patient. In what follows we interpret this result along three lines: what the ablations reveal about the structure of the task, what evidence indicates the model captures sleep-related behaviour rather than the button press itself, and where the model fails.

Ablations

The results of the ablation study presented in Table 3 reflect a series of inductive biases related to the structure of the sleep intent prediction task. The experiments show that the signal in the actigraphy is local in the time series, through the similar performance of the TCN. Adding positional encoding to this seems to help with accurately placing the event. Similarly, the superior performance of the distribution head shows that the soft labeling aids with learning.

Behaviour

Through our sampling rate experiment in Figure 8 and discussion of application in Section 6 we see that the LOTUS model behaves as desired. The model is able to maintain performance with considerably lower resolution in the data, meaning the behaviour learnt is not only derived from the button-press artefact. The OLS slope analysis further corroborates this by showing that the prediction reflects actual sleep intent rather than sleep onset. Compared against the current reference standards,

button-press and sleep diary, the predictions of the LOTUS model show good agreement. Overall, LOTUS shows promise as a method for predicting sleep-intent which can be highly valuable for the imputation of missing values. It can potentially replace button-press entirely allowing for a reduced patient burden. A more reliably measured sleep-intent also allows for more accurate measurement of treatment adherence during sleep restriction.

Robustness

LOTUS has shown promise in predicting sleep intent and through the analysis of the error distribution of the predictions in Figures 5 and 6 we get an increased understanding of the robustness of the model. We see the model' s limits, where the performance is considerably better in the more data rich, common bedtimes, but suffers in the extremes. This is reflected in the long tailed distribution of the errors. LOTUS outperforms three baselines on the mean score, largely due to it being able to handle the harder nights better. The median performance of LOTUS is similar to harnet10, but considerably outperforms on the mean. This could reflect that NorseMS is out-of-distribution for harnet10. Its encoder was pretrained on general-population data at 30 Hz and is used frozen, so its features may not capture the atypical nights that dominate the error tail in this cohort.

8 Limitations

The preprocessing needed to select a single nighttime button-press per night is fairly aggressive and may systematically discard the less typical nights, biasing both training and evaluation toward regular sleep patterns. Because the selection criterion requires a marked activity contrast around each press, retained nights are biased toward clearly visible transitions, which are also the easiest for the model to detect. Selecting the correct press is difficult in principle, and allowing the *LOTUS model* to predict multiple sleep bouts would remove one of the main sources of apparent error.

The training data span both arms of the trial, and sleep behaviour is expected to diverge between them over the course of the study. In the treatment group in particular, individual behaviour may change as the intervention takes effect, for example as sleep restriction shifts bedtimes. We do not model this drift.

Finally, the evaluation is anchored to button-presses throughout, and no external ground truth for sleep intent is available in this dataset. The diary comparison partly mitigates this, but conclusions about absolute accuracy should be read with that caveat in mind.

9 Conclusion

In this paper, we presented LOTUS, a temporal machine learning framework for locating transitions in locally stationary time series, applied to estimating the timing of sleep intent from wrist-worn wearable accelerometry alone. Trained on participants'

own button-presses in a trial of 513 adults with multiple sclerosis and insomnia, LOTUS located the reference event with a mean error of 14.0 minutes and a median of 6.7 minutes. Its mean error was lower than that of all three baselines, and its median error was comparable to that of the strongest baseline. The advantage of LOTUS lies mainly in its robustness on difficult nights. Agreement between LOTUS and the sleep diaries was comparable to agreement between the button-presses and the diaries. Downsampling analyses indicated that the *LOTUS model* responds to behaviour rather than to the button-press movement itself, which supports interpreting its output as sleep intent.

