Sleep Health: Journal of the National Sleep Foundation xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Sleep Health: Journal of the National Sleep Foundation

journal homepage: www.sleephealthjournal.org

Performance evaluation of a machine learning-based methodology using dynamical features to detect nonwear intervals in actigraphy data in a free-living setting

Jyotirmoy Nirupam Das [a,*,1], Linying Ji, PhD [b,2], Yuqi Shen [c,3], Soundar Kumara, PhD [a,4], Orfeu M. Buxton, PhD [d,5], Sy-Miin Chow, PhD [e,6]

[a] Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, Pennsylvania, USA
[b] Department of Psychology, Montana State University, Bozeman, Montana, USA
[c] Biobehavioral Health Department, The Pennsylvania State University, State College, Pennsylvania, USA
[d] Department of Biobehavioral Health Department, The Pennsylvania State University, University Park, Pennsylvania, USA
[e] Department of Human and Development and Family Studies, Pennsylvania State University, University Park, Pennsylvania, USA

ARTICLE INFO

Article history:
Received 2 June 2024
Received in revised form 28 August 2024
Accepted 6 October 2024

Keywords:
Nonwear detection
Machine learning
Actigraphy
Sleep

ABSTRACT

Goal and aims: One challenge using wearable sensors is nonwear time. Without a nonwear (e.g., capacitive) sensor, actigraphy data quality can be biased by subjective determinations confounding sleep/wake classification. We developed and evaluated a machine learning algorithm supplemented by dynamic features to discern wear/nonwear episodes.
Focus technology: Actigraphy data from wrist actigraph (Spectrum, Philips-Respironics).
Reference technology: The built-in nonwear sensor as "ground truth" to classify nonwear periods using other data, mimicking features of Actiwatch 2.
Sample: Data were collected over 1 week from employed adults (n = 853).
Design: Extreme gradient boosting (XGBoost), a tree-based classifier algorithm, was used to classify wear/nonwear, supplemented by dynamic features calculated over various time windows.
Core analytics: The performance of the proposed algorithm was tested over 30-second epochs.
Additional analytics and exploratory analyses: Evaluation of the SHapley Additive exPlanations (SHAP) values to find the effectiveness of the dynamic features.
Core outcomes: The XGBoost classifier yielded substantial improvements in balanced accuracy, sensitivity, and specificity, including dynamic features and comparison to default actiwatch classification algorithms.
Important supplemental outcomes: The proposed classifier effectively distinguished between valid and invalid days, and the duration of contiguous periods of nonwear correctly identified.
Core conclusion: Our findings highlight the potential of XGBoost using dynamic features of varying activity levels across the time series to provide insights on wear/nonwear classification using a large dataset. The methodology provides an alternative to laborious manual benchmarking of the data for similar devices that do not have a nonwear sensor.

Introduction

Data from wrist motion sensors for classifying sleep-wake patterns have grown exponentially due to ease of implementation and cost-effectiveness.[1] Wearable-derived digital biomarkers generally provide activity (movement) and light levels or other data to capture passive data over long periods[2] in free-living participants nonobtrusively.

One of the crucial challenges of studies using wearable sensors is nonwear time, a form of missing data potentially introducing bias. When not worn, the devices may still register low activity levels during

* Corresponding author: Jyotirmoy Nirupam Das, Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, Pennsylvania, USA.
E-mail address: jfd5895@psu.edu (J.N. Das).
[1] 0000-0003-4068-3988
[2] 0000-0003-1908-3718
[3] 0009-0004-3047-2209
[4] 0000-0002-7941-8818
[5] 0000-0001-5057-633X
[6] 0000-0003-1938-027X

https://doi.org/10.1016/j.sleh.2024.10.003

nonwear, which may be confounded with or indistinguishable from sleep or sedentary periods based on visual inspection. In devices without nonwear sensors, manually marking nonwear episodes is time-consuming and could introduce further bias. Common automated nonwear detection algorithms in the current literature use tri-axial accelerometers and other signals, such as skin temperature or respiration rate, to detect nonwear intervals accurately. Another approach is setting a time interval with low activity counts for nonwear detection. For example, using data from a cohort of children aged 11 years, 10 different nonwear criteria were evaluated; it was concluded that time intervals of 45-60 minutes with zero activity counts were sufficient to detect nonwear intervals,[3] consistent with similar time interval thresholds for different cohorts and accelerometers in adults.[4,5] Lok and Zeitzer[5] used a temporal threshold to distinguish wear and nonwear periods after manually scoring the periods as wear and nonwear.

Machine learning techniques have been used in nonwear detection studies, such as a logistic regression model to distinguish nonwear patterns during sleep and wake from 23 participants from the NHANES study.[6] A similar regression tree-based method classified wear/nonwear periods using movement and temperature sensor data[7] from different time windows as an input feature. Temperature changes increase their model's performance. Some studies generally set a threshold over activity levels or accelerometry sensors over an interval such as 30, 60, or 90 minutes.[8,9] Manual threshold constraints can be arbitrary and exclude the possibility of detecting shorter nonwear episodes.

Other studies have identified additional features from their data to improve the performance of their algorithms. A study using Random Forest for a sleep and nonwear classification study implemented in a clinical setting with 36 features extracted from triaxial accelerometer data, including 12 statistical measures (e.g., mean, median, standard deviation, t increased model accuracy).[10] Additional opportunities to maximize features not used in these studies include the stability of the time series, shifts in the variable level, and variability across multiple time resolutions.

