# Master Thesis

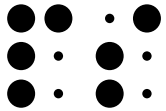
# Robust Recognition of Sleep Behavior Using Wearable Sensors

Milan 28 July 2021

STUDENT: LIDIA ALECCI
SUPERVISOR: PROF. SILVIA SANTINI
CO-SUPERVISOR: PROF. FRANCESCA GASPARINI AND PH.D. STUDENT SHKURTA GASHI

# Motivation

- **Sleep:** supporting humans in daily life, affects humans' health in short and long term
- **Wearable:** already used by consumers, nowadays almost unobtrusive, able to capture physiological signals
- **Robust:** the overall performance of a system is not impacted by noise o erroneous measurements.



# Thesis in a nutshell

Wearable sensors:

- Electrodermal activity
- Accelerometer
- Skin Temperature

Problems addressed:

- Sleep/Awake segments
- Sleep Quality (Very Poor, Poor, Normal, Good, Excellent)

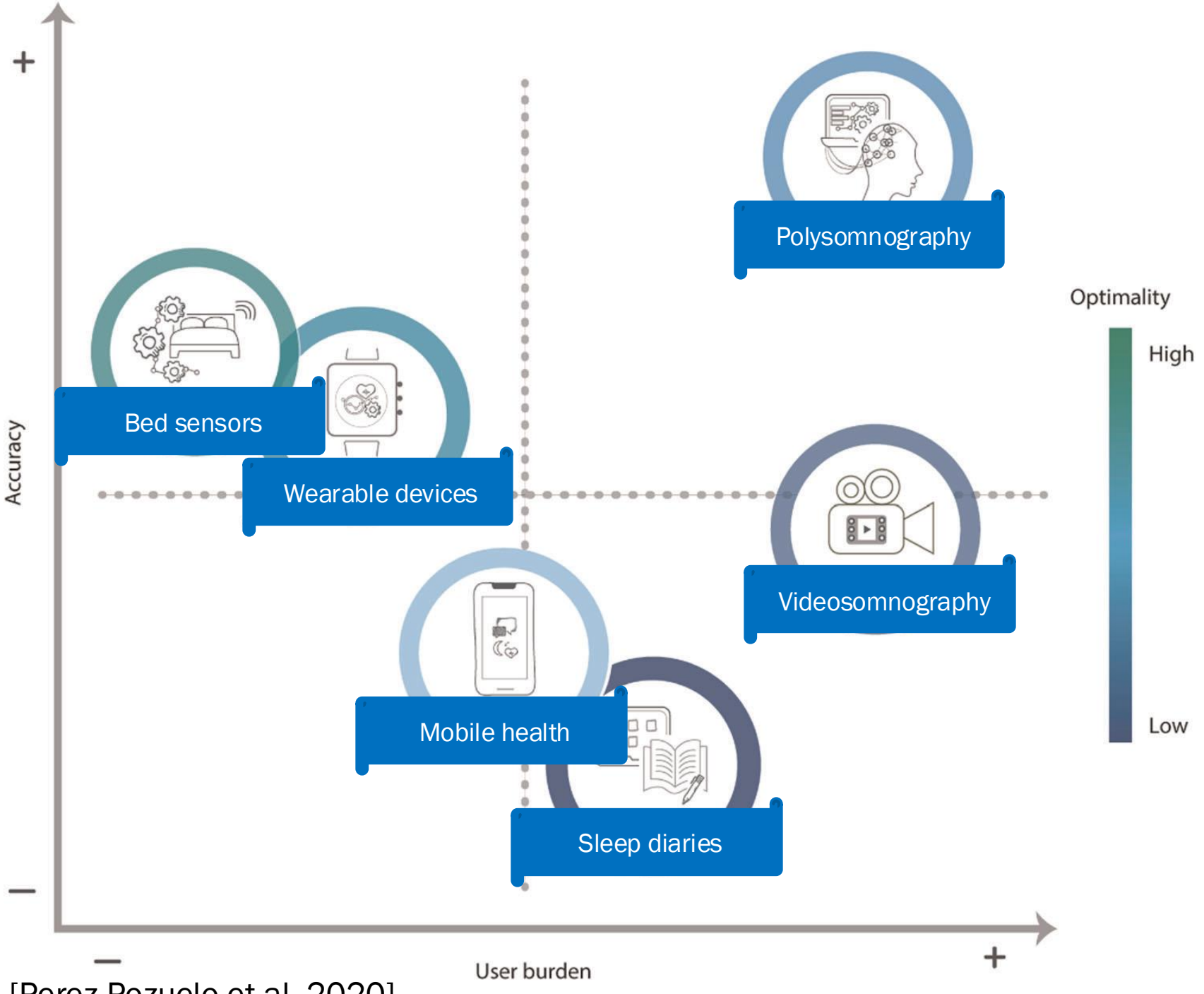
- Investigate the impact of artifact, peak epoch and storm
- Compare our performances with commercial devices (Garmin, FitBit, MiBand)

# Related work

Different methods: medical device, non-wearable [Min et al. 2014], wearable [Sadeghi et al. 2019]

