Handling Missing Data For Sleep Monitoring Systems

Shkurta Gashi
Università della Svizzera italiana
Lugano, Switzerland
shkurta.gashi@usi.ch

Lidia Alecci
Università della Svizzera italiana
Lugano, Switzerland
lidia.alecci@usi.ch

Martin Gjoreski
Università della Svizzera italiana
Lugano, Switzerland
martin.gjoreski@usi.ch

Elena Di Lascio
Università della Svizzera italiana
Lugano, Switzerland
elena.di.lascio@usi.ch

Abhinav Mehrotra
Samsung AI Center
Cambridge, United Kingdom
a.mehrotra1@samsung.com

Mirco Musolesi
UCL and University of Bologna
London, United Kingdom and Bologna, Italy
m.musolesi@ucl.ac.uk

Maike E. Debus
Université de Neuchâtel
Neuchâtel, Switzerland
maike.debus@unine.ch

Francesca Gasparini
Università degli Studi di Milano-Bicocca
Milan, Italy
francesca.gasparini@unimib.it

Silvia Santini
Università della Svizzera italiana
Lugano, Switzerland
silvia.santini@usi.ch

2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII) | 978-1-6654-5908-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACII55700.2022.9953832

***Abstract*—Sensor-based sleep monitoring systems can be used to track sleep behavior on a daily basis and provide feedback to their users to promote health and well-being. Such systems can provide data visualizations to enable self-reflection on sleep habits or a sleep coaching service to improve sleep quality. To provide useful feedback, sleep monitoring systems must be able to recognize whether an individual is sleeping or awake. Existing approaches to infer sleep-wake phases, however, typically assume *continuous* streams of data to be available at inference time. In real-world settings, though, data streams or data samples may be missing, causing severe performance degradation of models trained on complete data streams. In this paper, we investigate the impact of missing data to recognize *sleep* and *wake*, and use *regression-* and *interpolation-based* imputation strategies to mitigate the errors that might be caused by incomplete data. To evaluate our approach, we use a data set that includes *physiological traces* – collected using wristbands –, *behavioral data* – gathered using smartphones – and *self-reports* from 16 participants over 30 days. Our results show that the presence of missing sensor data degrades the balanced accuracy of the classifier on average by 10-35 percentage points for detecting sleep and wake depending on the missing data rate. The imputation strategies explored in this work increase the performance of the classifier by 4-30 percentage points. These results open up new opportunities to improve the robustness of sleep monitoring systems against missing data.**

***Index Terms*—Wearable Sensors, Missing Data, Sleep and Wake Recognition, Machine Learning**

I. Introduction

Sleep has a pivotal effect on people's performance, memory, recovery, mental health, and physical health [1, 2]. Given the potential risks of sleep deprivation and the benefits of good sleep routines, researchers have long studied this type of human behavior. The availability of truly unobtrusive wearable devices led to an increasing attention on the design and development of systems that monitor users' sleep-wake phase in real-world settings. Besides promoting introspection about sleep routine [1], systems able to capture users' sleep habits in the real world can also provide recommendations for a better sleep quality as well as prevent distractions by avoiding notifications to be delivered during sleep [3].

To enable the design and implementation of such features of sleep monitoring systems, it is necessary to recognize sleep in a robust manner, which is the focus of this work.

Sensor data collected in real-world settings is, however, rarely continuous and uninterrupted. Even in a controlled setting, and despite employing the best data collection practices, data losses often occur (e.g., in [4]). As a result, this leads to *missing data*, which refers to "*the data value that is not stored for a variable of the observation of interest*" [5]. For instance, accelerometer sensor is of interest in sleep monitoring systems and missing data refers to the data points that are not present from this sensor for a specific time point.

An example scenario to depict the missing data challenge may be an individual taking off their sleep monitoring device to take a shower. The sensors on the watch might still record accelerometer data, even if it is noisy (e.g., all flat), and after some time the device may run out of battery leading to missing accelerometer data. Missing data might occur also because the user does not wear the device that measures sensor data (e.g., due to forgetting, or charging) [6, 7, 8], the device gets broken [7, 8, 9], there are intermittent disconnections of the device to the network [10] or the device has hard energy saving constraints [7, 8] as well as a consequence of the presence and removal of noisy sensor data [11, 12].

Collection of mobile and wearable sensor data in real-world scenarios unfortunately leads to inevitable problems with miss-

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: Universita della Svizzera Italiana. Downloaded on January 27,2023 at 10:42:08 UTC from IEEE Xplore. Restrictions apply.

ing data [8, 12]. When more than one data stream is needed for sleep monitoring, the problem of missing data becomes even more pronounced. This is because the intersection of data streams with complete data becomes smaller and smaller, which leads to the loss of valuable information from the available sensor data. Indeed, long-term data collections have shown that "*only half of the collected data*" could be used for the data analysis [12].

A simple and common solution to handle missing data is *sample deletion* [6, 10, 13, 14]. Sample deletion refers to discarding the data samples with missing data points and run the model on the remaining data [13, 14]. While this technique is effective and straightforward, it leads to a significant amount of data loss that is crucial for data-hungry machine learning models. Additionally, it can introduce bias in the results if the missingness of the data is not completely at random (e.g., is not due to device malfunction or data transfer error) [13, 15].

An alternative approach is to *ignore the missing data* points and perform the data analysis on the remaining part of the data [10]. For instance, in classical machine learning pipelines, features can be extracted only on the available data. However, this technique could result in significant performance degradation of the model trained with all the data present.