In this paper, we propose and evaluate, consistent with best practice for the evaluation of sleep algorithms,[11] a machine learning-based methodology to detect nonwear time epochs (30-second) from actigraphy data. An eXtreme Gradient Boosting (XGBoost) model was applied to a dataset containing 853 participants, which led to a generalized machine learning-based approach for nonwear classification, including dynamic features, efficient tuning of hyperparameters, and visualization of results to facilitate interpretation of critical dynamic and time scale information. Due to the minimal data requirements, the method developed is versatile enough to be applied to a wide range of sleep wearable studies. Our methodology also addresses the limitations of thresholding used in earlier work, in which a hard threshold has to be chosen in terms of physical activity level and/or time interval for inactiveness. The optimal values for such fixed thresholds vary greatly from dataset to dataset, and do not capture the complexity of a dataset with varied nonwear intervals.

Methods

Sample

Data were collected from employed adults from two different industries, as previously described,[12-14] including an extended care provider (nursing home direct patient care workers; n = 545 individuals, 92% female, $Mean_{age}$: 38 years), and an information technology company (n = 308 participants, 44% female, $Mean_{age}$: 46 years). We excluded other participants from the larger dataset with activity counts during off-wrist periods over-written by an early, defective version of the manufacturer's software and set to "n/a," so "truth" was unavailable. Each participant was instructed to wear a wrist actigraph (Spectrum, Philips-Respironics, Murrysville, PA) on their nondominant wrist. The actiwatch provided wrist actigraphy measurements, including activity levels, white light, red light, blue light, and green light collected in 30-second epochs for a week. Nonwear periods were identified with the built-in wrist capacitive sensor, which, if the device was worn correctly, measured skin contact/proximity or its absence. The capacitive sensor could provide, in rare circumstances, both false on-wrist and false off-wrist values in stereotypical and visually identifiable ways, such as nearing battery failure, visually identified by two trained scorers. Scorers visually reviewed the entire recording for each file to better understand an individual's activity pattern and intensity levels across days and examined points of sudden change in activity levels and light levels in the recordings. After each scorer inspected the file independently, inter-rater reliability was assessed, and any discrepancies were resolved among the scoring team. The total number of off-wrist was 7.4% of epochs, and on-wrist was 92.5%, leading to an imbalanced dataset.

Machine learning approach

Data preparation. In this paper, we explore the use of a tree-based classifier known as XGBoost to classify wear/nonwear epochs.[15] Wear is labeled as 1, and nonwear has been labeled as 0 from the Spectrum data. Mimicking the data of the closely related and commonly used Actiwatch 2 model (which has no nonwear sensor), the two key inputs taken from the actiwatch data are the white light and the activity levels. A correlation study between the different light signals confirmed that the white light correlated at least 0.9 with all other light signals; hence, only the white light level was taken as input. To leverage information concerning short- and long-term time dependencies, dynamical features were extracted using the tsfeatures package in R.[16,17] Dynamical features created using the tsfeature package are spectral entropy, maximum level shift, maximum variance shift, maximum Kulback-Leiber divergence, hurst, spike, heterogeneity, the standard deviation of the first derivative of the time series, and activity lags. Spectral entropy, hurst, heterogeneity, and standard deviation of the first derivative of the time series capture if the time series is random/noisy or if it is stable and has a constant trend. Maximum level shifts and spikes keep track of any sudden changes in the time series. Maximum Kulback-Leiber divergence and maximum variance level shift detect any changes in data distribution in consecutive time windows. Dynamical features were created to provide statistical information to the XGBoost model and improve its ability to discern trends or patterns in the time series. The features have been made on the activity levels with a window size of 20 minutes and 1 hour. The 20-minute window captures shorter naps, and the 1-hour window size captures the average nap duration.[18] Furthermore, activity lags were created over 5, 10, 15, and 20 minutes. Next, we processed the time series to generate several temporal features such as minute, hour, quarter, day of the week, and time of the day. This information helps to interpret the timestamp in a useful representation, such as if the time of the day is night or the hour is early morning, then the person is more likely to be asleep. In addition, for each participant, the temporal features were converted into sinusoidal curves to capture prominent periodicities in the data. For each participant, employer information was added as a binary variable (1: extended care and 0: information technology). After processing each participant's data, the data were combined to form a larger dataset. This dataset was split into a 70:30 ratio to create the training and test sets. The random split was formed such that the training set has 597 unique participants and the test set has 256 unique participants.

Hyperparameter tuning and model training. The parameters for the XGBoost model were determined by using hyperparameter optimization and cross-validation. The Hyperopt Python library[19] was used to efficiently explore a range of hyperparameter values for the XGBoost model (see Fig. 1). The parameters were optimized on the average balanced accuracy across five folds of cross-validation

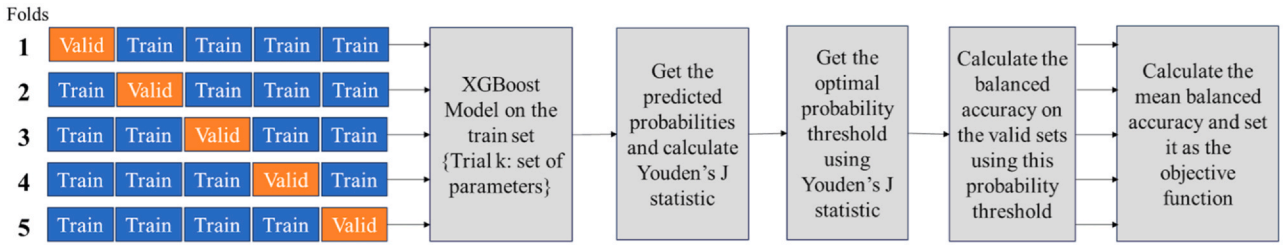

Fig. 1. Hyperparameter tuning and k-fold cross-validation to optimize XGBoost model parameters