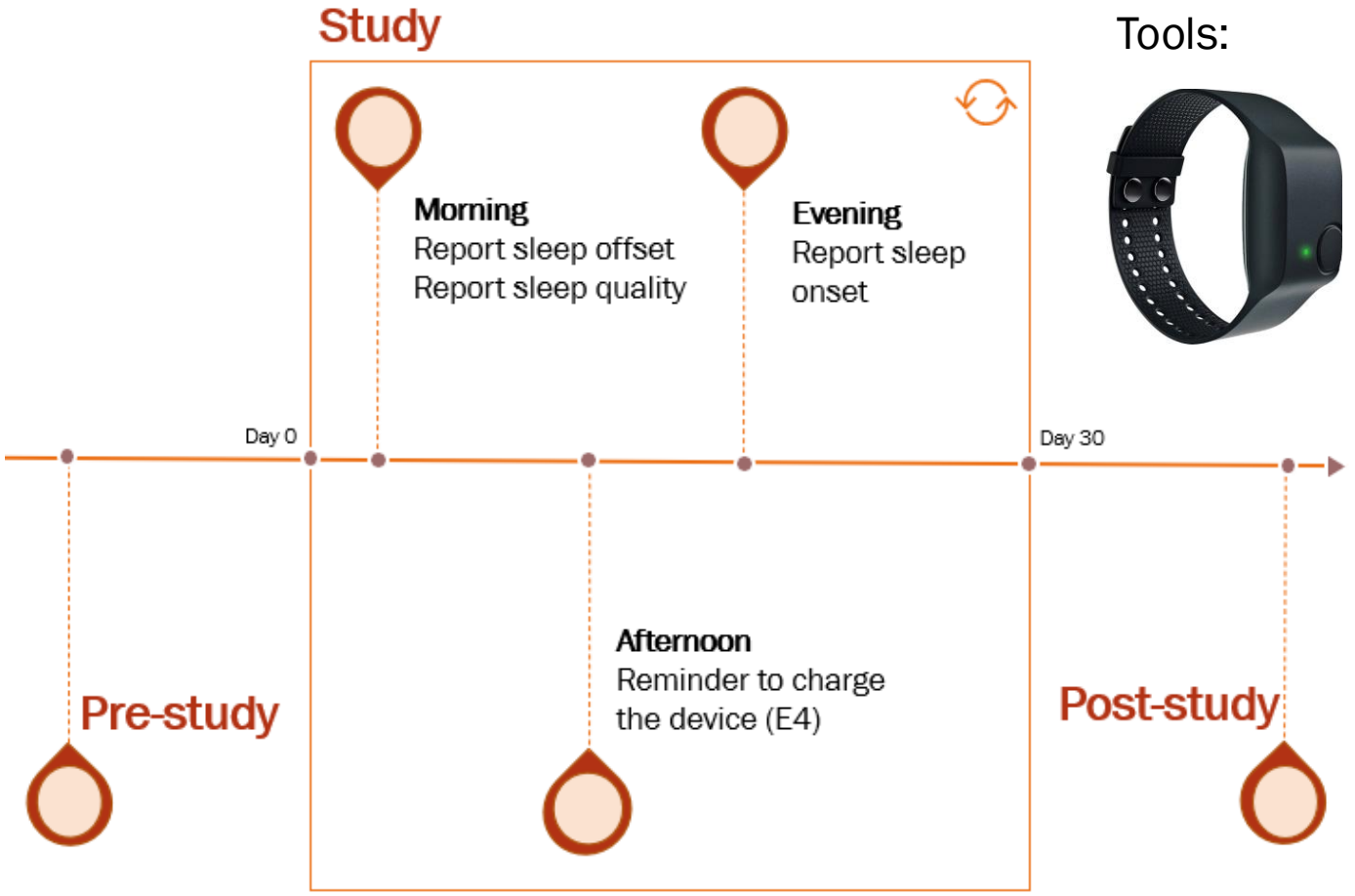
Other studies address only sleep/awake or sleep quality

Few approaches uses signal characteristics (storms, peak epochs and artifacts) in sleep detection [Sano et al. 2015]



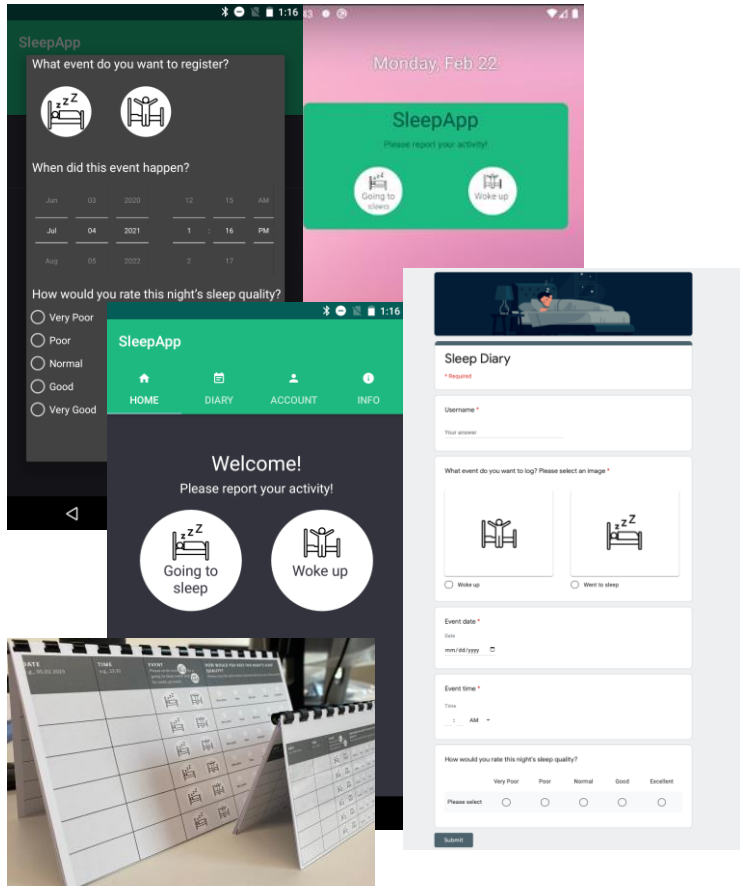

[Perez-Pozuelo et al. 2020]

# Data collection protocol




Tools:





# Data collected

**Behavioral data:**

- Phone lock/unlock
- Screen on/off
- Application usage
- Time and application of notifications
- Proximity
- Light