To address the limitations of the approaches mentioned above, recently researchers investigated the impact of *imputation strategies*, i.e., replacing missing sensor data with substituted values [7, 10, 12, 16]. Imputation strategies allow researchers to obtain a more complete data set and enable sleep monitoring systems to run continuously. Existing work that use imputation strategies, however, have mainly been tested using data from a context different of what we explore in this paper – e.g., mood prediction [12], physical activity recognition [6, 9], social anxiety [7] – or from data derived from a different set of sensors i.e., heart rate, breathing rate, number of steps [10], social media [17] – which might have different implications. Indeed, a simple data imputation technique might be sufficient for one data source, but not for another. For instance, skin temperature sensor data might be more static, thereby, filling missing data with a simple technique (e.g., most frequent value) might be sufficient. On the contrary, accelerometer data is more dynamic, so filling in missing data with a simple technique might not be sufficient. In addition, such techniques do not consider the impact of additional data sources to address the problem of missing data, which could help to maintain the robustness of sleep monitoring systems.

Several researchers have used sensor data to create models for recognizing sleep and wake [1, 18, 19, 20, 21]. However, to our knowledge there is no study that addresses the data messiness and algorithmic challenges of various data sources to achieve the goal of sleep detection. Thus, how missing data impacts model's performance and how best to handle incomplete data for maintaining the accuracy of sleep detection remains an open question.

In this work, we comprehensively investigate the impact of missing data and imputation strategies for sleep and wake classification using a data set collected over one month from 16 participants. We propose using sensor-specific missing data imputation techniques because the dynamics of a sensor might be different from another sensor.

In summary, we make the following contributions:

- We analyze the impact of missing wearable sensor data points for sleep and wake recognition and observe that missing data degrades the performance of the classifier by 4-30%. To the best of our knowledge, this is the first work to investigate this problem for sleep detection.
- We investigate *interpolation-* and *regression-based* imputation strategies to mitigate the problem of missing raw sensor data. Our results show that interpolation imputation strategy provides a higher classification accuracy by 10 percentage points in comparison to no imputation, even when 50% of data is missing.
- We explore whether other data sources (e.g., phone usage, hour of the day) could help to address the missing data problem. Our results show that in case of missing wearable sensor data, context information could be used to maintain the performance of the model to detect sleep.

II. Related Work

Several researchers have demonstrated the capability of sensor data to recognize whether a person is sleeping or awake [18, 21, 22, 23]. These techniques are tested in curated and homogeneous data. Thus, such models might be overoptimistic about what can be achieved in real-world settings. Indeed, it might be challenging to use such models in any real-world sleep monitoring system where the data is frequently missing.

A few existing approaches address the problem of missing data in sensor-based sensing systems [6, 7, 9, 12, 16, 17]. Saha et al. [17],propose a framework to predict missing social media data using other available data streams, such as, e.g., heart rate variability, stress, and physical activity. Their approach leads to an average improvement of 14% across all models to predict personality traits and affect. Jacques et al. [12] propose to use a multimodal autoencoder to impute missing features and achieve better mood prediction results. In comparison to these approaches, we address the missing *raw* sensor data points problem, instead of the missing features or missing data stream. Handling missing raw data points allows researchers to avoid the issue of missing data early on in the data analysis pipeline.

A few researchers have used a variety of methods to handle incomplete, missing data for sensor-based human behavior recognition [7, 16]. Rashid et al. [7] demonstrate the capability of data imputation strategies (e.g., matrix completion, multiple imputation) to increase the model's performance for predicting social anxiety using phone sensor data. The use of imputation strategies leads to a decrease of prediction error by 22%. In this work, we also investigate the impact of the imputation strategies to handle the missing data problem. In contrast, we focus on a different classification task using wearable sensor data (e.g., accelerometer, temperature, electrodermal activity).

Only a few studies have examined the impact of the data imputation on the performance of the model for detecting

Authorized licensed use limited to: Universita della Svizzera Italiana. Downloaded on January 27,2023 at 10:42:08 UTC from IEEE Xplore. Restrictions apply.

human behavior [7, 12, 16, 17]. For instance, researchers have shown the capability of imputation strategies to increase the performance of cooking activity recognition [16], mood prediction [12], as well as personality traits and affect recognition [17]. In contrast, we examine the impact of missing data and their imputation for sleep and wake recognition. Understanding the impact of missing data and imputation strategies is crucial to inform sleep monitoring systems about the type of strategy to follow in case of missing data. The imputation strategy to apply might differ from one sensor to another. For instance, for signals that are static (e.g., skin temperature) a simple imputation approach might be sufficient, however, the same method might not perform well for signals that are dynamic (e.g., accelerometer).

III. Method

A. Data set

In this work we use the data set presented in [24]. The goal of the work presented in [24] is to investigate the role of personalized and population models for sleep/wake and sleep quality recognition using sensor data. The data set contains physiological, behavioral and self-reports gathered from 16 participants (11 females and 5 males of age in the range 19 to 35 years old) over one month. The occupation of the participants is: students (10), workers (3), PhD students (2) and Post-Doc (1).

Physiological data. The data set contains physiological data of participants collected using the Empatica E4 wristband[1] [25] worn on the non-dominant hand. The E4 contains four sensors that measure: the electrodermal activity (EDA), skin temperature (TEMP), 3-axis acceleration (ACC) and blood volume pulse (BVP). The E4 measures such data with a sampling frequency of 4 Hz, 4Hz, 32Hz and 64Hz, respectively.

Behavioral data. The data set contains also behavioral data collected from phone sensors. Behavioral data was collected using an Android application, called *SleepApp*. The data set contains: time of phone lock/unlock events, screen on/off, application from which a notification arrived and whether the notification was clicked or not, proximity of the phone screen to any surface, and ambient light intensity. Information regarding the notifications was collected using the implementation of MyTraces app presented in [26].

Self-reports. Participants reported their sleep and wake up times as well as sleep quality score using the validated and standardized Pittsburgh Sleep Quality Index (PSQI) [27] questionnaire and diaries as a common procedure in the literature [19, 22, 28]. Participants chose between three tools to provide self-reports: the SleepApp installed on participant's smartphone; an online survey accessible form diary accessible from participant's laptop; or a pen-and-paper diary. Figure 1 shows an example of the *SleepApp* application.

Existing studies have shown that physiological and behavioral data collected from wearable sensors can be used to detect whether a user is sleeping or awake [19, 22, 28].