Table 1
Parameter search space

Parameter	Range
Maximum bin	10-500
Colsample_bytree	0.01-1
Learning rate	0.00001-1
Maximum depth	1-40
Reg_alpha	0-10
N_estimators	0-400
gamma	0-15
eta	0.01-1
Reg_lambda	0-10
Rate drop	0-1
Scale pos weight	0.05-1

training data. The initial ranges of hyperparameter values were expanded and re-evaluated when any hyperparameter value was observed near the search space's boundaries. The range of the search space of the parameters is given in Table 1. The final optimized hyperparameter values and parameter descriptions are summarized in Table 2. We used the histogram tree method, which buckets continuous features into bins to yield approximate tree-splitting solutions to enhance computational speed.[20] To further speed up the algorithm gbtree booster and GPU Quadro K4000 were used. Here we tuned the scale_pos_weight as part of the hyperparameter values to be optimized via Hyperopt to allow for heavier weighing of the rare instances of nonwear under imbalanced data (i.e., with the inflated presence of wear data) in the loss function.[21] The two regularization parameters helped reduce overfitting. From Fig. 1, we can see that after fitting the XGBoost model to the training set, we used Youden's J statistic[22] to select the probability threshold for classifying data as nonwear under imbalanced data optimal threshold was found and computed the final average balanced accuracy across the validation folds. The entire modeling process is shown in Fig. 1. Once we got the classification results from the XGBoost model based on the optimum parameters, we manually changed the probability threshold to get the final performance metrics on the training dataset. We manually changed the threshold instead of relying on Youden's J statistic because Yoden's J statistic increased the specificity at the expense of negative predictive value (NPV). The effects of using different probability thresholds for nonwear classification on the performance metrics are shown in Fig. 2. We observe that at probability threshold 0.55, there is a balance between the specificity and NPV value. Increasing or decreasing the probability threshold would lead to unequal tradeoffs in the performance metrics; hence, 0.55 was set as the optimum probability threshold for our dataset.

To further improve our model, we employed a temporal smoothing function to address abrupt (and rare/unlikely 1-epoch) changes in the output at an epoch level. This smoothing function checks neighborhood labels and assigns the label that has the majority in that time window. For example, in a window size of 3 epochs, if there are two nonwear labels and one wear label, all epochs are labeled as nonwear. A range of window sizes from 3 to 20 epochs were used to find the optimal window size. The optimal window size was found to be 3 for the training dataset. This window size gave the maximum specificity; as we increased the window size, specificity decreased.

Classification performance evaluation. The main aim of this paper was to distinguish nonwear epochs in actigraphy data. The XGBoost model provided probabilities at an epoch level and we determined the optimal probability threshold level based on different classification metrics. The classification metrics studied in the paper are sensitivity, specificity, balanced accuracy, F1 score, positive predictive value, and NPV. Area under the receiver operating characteristic curve was computed for the initial set of results (Table 3). All other results are based on the optimal probability threshold.

To further evaluate our methodology, we compared our result with a published methodology addressing the same general problem,[5] which showed that 95% of the inactive episodes are of length shorter than 20 minutes in an Actiwatch 2 with a resolution of data of 60 seconds. To create these nonwear/inactive episodes automatically, they used R studio to label any consecutive epochs with

Table 2
Optimal parameter values

Parameters	Value	Parameter purpose
tree_method	Hist	Can handle larger datasets and has been optimized for speed
booster	gbtree	The gbtree method is more suitable to fit nonlinear boundaries
normalize_type	Forest	Stabilizes learning
device	Cuda	The model can be trained on GPU
sampling_method	uniform	No bias in selecting the samples for each tree
colsample_bytree	0.2674	Low value reduces the risk of overfitting
gamma	12.7667	High value reduces overfitting by reducing tree complexity
learning_rate	0.0083	Lower learning rate helps in getting optimal loss
max_bin	489	Buckets continuous features to make a detailed yet computationally efficient representation of features
max_depth	19	A deeper tree can better learn the hidden patterns
n_estimators	91	Denotes the number of trees and a higher number can avoid underfitting
rate_drop	0.4`	Makes the model more robust
reg_alpha	5.44	Makes the weights sparse
reg_lambda	1.51	Smoothens the impact of weights
scale_pos_weight	0.157	Gives more weight for the negative class in the loss function

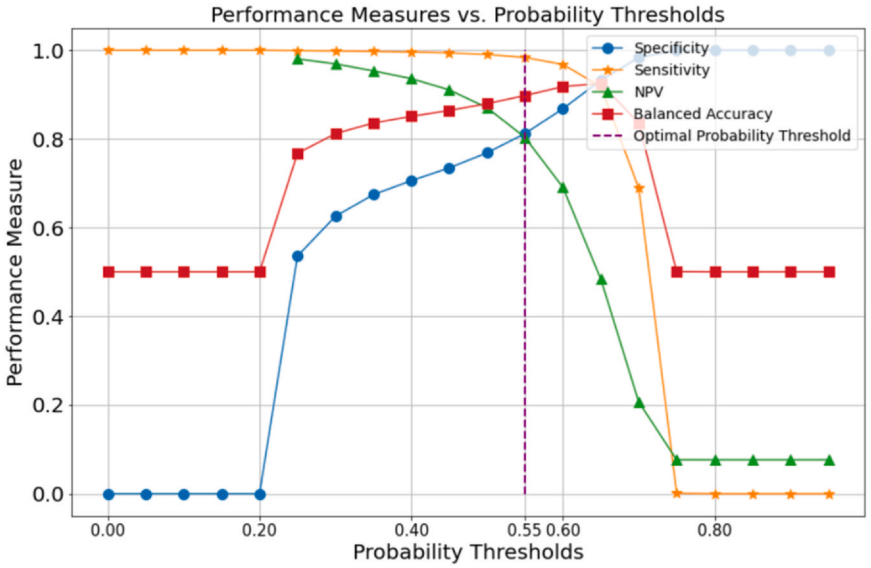

Fig. 2. Performance measures associated with different probability thresholds for nonwear classification. The optimal probability threshold (0.55) was selected since, at this probability, the best tradeoff between NPV and specificity was achieved. NPV, negative predictive value

Table 3
Classification metrics on the training and test set

	AUC	Sensitivity	Specificity	Balanced accuracy	F1 score	PPV	NPV
Training dataset	0.981	0.991	0.771	0.881	0.986	0.981	0.879
Test dataset	0.937	0.990	0.760	0.875	0.985	0.981	0.855

Abbreviations: AUC, area under the receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value.

zero activity counts as an episode of nonwear. The episodes were then visually labeled as wear and nonwear. They then generated predictions on the episodes by calculating the length of each episode and classifying it as nonwear if the episode had a length of more than 20 minutes. The optimal time threshold was computed from the receiver operating curve. Following their methodology, we derived an optimal threshold of 1 minute, which gave the highest accuracy in detecting nonwear episodes in our dataset. They used activity levels to classify wear/nonwear, and we followed similar approaches, adding dynamic features and white light to predict nonwear and wear epochs.