**Physiological data:**

- Skin temperature
- Electrodermal activity
- Accelerometer
- Blood Volume Pulse

**Self-reports:**

- Sleep onset
- Sleep offset
- Sleep quality

**Questionnaires:**

- Demographics
- Pittsburg sleep quality index (PSQI)
- Big five inventory (BFI)
- Munich chronotype questionnaire (MCTQ)

Pre

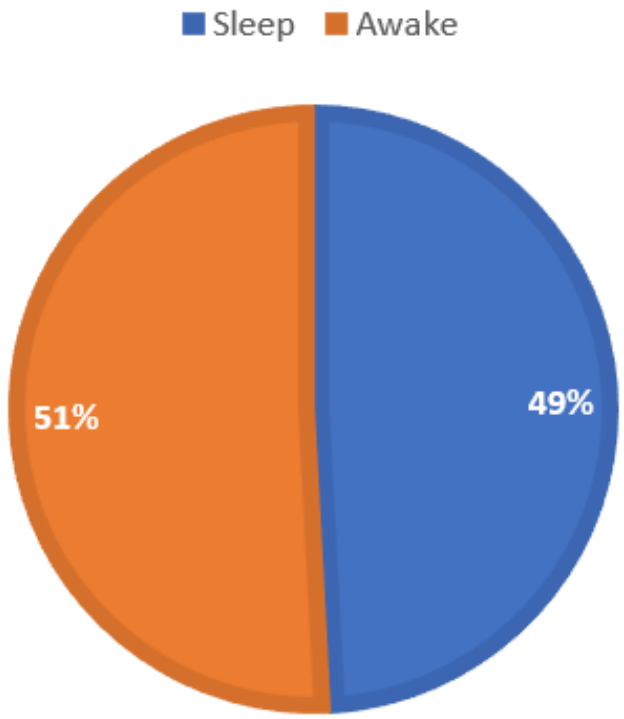
- Experience with study and tools

Post

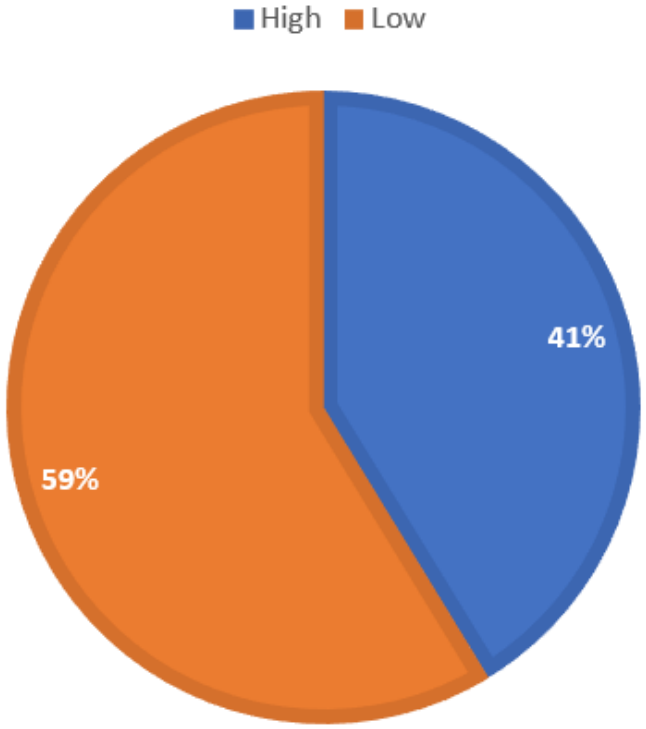
# Dataset

- 6557 hours distributed as follows:

Sleep/Awake distribution

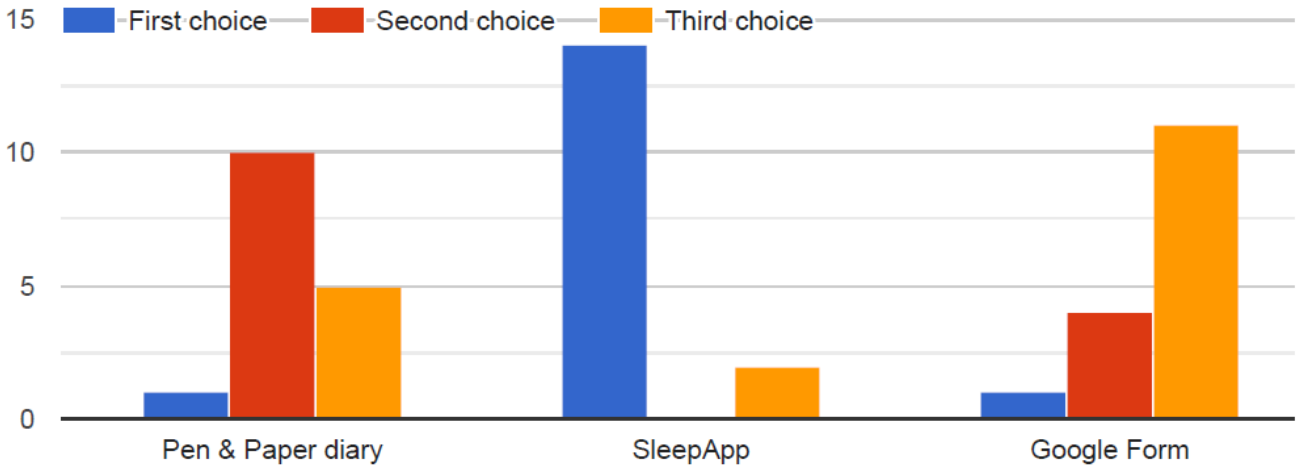



Sleep Quality two classes (High, Low)