Since no ground truth for sleep intent exists in this setting, we interpret these results as LOTUS standing in the same relation to self-report as the existing button-press measure does, rather than as an improvement over it. Even under this framing, the method has direct practical value for wearable-based sleep monitoring. Nights without a button-press account for 33.1% of the dataset, and LOTUS can be used to passively impute the missing values without additional burden on the participant. This would allow sleep onset latency and adherence to sleep restriction to be estimated on nights that would otherwise be discarded, supporting both remote clinical monitoring of insomnia and the delivery of CBT-I, including digital CBT-I programmes.

Ethics statement

This work is a secondary analysis of data collected in the NorseMS randomized controlled trial [35]. NorseMS was approved by the Norwegian Directorate for Medical Products (CIV-NO-23-03-042653) and the Norwegian Ethics Committees for Clinical Trials on Medicinal Products and Medical Devices (reference 623308), and was registered at ClinicalTrials.gov (NCT06113666) the 27th October 2023. All participants gave online/written informed consent before study entry, including consent for the use of their de-identified data in secondary research. The trial was conducted in accordance with the Declaration of Helsinki. The present analysis used only de-identified data and required no additional data collection from participants.

Funding

This work was supported by NTNU Health (NTNU Helse), the strategic research area for health at the Norwegian University of Science and Technology (NTNU). The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] Daley M, Morin CM, LeBlanc M, Grégoire JP, Savard J. The economic burden of insomnia: direct and indirect costs for individuals with insomnia syndrome, insomnia symptoms, and good sleepers. Sleep. 2009;32(1):55–64. https://doi.org/10.1093/sleep/32.1.55.

[2] Benjafield AV, Sert Kuniyoshi FH, Malhotra A, Martin JL, Morin CM, Maurer LF, et al. Estimation of the global prevalence and burden of insomnia: a systematic literature review-based analysis. Sleep Medicine Reviews. 2025 Aug;82:102121. https://doi.org/10.1016/j.smrv.2025.102121.

[3] Hillman D, Mitchell S, Streatfeild J, Burns C, Bruck D, Pezzullo L. The economic cost of inadequate sleep. Sleep. 2018 Aug;41(8). https://doi.org/10.1093/sleep/zsy083.

[4] Freeman D, Sheaves B, Waite F, Harvey AG, Harrison PJ. Sleep disturbance and psychiatric disorders. The Lancet Psychiatry. 2020;7(7):628–637. https://doi.org/https://doi.org/10.1016/S2215-0366(20)30136-X.

[5] Lichstein KL, Durrence HH, Taylor DJ, Bush AJ, Riedel BW. Quantitative criteria for insomnia. Behaviour Research and Therapy. 2003;41(4):427–445. https://doi.org/https://doi.org/10.1016/S0005-7967(02)00023-2.

[6] Van Hees VT, Sabia S, Jones SE, Wood AR, Anderson KN, Kivimäki M, et al. Estimating sleep parameters using an accelerometer without sleep diary. Scientific Reports. 2018 Aug;8(1):12975. https://doi.org/10.1038/s41598-018-31266-z.

[7] Sundararajan K, Georgievska S, te Lindert BHW, Gehrman PR, Ramautar J, Mazzotti DR, et al. Sleep classification from wrist-worn accelerometer data using random forests. Scientific Reports. 2021;11:24. https://doi.org/10.1038/s41598-020-79217-x.

[8] Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-Sleep: resilient high-frequency sleep staging. npj Digital Medicine. 2021;4(1):72. https://doi.org/10.1038/s41746-021-00440-5.

[9] Zibrandtsen IC, Hernandez C, Ibsen JD, Kjaer TW. Event marker compliance in actigraphy. Journal of Sleep Research. 2020 Feb;29(1):e12933. https://doi.org/10.1111/jsr.12933.

[10] Lehrer HM, Yao Z, Krafty RT, Evans MA, Buysse DJ, Kravitz HM, et al. Comparing polysomnography, actigraphy, and sleep diary in the home environment: The Study of Women's Health Across the Nation (SWAN) Sleep Study. SLEEP Advances. 2022 Jan;3(1):zpac001. https://doi.org/10.1093/sleepadvances/zpac001.