Results

Core analytics and main outcome variables

Epoch-by-epoch performance

In this paper, using data from 853 participants, 597 in the training dataset, and 256 in the test dataset. Each participant contributed ~20,000 epochs. Sensitivity refers to correctly predicting wear epochs whereas specificity gives the performance of the model on the nonwear epochs. The results from the optimum XGBoost model (Table 3) show that the balanced accuracy in the training is 0.881 and 0.876 in the test set. Due to the class imbalance in the data, we have a high sensitivity of 0.991 as opposed to a specificity of 0.771 in the training set. To balance this, we reclassify the probabilities generated by the XGBoost model using 0.55 as the new threshold and add the temporal smoothing function. The updated results are shown in Table 4, here, we can see that our specificity rises to 0.822 in the training data. The specificity is lower in the test at 0.791 but is still comparable to the training accuracy. Thus, we do not overfit or underfit our model. The sensitivity decreases from 0.991 to 0.985 in the training set but is still high enough to be determinantal in classifying wear time points. The NPV comes down when we use the probability threshold, but that is the best tradeoff we found in the model against increasing the specificity. The results of the comparison study can be seen in Table 4. We can see that our proposed model outperforms the temporal threshold method on every classification metric. This evaluation supports the claim that machine learning algorithms such as XGBoost can be an automation tool to classify wear and nonwear.

Table 4
Comparison study on the improved XGBoost methodology and the threshold method

	XGBoost with probability threshold		Threshold method	
	Train set	Test set	Train set	Test set
Sensitivity	0.985	0.983	0.957	0.947
Specificity	0.822	0.791	0.104	0.117
Balanced accuracy	0.904	0.887	0.526	0.532
F1 score	0.985	0.983	0.965	0.967
PPV	0.985	0.983	0.984	0.989
NPV	0.824	0.783	0.032	0.025

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

To validate the model further, we matched the invalid days in the dataset against the invalid days predicted by our algorithm. A day is defined as invalid if the person does not wear the actiwatch for more than 4 hours. We label invalid days as 1 and valid days as 0. From Table 5, we can calculate that we are predicting valid days with an accuracy of 0.997 and invalid days with 0.921 accuracy in the training dataset. We get comparable accuracy in the test set, as can be seen in Table 5. Table 6 shows a comprehensive list of classification metrics for predicting invalid and valid days. Most of the metrics are above 0.9; hence, our proposed methodology does not lead to a mismatch between valid and invalid days.

Table 5
Prediction of invalid days on the train and test set

	Train set		Test set	
	Predicted invalid	Predicted valid	Predicted invalid	Predicted valid
True invalid	389	33	156	18
True valid	8	3457	1	1505

Additional analytics and exploratory analyses

To visualize the predictive roles of the dynamic features, we plotted the SHAP (SHapley Additive exPlanations)[23] values of the 20 most important features (Fig. 3) to identify how a change in each observed value of a feature would change the prediction of the probability of wearing the device in the model. A high standard deviation in the first derivative of the activity levels in a time window of 1 hour leads to a higher probability of wear. That is, higher growth in activity levels was associated with an increase in the probability of wearing the device. In contrast, lower maximum level shift for that hour (denoted as max_level_shift_hr), maximum variance shift for that hour (denoted as max_var_shift_hr), and number of sharp spikes for that hour (spike_hr) were observed to play multiple roles: they were associated in higher probability for wear in some instances, but also higher probability for nonwear in other instances. In other words, while prolonged instances with little spikes, level, and variance shifts in an hourly window might be indicative of nonwear, these dynamical features might also capture episodes of sustained high physical activity. The abbreviation "short" in the feature labels in Fig. 3 corresponds to dynamical features created on the activity levels over a time window of 20 minutes. Hence, we have two-time windows of 20 minutes and 1 hour for each dynamic feature.

The standard deviation of the activity levels, spikes, and maximum variance shift over the shorter time window also contributed to the accuracy of the model in a similar pattern as the same features over a 1-hour time window. Both time windows work in tandem to give short and long-term information to the model. The shorter time window reflected changes in the recent history of the time series, while the hourly time window captured the general trend of the feature. A high Hurst value suggested high persistence in an individual activity level, the most likely occurrence of which might be during individuals' periods of nonwear. In contrast, a low Hurst value, suggesting low persistence, reflected sustained continuous changes in the process of interest, namely ongoing fluctuations in the individual's physical activity during wear. For the Hurst exponent feature, the SHAP value plot gives similar interpretations for both time windows. Volatile activity levels during physical activities are indicated by high entropy levels over the hour (entorpy_hr), whereas low entropy levels suggested no change in activity levels, which in turn, might be suggestive of nonwear. The maximum variance shift, maximum level shift, and spikes capture the sudden changes in the activity levels, in contrast, the Hurst and entropy indicate the flow of the activity levels. The SHAP value plot in Fig. 3 shows that the light signal, when compared to the dynamic features, was not important in facilitating nonwear classification. However, we kept the light signal as a feature since it did not significantly affect computational time. The above analysis demonstrates the relevance and insights dynamic features provide to the XGBoost model.