[1]https://www.empatica.com/research/e4/.

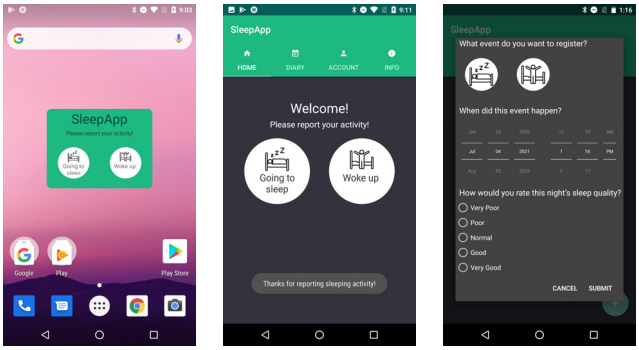

Fig. 1. Interface of the *SleepApp* Android application used to collect self-reports regarding sleep and wake up times as well as sleep quality.

In this paper we also focus on recognizing sleep and wake using physiological (e.g., EDA, TEMP and ACC data) and behavioral data (e.g., phone sensor data). We select a subset of the data set presented in [24] to be able to directly compare the classification results with existing work (e.g., [19, 24]) considering missing data.

B. Data analysis

To recognize sleep and wake, we set up a binary classification pipeline. The pipeline consists of signal processing and machine learning steps described as follows.

Data cleaning and preprocessing. We consider two types of data, namely, wristband and phones data.

Wristband sensors. To preprocess the EDA signals, we follow common preprocessing steps from the literature [11, 29, 30, 31, 32]. In particular, we filter the signal using a first order Butterworth low-pass filter with a cut-off frequency of 0.6 Hz similar to [11] to remove high frequency fluctuations. To obtain further information from the EDA signal, we decompose the signal into *tonic*, the slowly changing component, and *phasic* component, characterized by peaks in correspondence to a stimuli, using the cvxEDA method proposed by Greco et al. [33]. We down-sample the ACC signals to 4Hz similar to the sampling frequency of EDA and TEMP sensors.

Phone sensors. Proximity sensor embedded on the phone return the absolute distance of the phone to an object in *cm* or a categorical value representing whether the object was "*near*" or "*far*" from the phone. The maximum range of proximity sensor differs across phones. To make the data collected from different phones (participants) comparable to each other, we compute the ratio of the phone distance as measured by the proximity sensor to the maximum distance of the phone.

Segmentation and labeling. We use participants' self-reports to label the sensor data with the `sleep` and `wake` class. We first divide the continuous traces of sensor data into 10-minute non-overlapping windows. We use a window of 10 minutes because similar studies have shown that the median sleep latency or transition time from wake to sleep is approximately 10 minutes [22]. We assign the label `sleep` to all the windows of the signal included between the self-reported time indicating the user going to sleep and the corresponding self-report

Authorized licensed use limited to: Universita della Svizzera Italiana. Downloaded on January 27,2023 at 10:42:08 UTC from IEEE Xplore. Restrictions apply.

indicating the user to have waken up. The remaining windows are assigned to the `wake` class. The resulting data set includes 13519 windows in the sleep and 12774 in the wake class.

Missing data generation. To address the problem of missing data, it is crucial to understand the reason behind the absence of data. Feng and Narayanan [10] noticed two missing patterns of real-world sensor data: 1) random missing data points (`Random`) and 2) complete missing data over a continuous period of time (`Chunks`). To induce the missing data pattern, we eliminate sensor data points at 10, 25 and 50 percentages following these two techniques. There are three mechanisms of missing data: 1) *missing completely at random (MCAR)* – if the probability of missing is the same for all cases –, 2) *missing at random (MAR)* – if the probability of being missing is the same only within groups defined by observed data – and 3) *missing not at random (MNAR)* – when the missing values do not only depend on the observed values but also the unobserved ones – [5, 15]. To impute the missing data, most imputation strategies require having MCAR condition [5, 7, 15]. To ensure that the data set conforms to the MCAR condition, we intentionally select recordings without missing data in the data set and randomly mimic missing data. We eliminate 10, 25, and 50 percent of sensor data in a window as in [9, 10]. For instance, to eliminate a chunk of sensor data points we randomly select a data point within a window and discard the next 25% of the data points in the window. Similarly, to implement the `Random` technique, we randomly select 25% of the sensor data points within a window and discard them. When using multiple sensors, we discard the selected data points from all sensors assuming that the data from all sensors in a device would be missing at once. Data might be missing at `Random`, for instance, due to the presence and removal of noisy data and in `Chunks` due to user not wearing the device for a specific amount of time.

Missing data imputation. To handle missing data points, we investigate several imputation strategies which can be grouped into two categories: *regression-based* and *interpolation-based* [7, 13, 14]. Regression-based techniques predict the missing values of a target variable based on the other available variables in the data set. Interpolation-based methods instead fill in missing data with a placeholder value such as, e.g., zero, mean, median. In this work we explore the impact of *Zero*, *Most frequent* and *Padding* interpolation-based methods as well as *Iterative* regression-based method [13, 14]. *Zero* and *Most frequent* methods impute missing data points using a zero or the most frequent value of the sensor. *Padding* replaces missing data with the last measured value from the sensor, known also as "last observation carried forward" [7]. For the *Regressor* strategy, we consider the data points corresponding to missing values as a target of a regression model and use the data of other sensors as input to the model [14]. We apply the same technique to the test set as suggested in [14]. We model each data source with missing data points as a function of other available data sources. The missing data points of a sensor are then predicted using the regressor. We use the IterativeImputer[2] class from scikit-learn to implement the *Regressor* imputation strategy. Fig. 2 presents an example of the EDA signals and the missing data imputation strategies explored in this work.

We do not apply the missing data analysis steps to phone sensor data because it is difficult to determine whether there was no event or whether the data was actually missing. Instead, we investigate whether the available phone sensor data can help to maintain the classification performance in case of missing wristband sensor data.