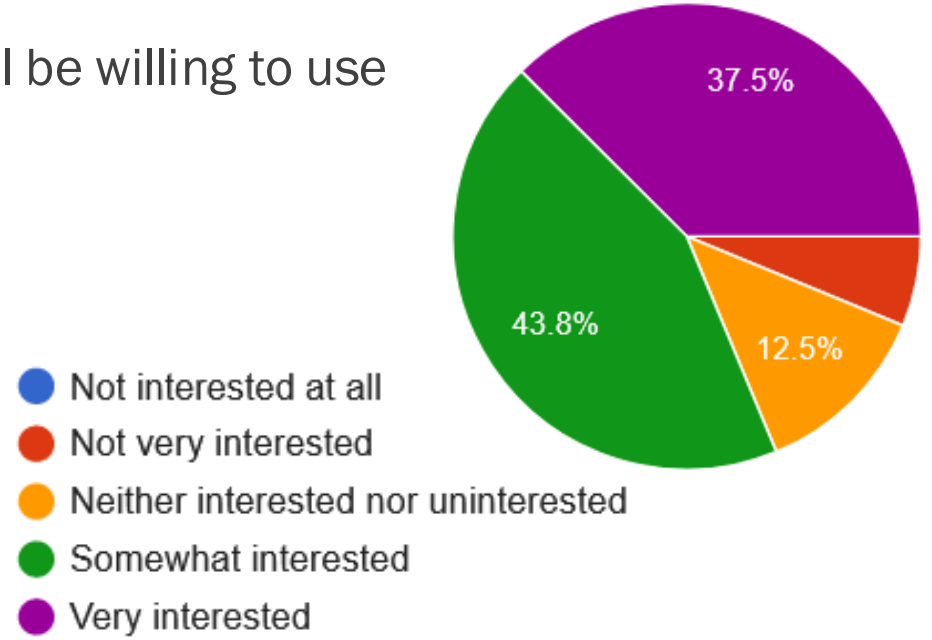


# Results from questionnaires

All users indicate wearable sensors as devices that they will be willing to use for measuring sleep behavior