To further analyze the results in terms of nonwear intervals of different durations, we created epoch intervals with different numbers of continuous nonwear epochs. An epoch interval length is defined as the length of the continuous set of nonwear epochs predicted by our methodology. Our nonwear epoch interval length varied from 2 to 19,970 epochs, corresponding to nonwear intervals from 1 minute to 7 days for 30-second epochs. Next, we grouped the different epoch intervals based on their length with a bin size of 20 epochs (10 minutes) as can be seen in Fig. 4, depicting the mean and median accuracy of the predicted nonwear epoch intervals against the actual number of nonwear epochs present in them. From the plot, we can observe that the mean accuracy gradually rises to 1 as the epoch interval length increases to 530 epochs. So for higher epoch interval length, the algorithm is better able to identify a set of continuous nonwear episodes. The median values approach 1 for epoch interval length of 110 epochs (55 minutes) to 530 epochs (nearly 4.5 hours). This suggests that the proposed methodology can discern between long naps or long periods of inactivity against nonwear epochs. Hence this methodology can be used in sleep studies where the main aim in preprocessing data is to identify valid days or long periods of nonwear in the dataset. To better illustrate the methodology's performance on longer time windows, two data samples of length 1 week each have been shown in Fig. 5. The black bands represent nonwear episodes, and the white bands show wear episodes. For each sample data, the ground truth from the capacitive sensor and the prediction from the methodology are depicted. The two bands from each sample are not identical, but the nonwear episodes on a longer horizon are effectively captured by the proposed methodology.

Discussion

In the current study, we addressed the problem of determining nonwear epochs in a sleep wearable (Actiwatch Spectrum) relevant to data from a different model of the same device (Actiwatch 2.0) that does not include a nonwear sensor feature, a commonly used device/data type in prior sleep studies. Addressing nonwear is particularly important because nonwear intervals near or during actual time in bed periods may be classified as sleep, leading to nonrandom bias in characterizing sleep and wake patterns. We used a large dataset using the Actiwatch Spectrum device, similar to the Actiwatch 2.0 but including a nonwear sensor that was used as the "ground truth" for nonwear. We developed and evaluated an XGBoost model to classify wear and nonwear epochs from an actiwatch worn by a participant in a free-living setting. We extracted different dynamical features from the activity level data from the actiwatch and also expanded the time stamps into sinusoidal curves to better utilize the temporal component of the data. This expanded dataset with 37 features was then used as input into the XGBoost model. The optimum parameters for the model were selected by running a cross-validation and hyperparameter tuning algorithm on the XGBoost model. Our systematic evaluation revealed classification performance in classifying epoch level nonwear in the test dataset of area under the receiver operating characteristic curve: 0.937, Sensitivity: 0.760, Specificity: 0.760, and Balanced Accuracy: 0.875,

Table 6
Classification metrics for predicting invalid days

	Sensitivity	Specificity	Balanced accuracy	F1 score	PPV	NPV
Training dataset	0.921	0.997	0.959	0.949	0.979	0.990
Test dataset	0.896	0.999	0.944	0.942	0.993	0.988

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

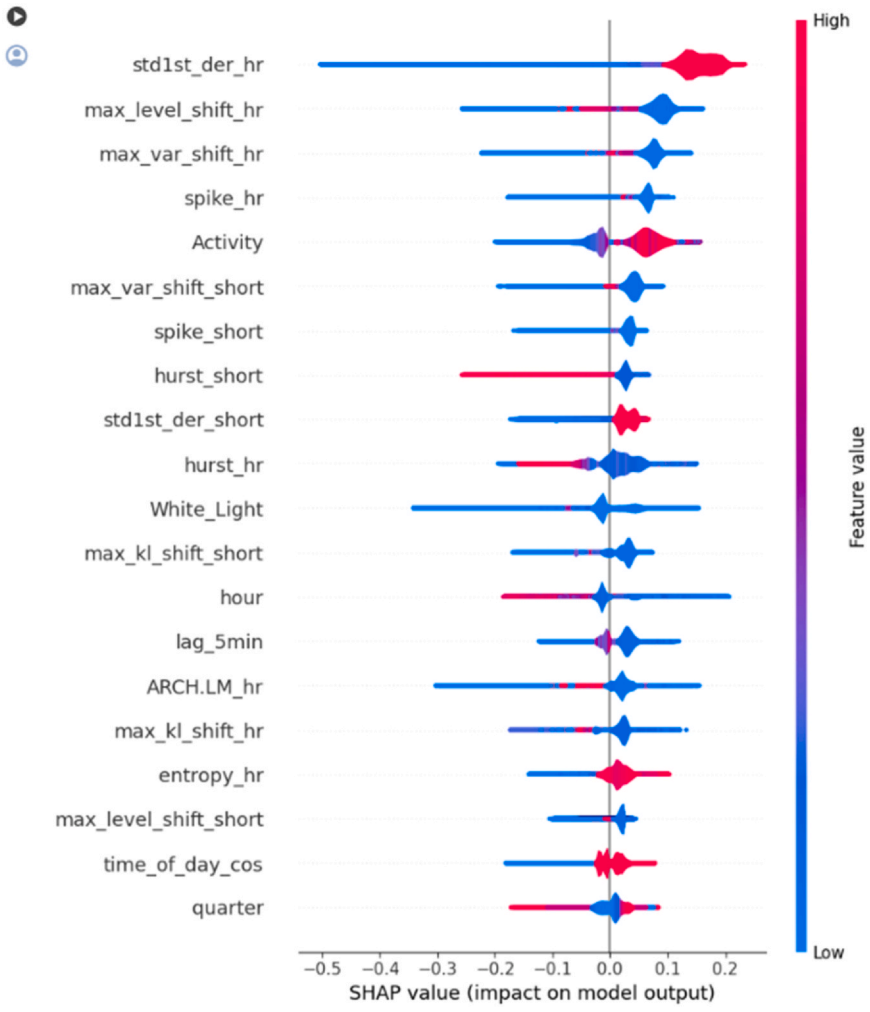

Fig. 3. The SHAP value plot is arranged according to the feature importance. Also, the SHAP values show the combined impact of each feature on the probability of wear and nonwear. Higher absolute SHAP values, defined as feature importance, conveying greater changes – either increases (in red) or decreases (in blue) in wear classification probability, are arranged in decreasing order from top to bottom according to their absolute SHAP value. SHAP, SHapley Additive exPlanations