Feature extraction. We consider three types of features, physiological, behavioral and context ones.

Physiological features. From each sensor of the wristband we extract three groups of features: *time*, *frequency*- and *time-frequency* domain, similar to [11, 24, 34, 35]. The features in time-domain representation include, statistical features such as, e.g., the min, max, median, variance, dynamic range, mean and standard deviation of the first derivative, difference between the last and first sensor value, and the slope of the signal. We derive the same statistical features also from wavelets coefficients extracted at three different time scales 4Hz, 2Hz and 1Hz, as in [11], to which we refer to as time-frequency domain features. To capture the periodicity of sensor data, we transform each data stream into the frequency-domain using the Fast Fourier Transform (FFT) [35] and extract features in this domain similar to [35]. We then compute features such as, e.g., the direct current component (DC), the sum of spectral coefficients, the information entropy and the energy of the signal. We compute the time-, frequency- and time-frequency domain features from the EDA, the tonic component of the EDA, the TEMP, and from each axis of the ACC sensor.

We extract further characteristics of the phasic component of the EDA using the *EDAExplorer*[3] [34] and *EDArtifact*[4] [11] tool-kits. We extract peaks-related features (e.g., the number of peaks, amplitude of the peak, peak rise and decay time, peak width) similar to [11, 34]. Several researchers have shown that EDA signals contain periods of high frequency activity, known as *storms*, during the night and in particular during deep sleep [19, 36]. Burch et al. [37] define storms as "*regions of EDA signal with a burst of high frequency peaks*". To detect storms and epochs, we use the definition from Sano et al. [38]. In particular, an EDA peak epoch refers to a region of EDA with more than 4 peaks per minute, whereas an EDA storm refers to regions where EDA peak epochs last for more than 10 minutes. We further extract features related to the storms and epochs such as a storm or epoch flag, similar to [19]. We extract in total 173 features from physiological signals.

Behavioral features. From phone sensor data we extract in total 13 features. The features include the total number of notifications clicked by the user during the 10-minute window, total number of times the user unlocks the phone, the number of times the screen is on, and the minimum, maximum, mean,

[2]https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.htmlIterativeImputer

[3]https://github.com/MITMediaLabAffectiveComputing/eda-explorer

[4]https://github.com/shkurtagashi/EDArtifact

Authorized licensed use limited to: Universita della Svizzera Italiana. Downloaded on January 27,2023 at 10:42:08 UTC from IEEE Xplore. Restrictions apply.

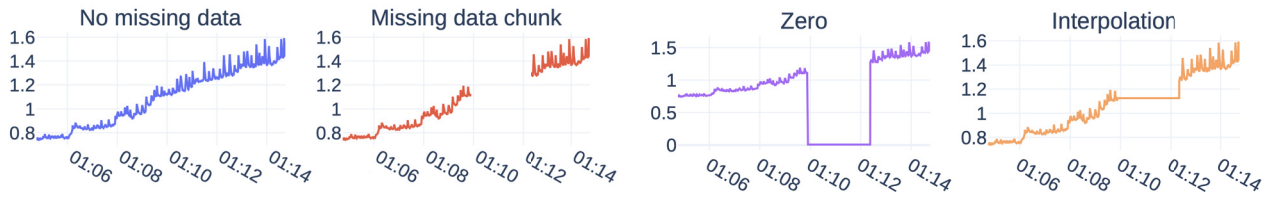

Fig. 2. Example of the electrodermal activity signals in the presence of all the data (`No missing data`), eliminating 25% of the data points (`Missing data chunk`) and replacing missing data with several imputation strategies (e.g., `Zero`, `Interpolation`).

mode and standard deviation of the proximity and the light sensors over the session.

Context features. There are two main components that determine sleep and wake up time: *circadian rhythm* and *homeostatic sleep dive* [1, 3]. Circadian rhythm promotes wakefulness during the day and sleep during the night and is synchronized to the time of the day. Homeostatic sleep drive refers to the pressure of sleep linearly building up in the brain while being awake and decreasing during sleep [1]. These two components have shown to improve sleep and wake recognition [1]. In this work, we investigate the impact of such information to recognize sleep and wake in case of missing sensor data. To encode the time information, we first extract the hour of the day from the timestamp of a sensor data point measurement. We then compute the `cosine` and `sine` functions on the hour of the day, by first normalizing it between 0 and 2π, to preserve the cyclical nature of the time of the day. Considering that sleep and wake up time changes throughout the week, we also extract the day of the week and day type (e.g., weekday or weekend) from the timestamp of a sensor data point measurement.

Classifiers. We investigate the performance of models built with the Gradient Boosting (`GB`) algorithm [39] because it has shown to perform best for sleep and wake classification in [24]. Additionally, it has shown competitive performance to neural networks in several machine learning competitions [40, 41]. We compare the performance of the GB classifier with and without missing data, to understand the impact of missing data in the classification task. We further investigate the impact of features in classifier's predictions using the SHapley Additive exPlanations (SHAP), as in [21].

Evaluation procedure. We follow common procedures in machine learning for sensor data to evaluate the performance of our approach [42], i.e., leave-one-subject-out (LOSO) approach. LOSO trains a classifier with the data of all users, except one which is kept as a test data. We repeat the same procedure for all the users. This technique ensures that the physiological data of the same user is not present both in train and test sets simultaneously, investigating the generalizability of the approach. In the training phase, we re-scale the features using the z-score standardization technique[5] and account for the imbalance nature of our data set by oversampling the minority class (e.g., wake class) to the majority class using the synthetic minority oversampling technique (SMOTE) [43], as a common procedure in machine learning [14].

