Making Sense of Sleep: Multimodal Sleep Stage Classification in a Large, Diverse Population Using Movement and Cardiac Sensing

BING ZHAI*, Newcastle University, UK
IGNACIO PEREZ-POZUELO*, University of Cambridge & The Alan Turing Institute, UK
EMMA A.D. CLIFTON, University of Cambridge, UK
JOAO PALOTTI, Massachusetts Institute of Technology, USA
YU GUAN, Newcastle University, UK

Traditionally, sleep monitoring has been performed in hospital or clinic environments, requiring complex and expensive equipment set-up and expert scoring. Wearable devices increasingly provide a viable alternative for sleep monitoring and are able to collect movement and heart rate (HR) data. In this work, we present a set of algorithms for sleep-wake and sleep-stage classification based upon actigraphy and cardiac sensing amongst 1,743 participants. We devise movement and cardiac features that could be extracted from research-grade wearable sensors and derive models and evaluate their performance in the largest open-access dataset for human sleep science. Our results demonstrated that neural network models outperform traditional machine learning methods and heuristic models for both sleep-wake and sleep-stage classification. Convolutional neural networks (CNNs) and long-short term memory (LSTM) networks were the best performers for sleep-wake and sleep-stage classification, respectively. Using SHAP (SHapley Additive exPlanation) with Random Forest we identified that frequency features from cardiac sensors are critical to sleep-stage classification. Finally, we introduced an ensemble-based approach to sleep-stage classification, which outperformed all other baselines, achieving an accuracy of 78.2% and F_1 score of 69.8% on the classification task for three sleep stages. Together, this work represents the first systematic multimodal evaluation of sleep-wake and sleep-stage classification in a large, diverse population. Alongside the presentation of an accurate sleep-stage classification approach, the results highlight multimodal wearable sensing approaches as scalable methods for accurate sleep-classification, providing guidance on optimal algorithm deployment for automated sleep assessment. The code used in this study can be found online at: https://github.com/bzhai/multimodal_sleep_stage_benchmark.git

CCS Concepts: • **Human-centered computing->Ubiquitous and mobile computing design and evaluation methods**;

Additional Key Words and Phrases: Sleep, sleep Stage, Actigraphy, Heart Rate, Heart Rate Variability, Multimodal Sensing, Multistage Classification, Neural Networks

1 INTRODUCTION

Sleep is a reversible physiological state that is essential for life, health and performance. Whilst the functions of sleep are not yet fully understood, it is known to restore energy, promote healing, rejuvenate physical systems, interact with the immune system and influence brain function, with manifold consequences including for memory consolidation and behaviour [16, 23, 42, 43]. As a result of its importance to vital human processes and the incomplete understanding of its function, accurate sleep monitoring is of interest to the understanding of human health and an active area of research for the *ubiquitous computing* community [1, 11, 34, 48].

*Both authors contributed equally to this research.

Authors' addresses: Bing Zhai, b.zhai2@newcastle.ac.uk, Newcastle University, Open Lab, Urban Sciences Building, Newcastle upon Tyne, UK; Ignacio Perez-Pozuelo, ip325@cam.ac.uk, University of Cambridge & The Alan Turing Institute, Department of Medicine, Cambridge, UK; Emma A.D. Clifton, emma.clifton@mrc-epid.cam.ac.uk, University of Cambridge, Department of Medicine, Cambridge, UK; Joao Palotti, palotti@mit.edu, Massachusetts Institute of Technology, CSAIL, Cambridge, USA; Yu Guan, Yu.Guan@newcastle.ac.uk, Newcastle University, Open Lab, Urban Sciences Building, Newcastle upon Tyne, UK.

2020. XXXX-XXXX/2020/6-ART0 $0.00
https://doi.org/xx.xxxxx/xxxxxxx

Traditionally, human sleep has been monitored in laboratory settings using polysomnography (PSG). PSG is a multi-sensor approach to monitoring, involving the collection and conveyance of a range of different signals from many sensors operating simultaneously. These sensors include electroencephalography (EEG), electromyography (EMG) and electrooculography (EOG), which together facilitate the measurement of brain activity, alongside both muscle and eye movement. Measurements of respiratory and cardiac activity are also often included. PSG recordings are processed and segmented into epochs, typically of 30-second duration, and an expert or technician then assigns a sleep stage to each epoch. This sleep stage "scoring" typically follows the rules set out by the American Association of Sleep Medicine (AASM), which defines five stages of sleep-wake cycles: wake (W), rapid eye movement (REM) sleep and three types of non-REM sleep (NREM), known as N1, N2 and N3 [7].