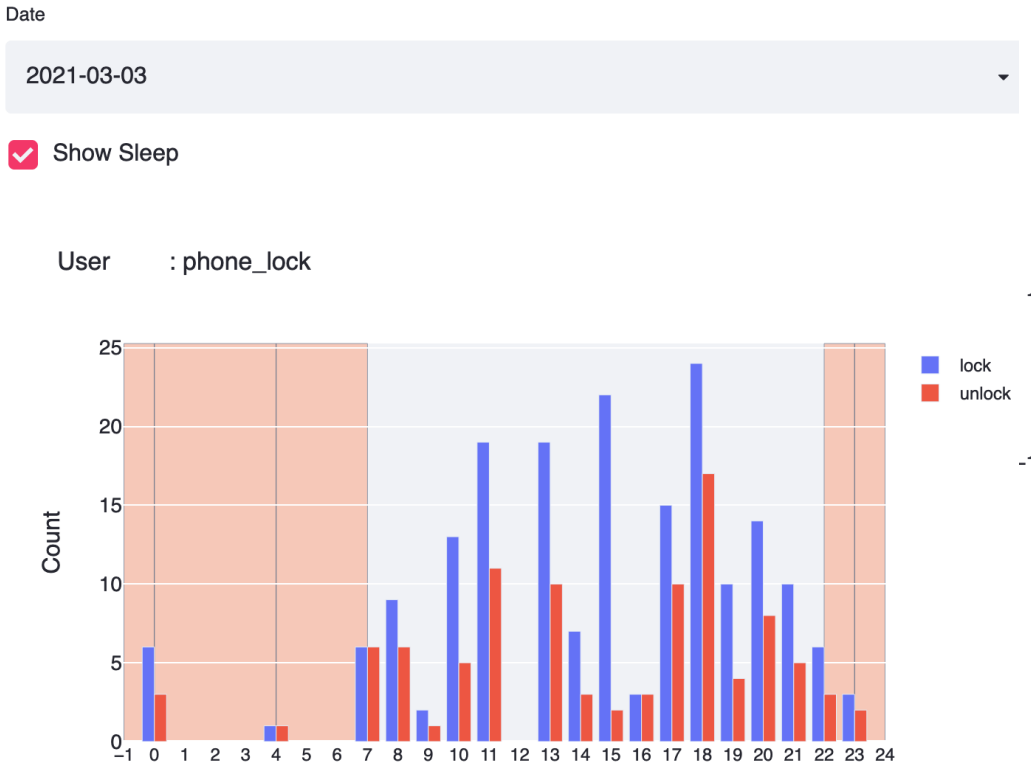


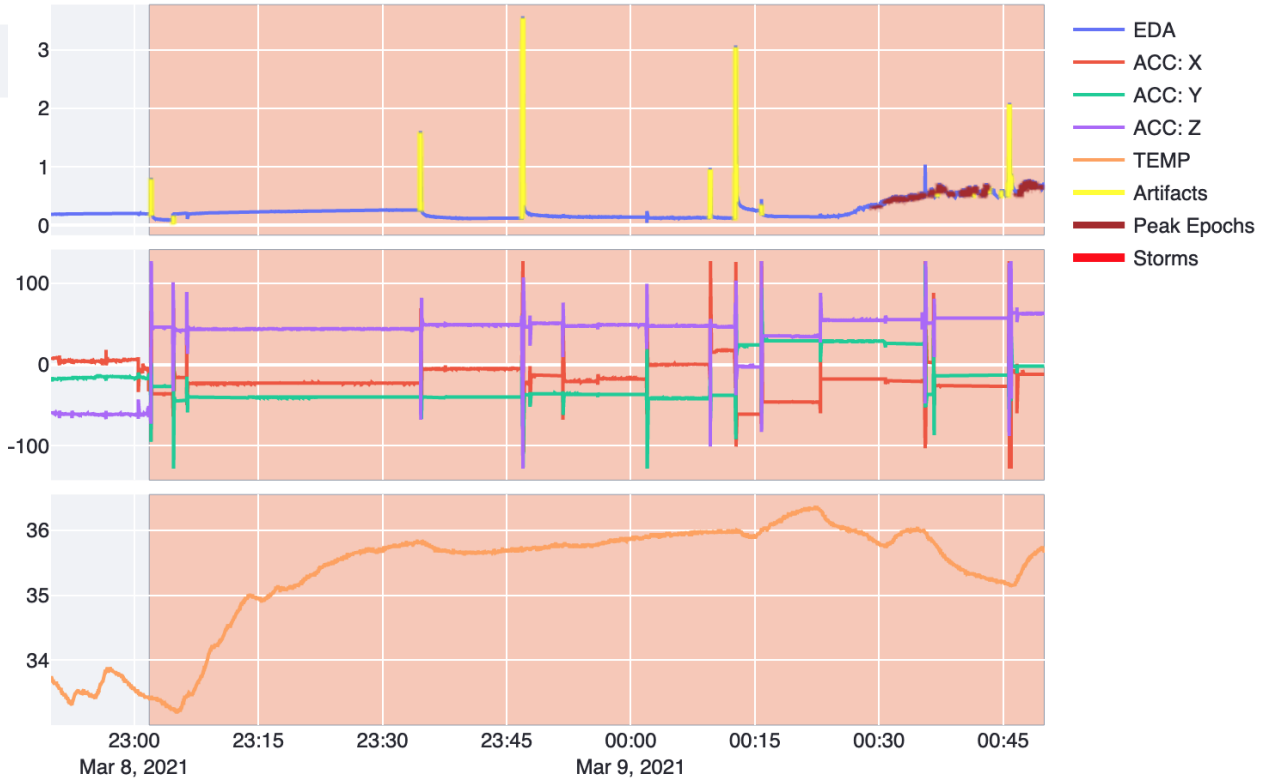

Favorite tool for reporting sleep/wake events and sleep quality




Interested on knowing physiological data (e.g., heart rate, body temperature, etc.) throughout the day and night

# Data visualization



# Data analysis

Sensors:
ACC
EDA
TEMP

Missing data: 0

Data processing

Segmentation

Feature extraction: statistical

Classification: XGBoost

Model interpretation: SHAP

# Feature extraction

Sensors:

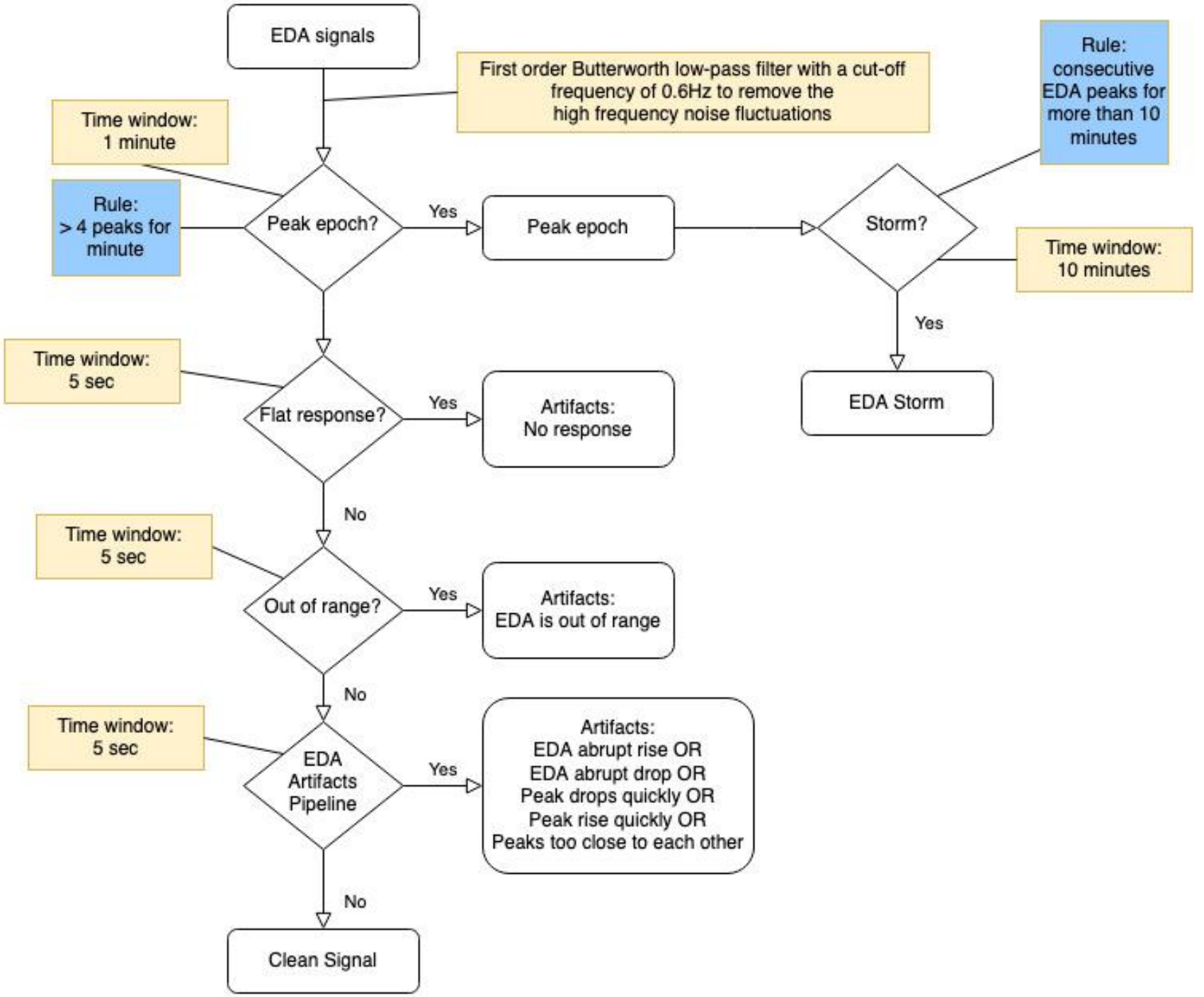
- Accelerometer: X, Y, Z
- Skin Temperature: TEMP
- Electrodermal activity: EDA
  Filtered (removing high frequency noise), Phasic, Tonic, Artifact, Peak Epoch, Storm

Segmentation windows: 10 minutes, 5 minutes, 1 minute.