[5]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

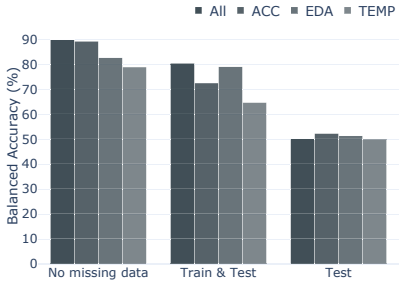

Fig. 3. Balanced accuracy of the `GB` classifier on the whole data set (`No missing data`), with 10% of the missing data in the training and test sets (`Train & Test`), and with missing data only on the test set (`Test`).

Evaluation metrics. We evaluate the classifiers using the balanced accuracy (BA) proposed in [44] and used in [45]. BA refers to the mean of recall score of a classifier for the both the majority and minority class [45].

IV. Results and Discussion

In this section, we first discuss the impact of missing data in sleep and wake classification. We then compare the performance of the GB classifier with and without missing data. We further investigate whether phone sensor data and context information can help to mitigate the missing data problem.

Impact of missing data in sleep and wake detection. In these experiments, we study the impact of missing data points within a window to classify sleep and wake using the GB model. Fig. 3 reports the accuracy using the features from the ACC, EDA, TEMP or the three combined (`All`) using the complete data set (`No missing data`) as well as the data set with missing data points both in the train and test set (`Train & Test`) or only the test set (`Test`). The BA when using all, ACC, EDA, or TEMP sensor is 90, 89, 82 and 79, respectively. In case of 10% of missing data points in both the train and test set, the accuracy of the GB classifier drops to 80, 72, 79, 64. A more realistic scenario is when the data set used to train the classifier is complete, while the test set contains missing data. We refer to this scenario as the `Test` in Fig. 3. From the figure we observe that the performance of the classifier for all sensors drops to 50%, which is not significantly higher than a random guess classifier. These results indicate that the model is not robust against incomplete data and hint at the need of addressing this problem.

Authorized licensed use limited to: Universita della Svizzera Italiana. Downloaded on January 27,2023 at 10:42:08 UTC from IEEE Xplore. Restrictions apply.

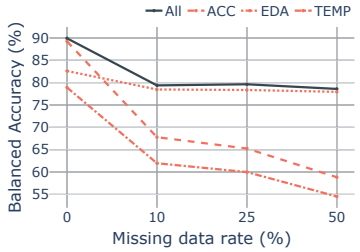

Fig. 4. Performance of the GB classifier, in terms of BA, using the whole data set (missing data rate = 0), and eliminating 10%, 25% and 50% of the sensor data points in a window using the `Random` technique.

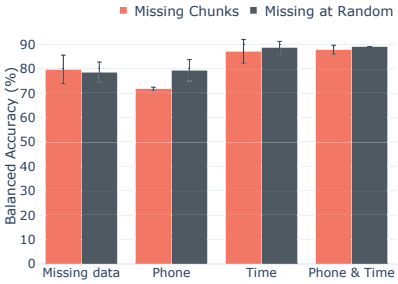

Fig. 5. Balanced accuracy of the GB classifier with 10% of the missing data (`Missing data`) as well as adding features extracted from the phone sensors (`Phone`) or context information (`Time`) to the classifier.

Fig. 4 presents the performance of sleep detection model depending on the percentage of missing data. The performance of the model to detect sleep is the highest when all sensor data is present. For instance, using the ACC, EDA and TEMP the BA of the GB classifier to detect sleep is 90%. Even a small percentage of missing data in a window (e.g., 10%) causes a reduction of 10 percentage points of the performance of the model. We observe that the multimodal approach (e.g., using ACC, EDA, and TEMP together) and EDA are more robust to missing data in comparison to the ACC and TEMP sensors. We believe the lower impact in the multimodal model is because of the effect of missing data in one data source to be compensated by the data from other data sources. The performance of the model using only ACC or TEMP sensor data decreases as the number of missing data points increases. The presence of missing data is more critical for dynamic signals (e.g., for ACC) rather than the EDA sensor.

Comparison of missing data imputation strategies. We investigate whether the imputation strategies can mitigate the impact of missing data to recognize sleep and wake from wearable sensors. Table I reports the performance, in terms of BA, of the `GB` classifier trained with features extracted from missing data (`Random` or `Chunks`) and replacing missing data with the interpolation- and regression-based imputation strategies. The majority of imputation strategies (e.g., interpolation, regressor, most frequent) perform better than no imputation at all hinting at the need of using such strategies for a more robust sleep and wake recognition. Imputing missing data with a zero value, which is a common procedure, leads to no significant improvement in the classification results. In all cases, except for EDA with a missing data rate of 25%, the interpolation imputation strategy provides the highest performance for distinguishing between sleep and wake classes. Imputing data that misses randomly with interpolation leads to an improvement of around 11, 4, 16 and 10 percentage points for `ACC`, `EDA`, `TEMP` and `All` sensors combined, in comparison to no imputation at all. Overall, the interpolation is comparably effective when data is missing in chunks. This implies that interpolation strategy not only recovers missing data, but also preserves the statistical characteristics of the original time series, which are essentially needed for post-analysis of the signal (e.g., classification task). According to the results from a `t-test` ($p < 0.05$), the classification results using padding strategy are significantly higher than zero and regression-based imputation. Padding is also higher than most frequent strategy in majority of experiments except for TEMP and EDA sensors.

The top five most selected features using the complete data set and the imputed data set (e.g., with 10% of missing data) is the same. The five most frequently selected features in all iterations are the standard deviation of the first derivative of x, y and z axis of the ACC sensor, the maximum value of the x-axis and the mean value of the tonic component of EDA. The most selected features when using the data set with missing data points (e.g., 10% of missing data) is different. The top five features in this case are the difference between the last and first sensor data point in a window for x, y, z and EDA as well as the number of EDA peaks in the window.