Whilst traditional PSG is considered the gold-standard for sleep monitoring, as a result of the need for sensing equipment, its use is limited to laboratory settings and typically to just one or two nights. These single nights of observed sleep in an unfamiliar environment may not reflect normal sleep. Further, it is impractical to measure sleep using this method for more than two consecutive nights as it is burdensome to patients or study participants. PSG is also expensive and requires expert set-up and analysis. For these reasons, efforts to monitor individual's typical sleep duration and quality longitudinally in large, free-living populations have generally relied upon sleep diaries or self-reported questionnaire data. Whilst sleep diaries are cost-effective, scalable and able to collect information regarding typical sleep patterns, there are concerns as to the validity and reliability of participant responses [17]. Wearable sensors offer a potential solution. Such sensors provide valuable, unobtrusive tools through which to objectively monitor physical activity in large population studies, with potential applications for sleep monitoring.

Conventional approaches to monitoring sleep using wearable devices are primarily based on actigraphy (count-based movement information) and accelerometry (raw, high frequency data which is often triaxial) [12, 39–41]. However, recent technological and battery life advances increasingly facilitate multimodal sensing (i.e., combining accelerometry with HR sensing). Multimodal sensing facilitates more intricate human activity recognition (HAR) tasks and has shown promise for sleep-stage classification [27]. The validity of actigraphy for the classification of sleep-wake transitions has been demonstrated over the past three decades [12, 39–41]. Algorithms applied to actigraphy for this purpose exploit differences in body movement between wakefulness and sleep. Recent work has demonstrated how different methods for binary sleep-wake classification using actigraphy compare when applied to the same, standardized dataset [32]. Furthermore, HR variability (HRV) metrics could be valuable for multistage classification as autonomic function fluctuations occur between non-REM sleep and waking/REM sleep, whilst these same functions are consistent when comparing wake to REM [2, 9, 15, 53].

Understanding time spent in different sleep stages (beyond binary sleep-wake classification) in free-living environments has important implications for commercial applications, as well as for research. For example, accurate sleep architecture inferences may provide better information to guide sleep-related behavioural changes and recommendations [13]. PSG is the gold-standard for sleep stage assessment, but is not scalable to large, population-based studies of free-living individuals with the power to make inferences regarding the implications of sleep for health and illness. Wearable devices are a potential inexpensive and scalable solution to monitor sleep in large populations. Limited literature exists regarding the performance of multimodal sensing using wearable technologies in sleep-wake and sleep-stage classification [56]. The development of a set of benchmarks to evaluate the performance of sleep-wake and sleep-stage classification methods on multimodal data using movement and cardiac sensing would address a major gap in the existing literature.

In order to address this gap, this work focused on five major contributions:

(1) We introduce a framework for pre-processing and analyzing multimodal sensor data from movement (actigraphy) and cardiac (RR intervals from ECG) sensors. We use the same signals that are derivable from research-grade ECG or photoplethysmogram (PPG) devices.

(2) We systematically compare single modality to combined sensing (actigraphy + HR/HRV) approaches for classifying sleep-wake using different machine learning models.
(3) We extend this systematic comparison to explore the performance of single modality approaches and combined sensors across three different multistage classification tasks: (A) Conventional three-stage classification (NREM, REM, wake), (B) Four-stage classification (light sleep, deep sleep, REM and wake) and (C) Five-stage classification (AASM-standard), also using traditional machine learning (ML) and deep learning (DL) models.
(4) We introduce an easy-to-interpret evaluation metric, namely, time deviation, which aims to be accessible to sleep practitioners. We also study the modality/feature importance by using Random Forest with SHAP, yielding some interesting findings (e.g., high frequency HRV is the most important feature in recognizing REM sleep).
(5) We introduce an ensemble structure for multistage classification of sleep based on multi-timescale and multimodal DL ensembles. This architecture aims to exploit the individual contributions and strengths of different classifiers. This approach can significantly improve the performance of the three-stage sleep classification task.