For each segmentation window statistical features: mean, standard deviation, sem (standard error of the mean of values within each group), maximum, minimum, median, variance, 7-quantiles.

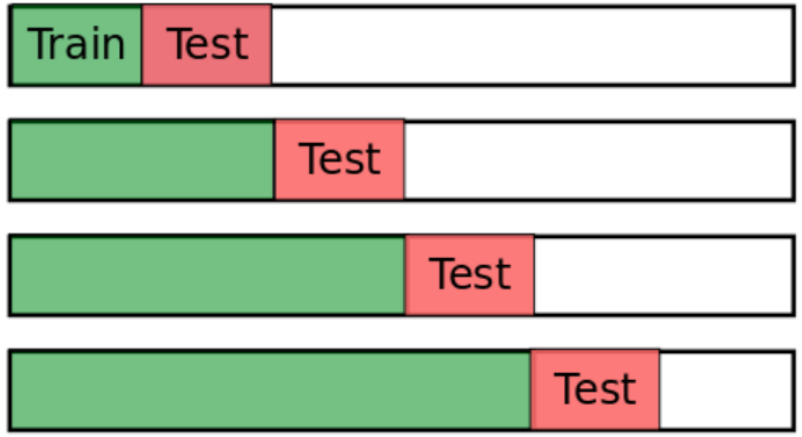
# Identifying EDA signal characteristics

- Artifacts: [Gashi et al. 2020]
- Peak Epoch: when there is a minimum of 4 peaks in a time window of 1 minute [Sano et al. 2012]
- Storm: peaks epochs that last more than 10 minutes [Sano et al. 2012]



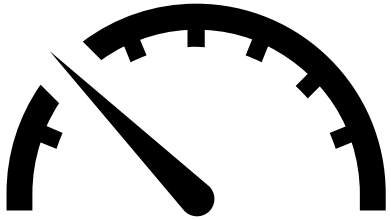
# Evaluation

**User-dependent:** for each user we select each future session as test (with at least 4 past sessions in chronological order) and only past sessions as training




**User-independent:** leave one subject out (LOSO)

**Metrics:** accuracy, balanced accuracy, recall and precision [Plotz et al. 2021]



# Results – Sleep vs Awake

| Time Window | Model | User-independent | User-dependent |
|---|---|---|---|
| 1 minute | EDA | 78.13% | 77.32% |
| | TEMP | 79.64% | 80.73% |
| | **ACC** | **88.11%** | **87.99%** |

Best one between each sensor alone: Accelerometer

# Results – Sleep vs Awake

| Time Window | Model | User-independent | User-dependent |
|---|---|---|---|
| 1 minute | EDA | 78.13% | 77.32% |
| | TEMP | 79.64% | 80.73% |
| | ACC | 88.11% | 87.99% |
| | EDA+ACC+TEMP | 89.75% | 89.93% |
| 5 minutes | EDA+ACC+TEMP | 90.14% | 90.58% |
| 10 minutes | **EDA+ACC+TEMP** | **90.58%** | **90.61%** |
| | SHAP top 20 | 90.42% | 90.35% |

- 10 minutes: best windows
- SHAP 20 features ≈ EDA+ACC+TEMP (56 features)
- There are not big differences between user-independent and user-dependent

# Results – Sleep vs Awake

| Time Window | Model | User-independent | User-dependent |
|---|---|---|---|
| 1 minute | EDA | 78.13% | 77.32% |
| | TEMP | 79.64% | 80.73% |
| | ACC | 88.11% | 87.99% |
| | EDA+ACC+TEMP | 89.75% | 89.93% |
| 5 minutes | EDA+ACC+TEMP | 90.14% | 90.58% |
| 10 minutes | **EDA+ACC+TEMP** | **90.58%** | **90.61%** |
| | SHAP top 20 | 90.42% | 90.35% |
| Biased Random Guess (based on the distribution) | | 49.99% | 50.00% |

# Results – Binary Sleep Quality

| Time Window | Model | User-independent | User-dependent |
|---|---|---|---|
| 1 minute | EDA | 50.27% | 63.60% |
| | TEMP | 48.89% | 63.70% |
| | ACC | 48.91% | 60.71% |
| | EDA+ACC+TEMP | 49.89% | 62.78% |
| 5 minutes | EDA+ACC+TEMP | 49.97% | 62.62% |
| 10 minutes | **EDA+ACC+TEMP** | **51.27%** | **62.63%** |
| | SHAP top 20 | 49.90% | 61.46% |