Impact of context information and phone sensor data. We investigate the impact of phone sensor data (e.g., amount of light, phone usage, number of notifications clicked) and context information (e.g., hour of the day, day of the week, weekday or weekend) to mitigate the missing data problem in wearable sensors. Fig. 5 presents the classification results of GB classifier to recognize sleep and wake trained with wearable sensor data with a missing rate of 50% (Missing data) as well as adding behavioral features extracted from the phone sensors (Phone) and context information regarding the time (Time). The time information – including the hour of the day, day of the week and type of the day (weekday or weekend) – is the most informative context information for detecting sleep stage. These results suggest that context information could be helpful to maintain the performance of the model to distinguish sleep and wake in case of missing wearable sensor data.

V. Limitations and Future Work

While our results show that it is feasible to use imputation strategies to handle missing data used to detect sleep, further research is needed to overcome the limitations of our work.

The first limitation stems from considering sleep detection as a binary classification problem. Our approach can be used to recognize sleep-wake stages, but not more detailed stages like, e.g., rapid eye movement (REM), non-rapid eye movement (NREM). However, to be able to recognize fine-grained sleep stages, it is first required to have a robust sleep-wake recognition approach, which is the focus of this work. Also, collecting

Authorized licensed use limited to: Universita della Svizzera Italiana. Downloaded on January 27,2023 at 10:42:08 UTC from IEEE Xplore. Restrictions apply.

TABLE I

MEAN (AND STANDARD DEVIATION) ACCURACY OF THE GB CLASSIFIER FOR SLEEP AND WAKE RECOGNITION USING DATA FROM ACC, EDA, TEMP SENSORS AND THE THREE COMBINED. THE PERFORMANCE IS INVESTIGATED WITH MISSING DATA (`None`) AT DIFFERENT RATES (E.G., 10, 25, AND 50%) AND REPLACING MISSING DATA USING THE DIFFERENT IMPUTATION TECHNIQUES (E.G., `Zero`, `Most frequent`, `Interpolation`, `Regressor`).

Missing data rate: 10%

	None		**Zero**		**Most frequent**		**Padding-interpolation**		**Regressor**	
	Random	Chunks	Random	Chunks	Random	Chunks	Random	Chunks	Random	Chunks
ACC	67.8 (9.2)	72.6 (3.3)	67.7 (9.2)	72.5 (3.3)	86.8 (3.6)	87.8 (3.5)	**89.2 (3.6)**	**88.9 (3.8)**	86.7 (3.7)	88.1 (3.7)
EDA	78.5 (6.0)	77.5 (7.3)	78.4 (6.1)	79.3 (6.4)	81.2 (5.9)	81.7 (6.5)	**82.4 (6.3)**	**82.2 (5.9)**	79.0 (5.4)	79.6 (6.4)
TEMP	62.0 (5.7)	64.7 (4.7)	62.1 (7.4)	64.8 (5.9)	77.5 (8.6)	76.5 (4.1)	**78.8 (8.1)**	**78.9 (8.1)**	71.4 (10.8)	68.2 (12.0)
All	79.4 (7.4)	79.5 (6.4)	79.8 (5.7)	80.5 (4.6)	88.3 (4.0)	88.7(8.1)	**89.7 (3.6)**	**89.8 (3.7)**	87.7 (4.5)	88.7 (3.9)

Missing data rate: 25%

	None		**Zero**		**Most frequent**		**Padding-interpolation**		**Regressor**	
	Random	Chunks	Random	Chunks	Random	Chunks	Random	Chunks	Random	Chunks
ACC	65.3 (11.0)	72.5 (3.4)	65.3 (11.1)	72.7 (3.5)	86.7 (3.3)	87.4 (3.4)	**89.0 (3.4)**	**88.6 (3.7)**	86.4 (3.9)	87.4 (4.5)
EDA	78.4 (6.8)	79.0 (6.4)	78.6 (6.5)	78.9 (6.3)	**81.7 (6.1)**	80.5 (6.6)	81.6 (6.6)	**81.5 (6.1)**	77.6 (6.3)	79.7 (6.3)
TEMP	60.0 (8.3)	64.7 (5.3)	60.0 (8.3)	64.8 (5.6)	76.9 (7.6)	76.7 (7.9)	**78.9 (8.1)**	**78.7 (7.7)**	70.9 (7.9)	70.6 (9.6)
All	79.6 (6.1)	80.4 (4.6)	80.1 (5.8)	80.6 (4.7)	87.8 (3.9)	88.4 (4.1)	**89.4 (4.0)**	**89.3 (3.9)**	86.9 (4.4)	88.4 (4.3)

Missing data rate: 50%

	None		**Zero**		**Most frequent**		**Padding-interpolation**		**Regressor**	
	Random	Chunks	Random	Chunks	Random	Chunks	Random	Chunks	Random	Chunks
ACC	58.8 (10.4)	72.6 (3.3)	58.9 (10.5)	72.9 (3.4)	85.2 (3.6)	85.4 (3.8)	**89.1 (3.6)**	**87.7 (3.1)**	85.5 (4.2)	86.1 (3.6)
EDA	77.9 (5.5)	78.8 (6.3)	77.8 (5.7)	78.3 (6.7)	78.8 (6.8)	79.0 (7.1)	**81.3 (6.3)**	**80.5 (6.4)**	71.5 (6.9)	76.0 (9.4)
TEMP	54.5 (5.8)	64.8 (4.5)	54.6 (6.0)	64.7 (4.6)	76.8 (9.1)	77.0 (8.5)	**78.8 (7.8)**	**77.9 (8.0)**	68.1 (11.0)	68.1 (10.7)
All	78.6 (4.9)	79.7 (6.0)	78.3 (4.8)	80.2 (5.2)	87.1 (3.9)	87.2 (3.1)	**89.4 (3.8)**	**88.8 (3.6)**	85.8 (6.2)	87.1 (4.2)

ground-truth data for detailed sleep stages requires the use of cumbersome sensors, which might interfere with the behavior of users in everyday life settings [19, 21]. Further, we explored the impact of missing data using only the GB classifier. While GB has shown to perform best in similar classification tasks [11, 24, 41], in future work, we plan to investigate the impact of missing data for other types of classifiers (e.g., deep neural networks) and the generalizability of our approach to other classification problems. Lastly, our approach has been tested with data of healthy participants and these findings might not generalize to other types of populations (e.g., people with sleep disorders). Further research is needed to understand the generalizability of our approach to different populations.