[11] Spielman AJ, Saskin P, Thorpy MJ. Treatment of chronic insomnia by restriction of time in bed. Sleep. 1987;10(1):45–56. https://doi.org/10.1093/sleep/10.1.45.

[12] Matthews EE, Arnedt JT, McCarthy MS, Cuddihy LJ, Aloia MS. Adherence to cognitive behavioral therapy for insomnia: a systematic review. Sleep Medicine Reviews. 2013 dec;17(6):453–464. https://doi.org/10.1016/j.smrv.2013.01.001.

[13] Vincent N, Lewycky S, Finnegan H. Barriers to engagement in sleep restriction and stimulus control in chronic insomnia. Journal of Consulting and Clinical Psychology. 2008;76(5):820–828. https://doi.org/10.1037/0022-006X.76.5.820.

[14] Ogasawara E, Salles R, Porto F, Pacitti E. Event Detection in Time Series. 1st ed. Synthesis Lectures on Data Management. Cham: Springer Nature Switzerland; 2025.

[15] Killick R, Fearnhead P, Eckley IA. Optimal Detection of Changepoints With a Linear Computational Cost. Journal of the American Statistical Association. 2012 Oct;107(500):1590–1598. https://doi.org/10.1080/01621459.2012.737745.

[16] Yoshizawa G. Bayesian Online Change Point Detection for Baseline Shifts. Statistics, Optimization & Information Computing. 2020;9(1):1–16. https://doi.org/10.19139/soic-2310-5070-1072.

[17] Yao R, Lin G, Shi Q, Ranasinghe DC. Efficient dense labelling of human activity sequences from wearables using fully convolutional networks. Pattern Recognition. 2018;78:252–266. https://doi.org/https://doi.org/10.1016/j.patcog.2017.12.024.

[18] Bai S, Kolter JZ, Koltun V.: An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. Available from: https://arxiv.org/abs/1803.01271.

[19] Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In: The Eleventh International Conference on Learning Representations; 2023. Available from: https://openreview.net/forum?id=Jbdc0vTOcol.

[20] Tompson J, Jain A, LeCun Y, Bregler C. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In: Advances in Neural Information Processing Systems (NIPS). vol. 27; 2014. Available from: https://neurips.cc.

[21] Saint-Maurice PF, Freeman JR, Russ D, Almeida JS, Shams-White MM, Patel S, et al. Associations between Actigraphy-Measured Sleep Duration, Continuity, and Timing with Mortality in the UK Biobank. Sleep. 2024;47(3):zsad312.

[22] Luo W, Scharf MT, Androulakis IP. Aging and Activity Patterns: Actigraphy Evidence from NHANES Studies. Frontiers in Systems Biology. 2025;5:1632110.

[23] Åsvold BO, Langhammer A, Rehn TA, Kjelvik G, Grøntvedt TV, Sørgjerd EP, et al. Cohort Profile Update: The HUNT Study, Norway. International Journal of Epidemiology. 2023;52(1).

[24] Carpenter JS, Crouse JJ, Park SH, Shin M, Tonini E, Guo W, et al. Actigraphy-derived circadian rhythms, sleep-wake patterns, and physical activity across clinical stages and pathophysiological subgroups in young people presenting for mental health care. Journal of Psychiatric Research. 2025 Jun;186:396–406. https://doi.org/10.1016/j.jpsychires.2025.03.003.

[25] Kim H, Kim D, Oh J. Automation of classification of sleep stages and estimation of sleep efficiency using actigraphy. Frontiers in Public Health. 2023 Jan;10:1092222. https://doi.org/10.3389/fpubh.2022.1092222.