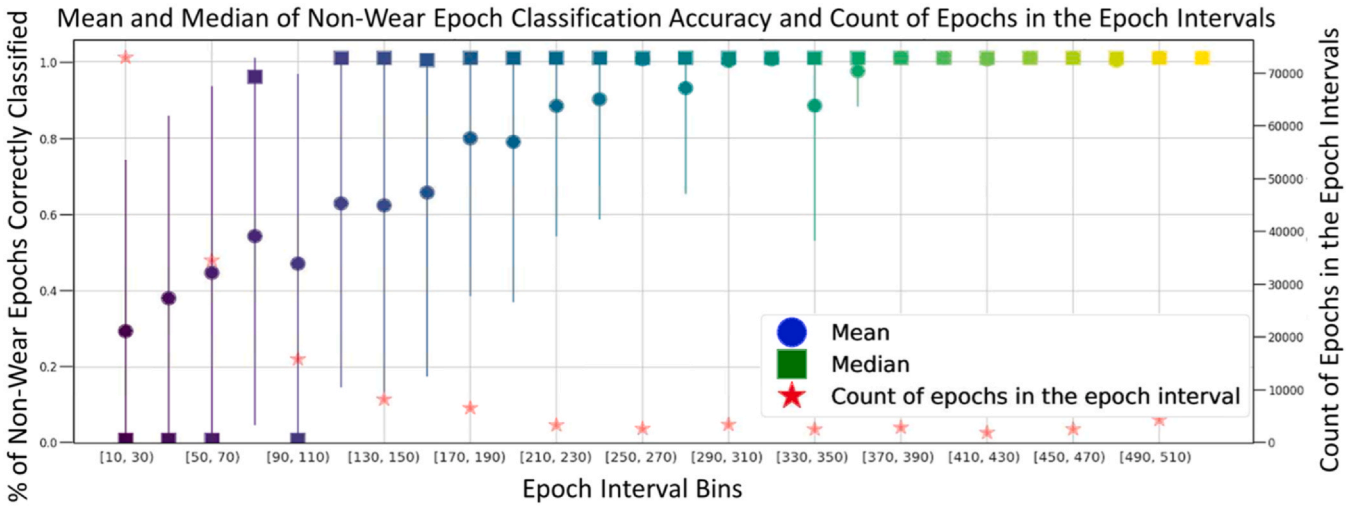

Fig. 4. Each epoch interval represents a nonwear episode of certain time period. Here [10,30) represents nonwear episodes ranging from 5 to 15 minutes. The Mean and Median provide the model classification accuracy on these epoch interval bins. On the 2nd y-axis shows the number of epochs in each epoch intervals

among other favorable evaluation metrics (Table 3). The results were further optimized by the application of manual probability thresholding and the temporal smoothing function, leading to a 0.887 balanced accuracy score in the test dataset. In comparison to the temporal thresholding methodology described in Lok and Zeitzer[5] our methodology performed better in all performance metrics. Since

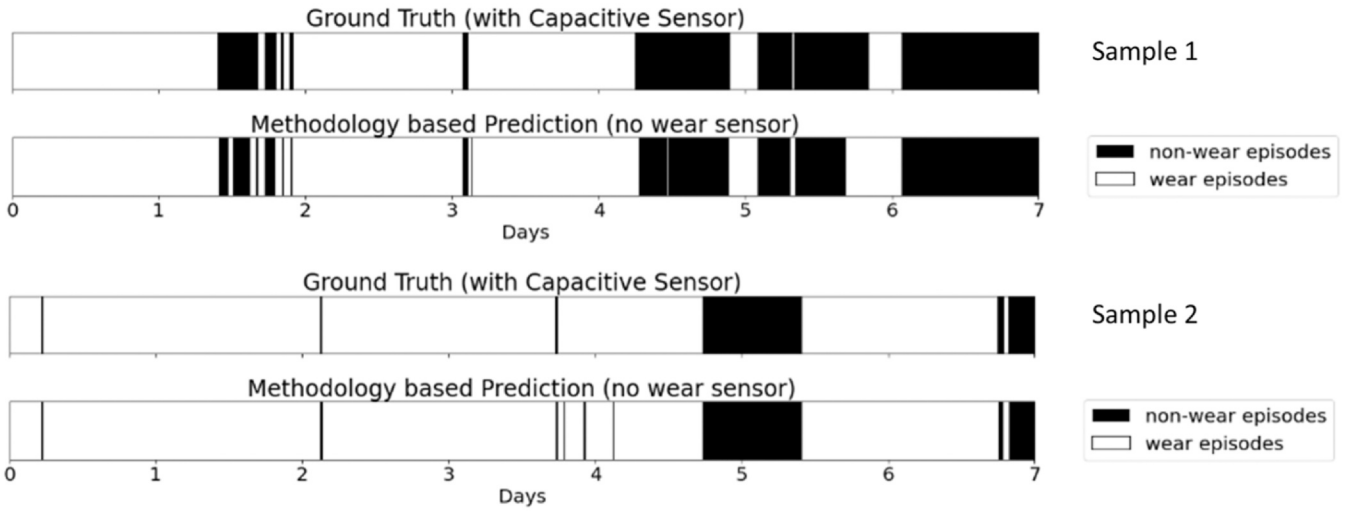

Fig. 5. Two data samples over 1 week showcase the comparison between methodology-based prediction and ground truth

we achieved high balanced accuracy on the test set, the proposed model and dynamic features can be implemented to detect nonwear on other datasets. As a practical outcome, if an "invalid day" is deemed to require at least 24 hours of wear time or < 4 hours of nonwear, we calculated that we predicted valid days with an accuracy of 0.997 and invalid days with an accuracy of 0.921. Hence, in studies where the determination of valid days plays a major role, our methodology can be used to identify the days efficiently. Importantly, we further evaluated the threshold of the duration of a nonwear period of consecutive nonwear epochs where classification approached 100% correct at a contiguous duration of ~ 1 hour. This evaluation supports the use of this nonwear detection algorithm for rigorous detection of sleep in the absence of a nonwear sensor on a wearable.