VI. CONCLUSIONS

In this work we investigate the impact of missing data for recognizing sleep and wake using wearable sensor data. Our results show that the presence of missing data significantly decreases the accuracy of gradient boosting classifier by 10-35 percentage points. To mitigate this problem, we use interpolation- and regression-based imputation strategies. Our results show that the model's performance to recognize sleep and wake is higher by 4-30 percentage points when using imputation strategies. We believe these results open up new opportunities for designing and developing robust sleep monitoring systems to provide continuous feedback to the user and to maintain their accuracy in case of missing sensor data.

ETHICAL IMPACT STATEMENT

While the work presented in this paper provides insights about the impact of missing data in sleep monitoring systems, the deployment of such systems can raise significant concerns about users' privacy. In this work, we have followed the ethical approval process at our institution to maintain participants' privacy. Our long-term goal is to design and develop sleep monitoring systems that can support people in everyday life in a privacy-preserving manner.

ACKNOWLEDGEMENT

This work is partially supported by the Swiss National Science Foundation (SNSF) through the grant 205121_197242 for the project "PROSELF: Semi-automated Self-Tracking Systems to Improve Personal Productivity".

REFERENCES

[1] M. Altini and H. Kinnunen, "The Promise of Sleep: A Multi-Sensor Approach for Accurate Sleep Stage Detection Using the Oura Ring," *Sensors*, vol. 21, no. 13, p. 4302, 2021.

[2] M. Borazio and K. Van Laerhoven, "Predicting Sleeping Behaviors in Long-term Studies With Wrist-Worn Sensor Data," in *International Joint Conference on Ambient Intelligence*. Springer, 2011, pp. 151–156.

[3] F. Wahl and O. Amft, "Data and Expert Models for Sleep Timing and Chronotype Estimation from Smartphone Context Data and Simulations," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, vol. 2, no. 3, pp. 1–28, 2018.

[4] E. D. Chinoy, J. A. Cuellar, K. E. Huwa, J. T. Jameson, C. H. Watson, S. C. Bessman, D. A. Hirsch, A. D. Cooper, S. P. Drummond, and R. R. Markwald, "Performance of Seven Consumer Sleep-tracking Devices Compared with Polysomnography," *Sleep*, vol. 44, no. 5, 2021.

[5] H. Kang, "The Prevention and Handling of the Missing Data," *Korean Journal of Anesthesiology*, vol. 64, no. 5, p. 402, 2013.

[6] A. Saeed, T. Ozcelebi, and J. Lukkien, "Synthesizing and Reconstructing Missing Sensory Modalities in Behavioral Context Recognition," *Sensors*, vol. 18, no. 9, p. 2967, 2018.

[7] H. Rashid, S. Mendu, K. E. Daniel, M. L. Beltzer, B. A. Teachman, M. Boukhechba, and L. E. Barnes, "Predicting Subjective Measures of Social Anxiety from Sparsely Collected Mobile Sensor Data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, vol. 4, no. 3, 2020.

[8] H. Luo, P.-A. Lee, I. Clay, M. Jaggi, and V. De Luca, "Assessment of Fatigue Using Wearable Sensors: A Pilot Study," *Digital Biomarkers*, vol. 4, no. 1, pp. 59–72, 2020.

Authorized licensed use limited to: Universita della Svizzera Italiana. Downloaded on January 27,2023 at 10:42:08 UTC from IEEE Xplore. Restrictions apply.

[9] T. Hossain, M. Ahad, A. Rahman, and S. Inoue, "A Method for Sensor-Based Activity Recognition in Missing Data Scenario," *Sensors*, vol. 20, no. 14, p. 3811, 2020.

[10] T. Feng and S. Narayanan, "Imputing Missing Data in Large-Scale Multivariate Biomedical Wearable Recordings Using Bidirectional Recurrent Neural Networks With Temporal Activation Regularization," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 2529–2534.

[11] S. Gashi, E. Di Lascio, B. Stancu, V. D. Swain, V. Mishra, M. Gjoreski, and S. Santini, "Detection of Artifacts in Ambulatory Electrodermal Activity Data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, vol. 4, no. 2, pp. 1–31, 2020.

[12] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multimodal Autoencoder: A Deep Learning Approach to Filling in Missing Sensor Data and Enabling Better Mood Prediction," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 202s–208.

[13] A. Burkov, *The Hundred-page Machine Learning Book*. Andriy Burkov Quebec City, Can., 2019.

[14] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

[15] J. Di, C. Demanuele, A. Kettermann, F. I. Karahanoglu, J. C. Cappelleri, A. Potter, D. Bury, J. M. Cedarbaum, and B. Byrom, "Considerations to Address Missing Data When Deriving Clinical Trial Endpoints From Digital Health Technologies," *Contemporary Clinical Trials*, vol. 113, p. 106661, 2022.

[16] S. Gashi, E. Di Lascio, and S. Santini, "Multi-class Multi-label Classification for Cooking Activity Recognition," in *Human Activity Recognition Challenge*. Springer, 2021, pp. 75–89.

[17] K. Saha, M. D. Reddy, V. das Swain, J. M. Gregg, T. Grover, S. Lin, G. J. Martinez, S. M. Mattingly, S. Mirjafari, R. Mulukutla, K. Nies, P. Robles-Granda, A. Sirigiri, D. W. Yoo, P. Audia, A. T. Campbell, N. V. Chawla, S. K. D'Mello, A. K. Dey, K. Jiang, Q. Liu, G. Mark, E. Moskal, A. Striegel, and M. de Choudhury, "Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019.

[18] A. Sano and R. W. Picard, "Comparison of Sleep-Wake Classification Using Electroencephalogram and Wrist-Worn Multi-Modal Sensor Data," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014.