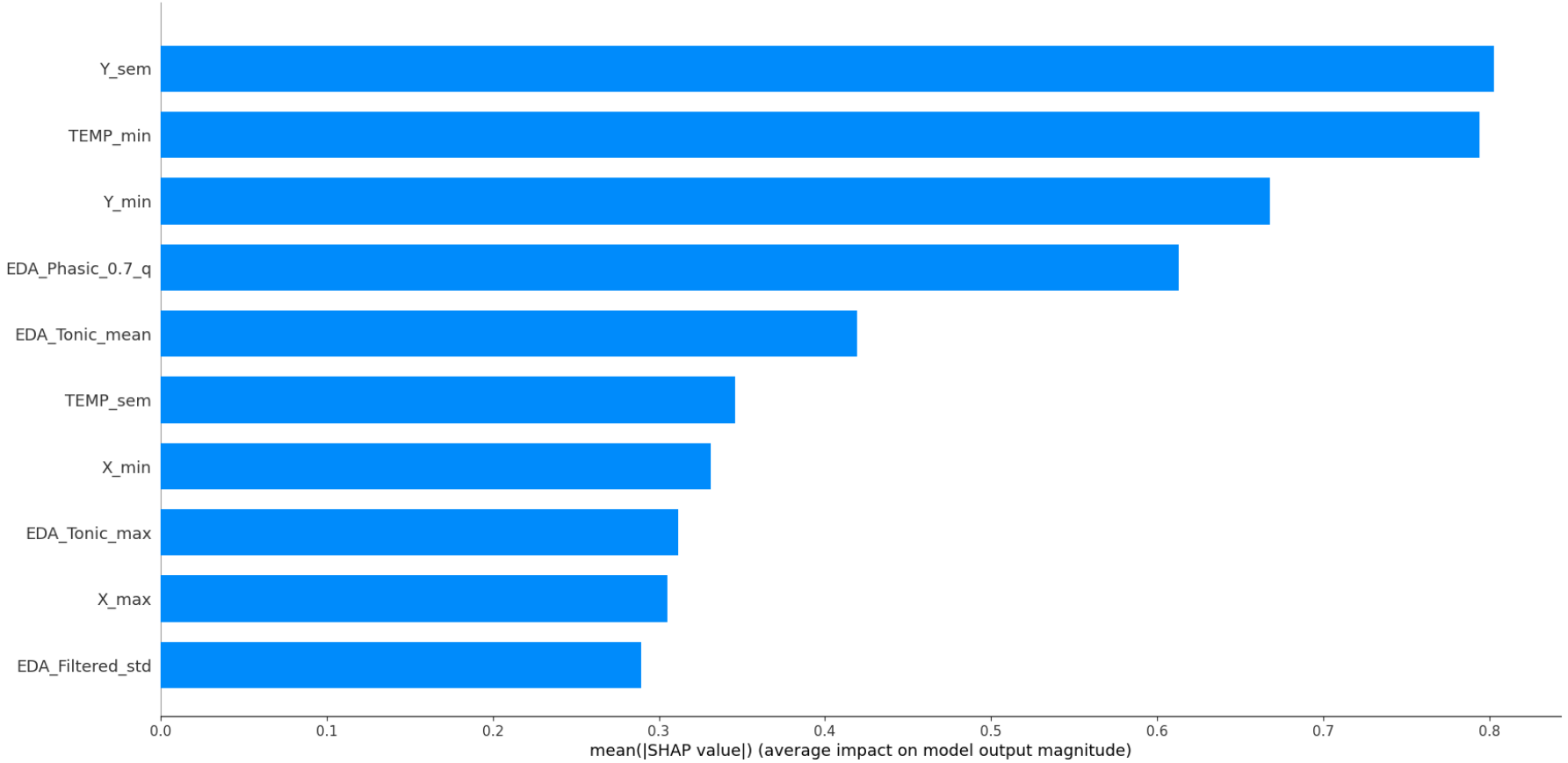
- 10 minutes: best windows
- Almost 10 percentage points between user-dependent and user-independent
- SHAP 20 features ≈ EDA+ACC+TEMP (56 features)

# Results – Binary Sleep Quality

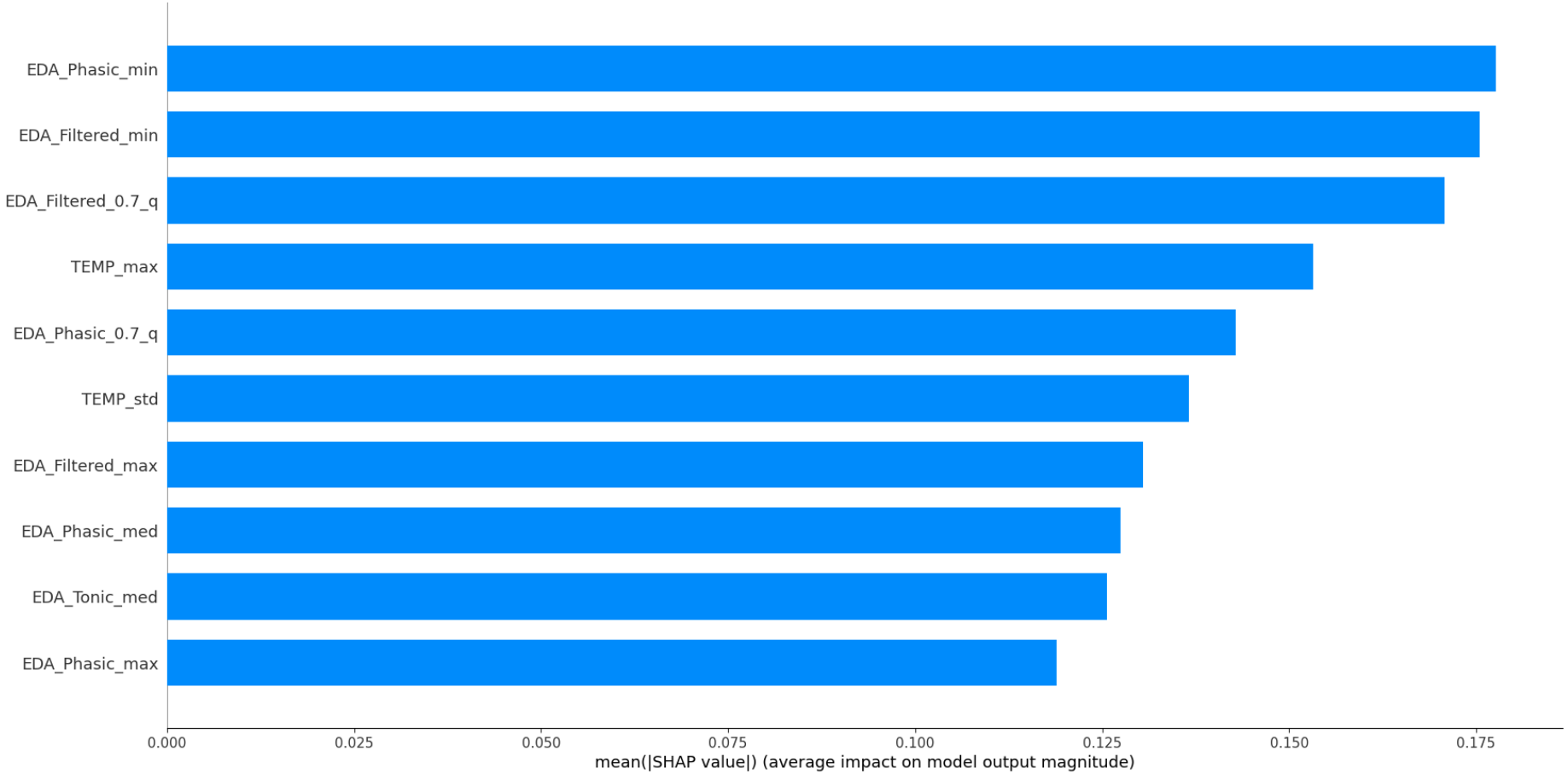
| Time Window | Model | User-independent | User-dependent |
|---|---|---|---|
| 1 minute | EDA | 50.27% | 63.60% |
| | TEMP | 48.89% | 63.70% |
| | ACC | 48.91% | 60.71% |
| | EDA+ACC+TEMP | 49.89% | 62.78% |
| 5 minutes | EDA+ACC+TEMP | 49.97% | 62.62% |
| 10 minutes | EDA+ACC+TEMP | 51.27% | 62.63% |
| | SHAP top 20 | 49.90% | 61.46% |
| | **Storm + Peak Epoch** | **60.61%** | **65.47%** |
| Biased Random Guess (based on the distribution) | | 20.03% | 29.36% |