Limitations and future perspectives

One of the limitations of our study can be observed in Fig. 4, where we can see that the median accuracy is near zero for epoch lengths of less than 110 epochs (55 minutes). Brief periods of time when the activity level is low, leading to misclassification of nonwear. Due to this, the proposed methodology has an NPV of 0.783. The median accuracy might be low due to short naps, which might be confused with nonwear epochs due to low activity levels. In addition to this short duration, epoch lengths might be inaccurately classified due to a lack of knowledge of the recent history of the time series. Alternative machine learning variations, including long short-term memory networks and other recurrent neural network variations,[24] and a wavelet scattering-based model[25] might be an approach for future work. Due to the unbalanced nature of the data, a high window length for the RNN model might be necessary to capture nonwear periods with the limitation of high computational resources, which might be infeasible in large-scale studies. In this manuscript, we extracted dynamic features from activity levels. One may further analyze the activity levels regarding the hierarchical organization of human physical activity. Depending on the time scale of a physiological process, it can be characterized in 4 ranges. Each range has specific characteristics that might add insights into the activity pattern.[26] Similarly, if the data have sleep periods, one can find patterns in locomotor inactivity to study sleep dynamics.[27] Such patterns can be used as additional data for the machine learning model but it might require high computational memory. Another limitations is due to the nature of the capacitive sensor on the wrist which may lead to brief false positives.

Core conclusion

An XGBoost model was built on a dataset to classify nonwear and wear epochs of a commonly used sleep wearable without a nonwear sensor using the data from a nearly identical version of the device with a nonwear sensor as a reference. A set of dynamic features was created to augment the dataset and substantially increase the accuracy of the model. Even though the nonwear classification algorithm proposed in this article was applied to data from one specific device, the key innovation resides in the extraction and use of general, device-agnostic dynamic features that summarize differences in the trends and patterns of movement data characterizing wear and nonwear episodes. Thus, the algorithm is also applicable to other accelerometry devices. When compared to a temporal thresholding method, our methodology performed better on five out of the six classification metrics used in this paper. Our methodology was able to detect invalid days in the dataset and contiguous nonwear periods over 55 minutes, meaning night sleep and longer naps can be detected using this algorithm in data from a sleep wearable without a nonwear sensor. Future work may involve using complex neural network models to improve accuracy.

Author contributions

Jyotirmoy Das: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Linying Ji:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – review & editing, Project administration, Writing – original draft. **Yuqi Shen:** Data Curation, Investigation, Formal analysis, Software, Conceptualization, Writing – review & editing. **Soundar Kumara:** Writing – review & editing, Conceptualization, Supervision. **Orfeu M. Buxton:** Funding acquisition, Resources, Data curation, Writing – review & editing, Supervision, Conceptualization. **Sy-Miin Chow:** Funding acquisition, Project administration, Supervision, Writing – review & editing, Resources, Conceptualization, Methodology, Investigation.

Data and code availibility

Preprocessing code, model development code, and dynamic feature generation code are available at https://github.com/nj239/Nonwear_classification_XGBoost_dynamic_feature.git.

Funding

The source of these data, the Work Family and Health Study (www.WorkFamilyHealthNetwork.org), was funded by a cooperative agreement through the National Institutes of Health and the Centers for Disease Control and Prevention: Eunice Kennedy Shriver National Institute of Child Health and Human Development (U01HD051217, U01HD051218, U01HD051256, and U01HD051276); National Institute on Aging (U01AG027669); Office of Behavioral and Social Sciences Research and National Institute for Occupational Safety and Health (U01OH008788 and U01HD059773). Grants from the National Heart, Lung, and Blood Institute (R01HL107240), William T. Grant Foundation, Alfred P. Sloan Foundation, and the Administration for Children and Families have provided additional funding. Specific funding for this project is also provided by the Penn State Social Science Research Institute. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of these institutes and offices. National Institutes of Health grants U24AA027684, OT2 HL161847, and R01 DK134863; additional support from the Pennsylvania State University Social Science Research Institute; and the Pennsylvania State University Quantitative Social Sciences Initiative and UL1TR002014-06 from the National Center for Advancing Translational Sciences.

Declaration of conflicts of interest

Dr Orfeu M. Buxton: Outside of the current work, Dr Orfeu M. Buxton discloses that he received subcontract grants to Penn State from Proactive Life (formerly Mobile Sleep Technologies), doing business as SleepSpace (National Science Foundation grant #1622766 and NIH/National Institute on Aging Small Business Innovation Research Program R43AG056250, R44 AG056250), received honoraria/travel support for lectures from Tufts School of Dental Medicine, New York University, University of Miami, University of South Florida, University of Utah, University of Arizona; consulting fees from Georgia State University and Harvard Chan School of Public Health; and receives an honorarium for his role as the Editor in Chief of the journal Sleep Health. Dr Chow: Dr Chow reports grants from National Institutes of Health, during the conduct of the study. No Disclosures to report: JND, LJ, YS.