[19] A. Sano, W. Chen, D. Lopez-Martinez, S. Taylor, and R. W. Picard, "Multimodal Ambulatory Sleep Detection Using LSTM Recurrent Neural Networks," *IEEE Journal of Biomedical and Health Informatics*, 2018.

[20] I. Perez-Pozuelo, B. Zhai, J. Palotti, R. Mall, M. Aupetit, J. M. Garcia-Gomez, S. Taheri, Y. Guan, and L. Fernandez-Luque, "The Future of Sleep Health: A Data-driven Revolution in Sleep Science and Medicine," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–15, 2020.

[21] B. Zhai, I. Perez-Pozuelo, E. A. Clifton, J. Palotti, and Y. Guan, "Making Sense of Sleep: Multimodal Sleep Stage Classification in a Large, Diverse Population Using Movement and Cardiac Sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, vol. 4, no. 2, pp. 1–33, 2020.

[22] J.-K. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. I. Hong, "Toss'N'Turn: Smartphone as Sleep and Sleep Quality Detector," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2014, pp. 477–486.

[23] C.-Y. Hsu, A. Ahuja, S. Yue, R. Hristov, Z. Kabelac, and D. Katabi, "Zero-effort In-home Sleep and Insomnia Monitoring Using Radio Signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, vol. 1, no. 3, pp. 1–18, 2017.

[24] S. Gashi, L. Alecci, E. Di Lascio, M. E. Debus, F. Gasparini, and S. Santini, "The Role Of Model Personalization for Sleep Stage and Sleep Quality Recognition Using Wearables," *IEEE Pervasive Computing*, 2022.

[25] M. Garbarino, M. Lai, D. Bender, R. W. Picard, and S. Tognetti, "Empatica E3–A Wearable Wireless Multi-sensor Device for Real-time Computerized Biofeedback and Data Acquisition," in *Proceedings of the International Conference on Wireless Mobile Communication and Healthcare (MobiHealth 2014)*, 2014.

[26] A. Mehrotra, F. Tsapeli, R. Hendley, and M. Musolesi, "MyTraces: Investigating Correlation and Causation Between Users' Emotional States and Mobile Phone Interaction," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, vol. 1, no. 3, pp. 1–21, 2017.

[27] D. J. Buysse, C. F. Reynolds III, T. H. Monk, S. R. Berman, and D. J. Kupfer, "The Pittsburgh Sleep Quality Index: A New Instrument for Psychiatric Practice and Research," *Psychiatry research*, vol. 28, no. 2, pp. 193–213, 1989.

[28] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell, "Unobtrusive Sleep Monitoring Using Smartphones," in *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. IEEE, 2013, pp. 145–152.

[29] J. Hernandez, I. Riobo, A. Rozga, G. D. Abowd, and R. W. Picard, "Using Electrodermal Activity to Recognize Ease of Engagement in Children During Social Interactions," in *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2014)*, 2014.

[30] E. Di Lascio, S. Gashi, and S. Santini, "Unobtrusive Assessment of Students' Emotional Engagement During Lectures Using Electrodermal Activity Sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, vol. 2, no. 3, 2018.

[31] W. Boucsein, *Electrodermal Activity*. Springer Science & Business Media, 2012.

[32] Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures, W. Boucsein, D. C. Fowles, S. Grimnes, G. Ben-Shakhar, W. T. Roth, M. E. Dawson, and D. L. Filion, "Publication recommendations for electrodermal measurements," *Psychophysiology*, vol. 49, no. 8, pp. 1017–1034, 2012.

[33] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 797–804, 2015.

[34] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard, "Automatic Identification of Artifacts in Electrodermal Activity Data," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015.

[35] D. Figo, P. C. Diniz, D. R. Ferreira, and J. M. Cardoso, "Preprocessing Techniques for Context Recognition from Accelerometer Data," *Personal and Ubiquitous Computing*, vol. 14, no. 7, pp. 645–662, 2010.

[36] K. Asahina and K. Omura, "Phenomenological Study of Paradoxical Phase and Reverse Paradoxical Phase of Sleep," *The Japanese Journal of Physiology*, vol. 14, no. 4, pp. 365–372, 1964.

[37] N. R. Burch, "Data Processing of Psychophysiological Recordings (Discussant: Harold W. Shipton)," *NASA Special Publication*, vol. 72, p. 165, 1965.

[38] A. Sano and R. W. Picard, "Quantitative Analysis of Electrodermal Activity During Sleep," *Sleep*, vol. 1, no. 2Q, p. 3Q, 2012.

[39] T. Chen and C. Guestrin, "Xgboost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*. ACM, 2016, pp. 785–794.

[40] F. Chollet, *Deep Learning with Python*. Simon and Schuster, 2017.

[41] M. Gjoreski, V. Janko, G. Slapničar, M. Mlakar, N. Reščič, J. Bizjak, V. Drobnič, M. Marinko, N. Mlakar, M. Luštrek *et al.*, "Classical and Deep Learning Methods for Recognizing Human Activities and Modes of Transportation With Smartphone Sensors," *Information Fusion*, vol. 62, pp. 47–62, 2020.

[42] T. Plötz, "Applying Machine Learning for Sensor Data Analysis in Interactive Systems: Common Pitfalls of Pragmatic Use and Ways to Avoid Them," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–25, 2021.

[43] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[44] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The Balanced Accuracy and Its Posterior Distribution," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 3121–3124.

[45] A. Saeed, T. Ozcelebi, S. Trajanovski, and J. J. Lukkien, "End-to-end Multi-Modal Behavioral Context Recognition in a Real-Life Setting," in *2019 22th International Conference on Information Fusion (FUSION)*. IEEE, 2019, pp. 1–8.

Authorized licensed use limited to: Universita della Svizzera Italiana. Downloaded on January 27,2023 at 10:42:08 UTC from IEEE Xplore. Restrictions apply.