Just Storm + Peak Epoch (2 features) performed better than EDA+ACC+TEMP

# Best features (SHAP) – Sleep/Awake



# Best features (SHAP) – Low/High Sleep Quality



# Comparison to existing devices

| Device | Balanced Accuracy |
|---|---|
| MiBand1 | 97.82% |
| MiBand2 | 96.41% |
| **Fitbit** | **97.89%** |
| Garmin1 | 93.59% |
| Garmin2 | 92.82% |
| Our approach | 90.61% |

Sleep/Awake problem

| Device | Balanced Accuracy |
|---|---|
| MiBand1 | 5.71% |
| MiBand2 | 17.14% |
| **Our approach** | **46.54%** |

Sleep quality problem with 5 classes

# Limitations and future work

- No distinction between nap and nightly sleep
  - Use different models
- Preprocessing only on electrodermal activity
  - Apply filtering methods also to accelerometer data
- Do not consider the temporal aspect of the data
  - Use long short-term memory networks (LSTM)
- No use of phone features
  - Combined physiological features with behavioral ones

# Conclusions and implications

- Wearable sensors are very promising in future health monitor systems
- Sleep/awake problem reaches a balanced accuracy above 90%
- Sleep quality is more a user dependent model and depends on storm and peak epoch (just using storm and peak epoch the balanced accuracy is 65.47% in user-dependent)

# Thank you!

# Backup Slides

# Limitations and future work

- Definition of sleep quality no trivial
  - Understand the rest level of the user
- No distinction between nap and nightly sleep
  - Use different models
- Preprocessing only on electrodermal activity
  - Apply filtering methods also to accelerometer data
- Signals as frequency but only statistical features
  - Extracting features also from frequency domain (e.g., slope changes)
- Do not consider the temporal aspect of the data
  - Use long short-term memory networks (LSTM)
- No use of phone features
  - Combined physiological features with behavioral ones

# Contributions

- Design and carry out a data collection in a real-world setting
- Dedicated tools to monitor data quality and quantity during data collection
- Dashboard to visualize collected data and visually inspect it
- Extension of EDArtifact by adding peak epochs and storms detection with definition based on literature studies
- A machine learning pipeline to detect sleep/awake and subject sleep quality using electrodermal activity, skin temperature and acceleration data collected with wristbands
- Evaluation of the model by comparing its performance with: its development as a user-dependent model and a user-independent model, different variants with different sensors and features, different time windows used, three baselines
- Evaluation of commercial devices (two MiBand, one Fitbit and two Garmin) comparing them with self-reports collected during our study
- Understand limitation of the current work and suggest future improvements

# Metrics (1)

**Accuracy** [Gron, 2017] is the percentage of correctly classified instances. It is calculated as

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

**Recall** [Gron, 2017] is also called positive predictive value (PPV) and it can be considered as how many of the actual positives are true positive. Recall is a crucial metric to look at when there is a high cost associated to a false negative (e.g., disease detection). It is computed as

$$Recall = \frac{TP}{TP + FN}$$

| | Actual positive (1) | Actual negative (0) |
|---|---|---|
| Predict positive (1) | True Positive (TP) | False Positive (FP) |
| Predict negative (0) | False Negative (FN) | True Negative (TN) |

# Metrics (2)

**Precision** [Gron, 2017] is also called true positive rate or sensitivity and can be defined as how precise and accurate the model is, by looking at how many of those predicted positive are actual positive. The formula to obtain this metric is

$$Precision = \frac{TP}{TP + FP}$$

**Balanced accuracy** [Gron, 2017] explains as a percentage how good a classifier is by also taking into account the classes balance. It is computed as

$$Balanced\ accuracy = \frac{1}{2} * \frac{TP}{TP + FN} + \frac{1}{2} * \frac{TN}{FP + TN}$$