[26] Smith MT, McCrae CS, Cheung J, Martin JL, Harrod CG, Heald JL, et al. Use of Actigraphy for the Evaluation of Sleep Disorders and Circadian Rhythm Sleep-Wake Disorders: An American Academy of Sleep Medicine Systematic Review, Meta-Analysis, and GRADE Assessment. Journal of Clinical Sleep Medicine. 2018 Jul;14(7):1209–1230. https://doi.org/10.5664/jcsm.7228.

[27] Weitz M, Syed S, Hopstock LA, Morseth B, Henriksen A, Horsch A.: Automatic time in bed detection from hip-worn accelerometers for large epidemiological studies: The Tromsø Study. Health Informatics. Available from: http://medrxiv.org/lookup/doi/10.1101/2022.03.07.22270992.

[28] Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic Sleep/Wake Identification From Wrist Activity. Sleep. 1992 09;15(5):461–469. https://doi.org/10.1093/sleep/15.5.461. https://academic.oup.com/sleep/article-pdf/15/5/461/13688763/sleep-15-5-461.pdf.

[29] Sadeh A, Sharkey M, Carskadon MA. Activity-Based Sleep-Wake Identification: An Empirical Test of Methodological Issues. Sleep. 1994 05;17(3):201–207. https://doi.org/10.1093/sleep/17.3.201. https://academic.oup.com/sleep/article-pdf/17/3/201/13659930/sleep-17-3-201.pdf.

[30] Logacjov A, Skarpsno E, Kongsvold A, Bach K, Mork PJ. A Machine Learning Model for Predicting Sleep and Wakefulness Based on Accelerometry, Skin Temperature and Contextual Information. Nature and Science of Sleep. 2024 Jun;Volume 16:699–710. https://doi.org/10.2147/NSS.S452799.

[31] Yuan H, Plekhanova T, Walmsley R, Reynolds AC, Maddison KJ, Bucan M, et al. Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality. npj Digital Medicine. 2024 Jun;7(1):86. https://doi.org/10.1038/s41746-024-01065-0.

[32] Carney CE, Buysse DJ, Ancoli-Israel S, Edinger JD, Krystal AD, Lichstein KL, et al. The Consensus Sleep Diary: Standardizing Prospective Sleep Self-Monitoring. Sleep. 2012 Feb;35(2):287–302. https://doi.org/10.5665/sleep.1642.

[33] Withrow D, Roth T, Koshorek G, Roehrs T. Relation between ambulatory actigraphy and laboratory polysomnography in insomnia practice and research.

Journal of Sleep Research. 2019 Aug;28(4):e12854. https://doi.org/10.1111/jsr.12854.

[34] Fekedulegn D, Andrew ME, Shi M, Violanti JM, Knox S, Innes KE. Actigraphy-Based Assessment of Sleep Parameters. Annals of Work Exposures and Health. 2020 Apr;64(4):350–367. https://doi.org/10.1093/annweh/wxaa007.

[35] Saksvik SB, Rønhovde LM, Ytrehus-Lynum K, Langsrud K, Olsen A, Das Nair R, et al. Digital cognitive behavioral therapy for insomnia compared with digital patient education about insomnia in people with multiple sclerosis in Norway: study protocol for a randomized controlled trial. SLEEP Advances. 2025 Oct;6(4):zpaf075. https://doi.org/10.1093/sleepadvances/zpaf075.

[36] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for Hyper-Parameter Optimization. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K, editors. Advances in Neural Information Processing Systems. vol. 24. Curran Associates, Inc.; 2011. Available from: https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

[37] Yuan H, Chan S, Creagh AP, Tong C, Acquah A, Clifton DA, et al. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. NPJ digital medicine. 2024;7(1):91.

[38] Santorelli A, Perez F, de Wael RV, Rupprecht F, d'Antonio Bertagnolli JV, Franco A, et al. wristpy: Fast, User-Friendly Python Processing of Wrist-worn Accelerometer Data. Journal of Open Source Software. 2025;10(114):8637. https://doi.org/10.21105/joss.08637.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- appendixintent.pdf