References

1. Danilenko KV, Stefani O, Voronin KA, et al. Wearable light-and-motion data-loggers for sleep/wake research: a review. *Appl Sci.* 2022;12(22):11794. https://doi.org/10.3390/app122211794
2. Nakagata T, Murakami H, Kawakami R, et al. Step-count outcomes of 13 different activity trackers: results from laboratory and free-living experiments. *Gait Posture.* 2022;98:24–33. https://doi.org/10.1016/j.gaitpost.2022.08.004
3. Aadland E, Andersen LB, Anderssen SA, Resaland GK. A comparison of 10 accelerometer non-wear time criteria and logbooks in children. *BMC Public Health.* 2018;18(1):323. https://doi.org/10.1186/s12889-018-5212-4
4. Berendsen BAJ, Hendriks MRC, Willems P, et al. A 20 min window is optimal in a non-wear algorithm for tri-axial thigh-worn accelerometry in overweight people. *Physiol Meas.* 2014;35(11):2205–2212. https://doi.org/10.1088/0967-3334/35/11/2205
5. Lok R, Zeitzer JM. A temporal threshold for distinguishing off-wrist from inactivity periods: a retrospective actigraphy analysis. *Clocks Sleep.* 2020;2(4):466–472. https://doi.org/10.3390/clockssleep2040034
6. Thapa-Chhetry B, Arguello DJ, John D, Intille S. Detecting sleep and nonwear in 24-h wrist accelerometer data from the National Health and Nutrition Examination Survey. *Med Sci Sports Exerc.* 2022;54(11):1936–1946. https://doi.org/10.1249/MSS.0000000000002973
7. Vert A, Weber KS, Thai V, et al. Detecting accelerometer non-wear periods using change in acceleration combined with rate-of-change in temperature. *BMC Med Res Methodol.* 2022;22(1):147. https://doi.org/10.1186/s12874-022-01633-6
8. Choi L, Liu Z, Matthews CE, Buchowski MS. Validation of accelerometer wear and nonwear time classification algorithm. *Med Sci Sports Exerc.* 2011;43(2):357–364. https://doi.org/10.1249/MSS.0b013e3181ed61a3
9. Hecht A, Ma S, Porszasz J, Casaburi R. For The COPD Clinical Research Network. Methodology for using long-term accelerometry monitoring to describe daily activity patterns in COPD. *COPD J Chronic Obstr Pulm Dis.* 2009;6(2):121–129. https://doi.org/10.1080/15412550902755044
10. Sundararajan K, Georgievska S, Te Lindert BHW, et al. Sleep classification from wrist-worn accelerometer data using random forests. *Sci Rep.* 2021;11(1):24. https://doi.org/10.1038/s41598-020-79217-x
11. De Zambotti M, Menghini L, Grandner MA, et al. Rigorous performance evaluation (previously, "validation") for informed use of new technologies for sleep health measurement. *Sleep Health.* 2022;8(3):263–269. https://doi.org/10.1016/j.sleh.2022.02.006
12. Bray J, Kelly E, Hammer L, et al. *An Integrative, Multilevel, and Transdisciplinary Research Approach to Challenges of Work, Family, and Health.* Methods Rep RTI Press; 2013. https://doi.org/10.3768/rtipress.2013.mr.0024.1303
13. Olson R, Crain TL, Bodner TE, et al. A workplace intervention improves sleep: results from the randomized controlled Work, Family, and Health Study. *Sleep Health.* 2015;1(1):55–65. https://doi.org/10.1016/j.sleh.2014.11.003
14. Berkman LF, Kelly EL, Hammer LB, et al. Employee cardiometabolic risk following a cluster-randomized workplace intervention from the Work, Family and Health Network, 2009–2013. *Am J Public Health.* 2023;113(12):1322–1331. https://doi.org/10.2105/AJPH.2023.307413
15. Chen T, Guestrin C. *XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM; 2016:785–794. https://doi.org/10.1145/2939672.2939785
16. Hyndman RJ, Wang E, Laptev N. *Large-scale unusual time series detection. 2015 IEEE International Conference on Data Mining Workshop (ICDMW).* IEEE; 2015:1616–1619. https://doi.org/10.1109/ICDMW.2015.104
17. tsfeatures.pdf. Available at: https://cran.r-project.org/web/packages/tsfeatures/tsfeatures.pdf. Accessed April 4, 2024.
18. Milner CE, Cote KA. Benefits of napping in healthy adults: impact of nap length, time of day, age, and experience with napping. *J Sleep Res.* 2009;18(2):272–281. https://doi.org/10.1111/j.1365-2869.2008.00718.x
19. Bergstra J, Yamins D, Cox DD. *Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. International Conference on Machine Learning.* PMLR; 2013:115–123.
20. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30 (2017).
21. XGBoost Parameters — xgboost 2.1.0-dev documentation. Available at: https://xgboost.readthedocs.io/en/latest/parameter.html. Accessed April 4, 2024.
22. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3(1):32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3
23. Lundberg S, Lee SI. A unified approach to interpreting model predictions. Published online November 24, 2017. Available at: http://arxiv.org/abs/1705.07874. Accessed February 21, 2024.
24. Schmidt RM. Recurrent Neural Networks (RNNs): a gentle introduction and overview. Published online November 23, 2019. Available at: http://arxiv.org/abs/1912.05911. Accessed April 23, 2024.
25. Sharma M, Lodhi H, Yadav R, Acharya UR. Sleep disorder identification using wavelet scattering on ECG signals. *Int J Imaging Syst Technol.* 2024;34(1):e22980. https://doi.org/10.1002/ima.22980
26. Búzás A, Makai A, Groma GI, et al. Hierarchical organization of human physical activity. *Sci Rep.* 2024;14(1):5981. https://doi.org/10.1038/s41598-024-56185-0
27. Winnebeck EC, Fischer D, Leise T, Roenneberg T. Dynamics and ultradian structure of human sleep in real life. *Curr Biol.* 2018;28(1):49–59.e5. https://doi.org/10.1016/j.cub.2017.11.063