Our work presents a systematic multimodal and multistage evaluation of sleep-wake cycles and sleep stages in a large, diverse population. We examine each individual method and modality, as well as exploring how sensor fusion leads to improved performance. Additionally, we explore the features that contribute the most to the different classification tasks, developing an understanding of the physiological underpinnings of our models.

2 RELATED WORK

Since the 1980s, a vast number of studies have explored new methods and techniques to infer sleep-wake cycles using actigraphy with either single-axial [12, 40, 41] or, more recently, tri-axial accelerometry [20].While these methods have proven valuable, they were often derived in small cohorts or by using non-clinical grade equipment or sleep diaries. The recent availability of large datasets, provided by initiatives such as the National Sleep Research Resource [14, 57][1], makes it possible for researchers to create large standardized benchmarks, such as that proposed in our work. For example, Palotti et al. leveraged one of the available datasets, the Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Study[2], to compare the performance of the most relevant heuristic approaches and ML methods for binary sleep-wake classification [32]. Whilst novel, their work was limited by: (1) exclusively comparing methods for sleep-wake classification, rather than multistage classification; (2) only using actigraphy data. Here we address these limitations in the MESA Sleep Study dataset, the only dataset suitable for such experiments to-date.

Beyond actigraphy data, in this work, we explore the use of HR and HRV data. HR can be defined as the average number of heartbeats per minute, while HRV is a measure of the variability in beat-to-beat intervals, known as RR intervals. These measurements are powerful biomarkers that have been used to understand training and recovery, address chronic disease and monitor stress and sleep [24, 45, 52]. HRV is typically higher during the night, reflecting the fact that sleep is a state in which vagal activity, characterized by rapid fluctuations in activity controlling coronary artery tone, HR and systolic blood pressure, is dominant [24, 45, 52]. Thus, HRV shows a nocturnal increase in the deviation of mean RR intervals. These deviations also differ between sleep stages. Conversely, several studies have shown that HR does not change significantly between sleep stages, although some work has suggested a rise during REM sleep [22, 46]. HRV analysis has demonstrated that the High Frequency (HF) band doubles in relative power when going from quiet wakefulness to non-REM sleep [53]. Hence, a full-feature set of HRV-relevant features is a powerful tool for sleep-stage classification [10]. Recently,

[1]https://sleepdata.org
[2]https://sleepdata.org/datasets/mesa

Radha et al. have reported that HRV has great potential to classify sleep stages [37]. However, their work was performed in a private dataset and conducted using some features that are often not present on wearable devices. In our work, we devise HR/HRV features that could be extracted from research-grade wearable sensors and evaluate their performance in the largest public dataset to date.

Aside from ML and DL models, ensemble architectures are becoming increasingly prevalent for HAR tasks. For instance, in 2015, Single et al. adopted an approach consisting of three variants of long-short term memory (LSTM) networks that worked in parallel to tackle a biological sequence analysis task and then used *majority voting* to decide upon the final classification prediction [44]. In [18], Guan and Ploetz developed a LSTM ensemble model via epoch-wise bagging for efficient training. They injected several random factors to increase the diversity of the classifiers and improve performance in several HAR tasks. Recent work has explored the application of ensemble models for automatic sleep classification using PSG/EEG signals. These studies have shown promising results, improving the performance of shallow ML and even DL approaches [3, 25, 47]. Koley et al. used an ML ensemble architecture approach consisting of five binary support vector machine (SVM) classifiers to classify different sleep stages [25]. Using a *"winner-takes-all"* ensemble method [30] the researchers managed to extract more discriminant patterns from EEG. Recently, Huy et al. applied an ensemble method to multimodal PSG data (EOG, EEG and electromyography (EMG)) by fusing classifiers [35]. All these previously reported models are based on sleep epoch (30 seconds) level feature extraction protocols and use classifier ensembles derived in sensors which often exceeded 100Hz sampling rates (EEG data, etc). These methods demand high specifications with regards to computing power. Currently, the limited data storage and processing capabilities of wearable devices mean that using methods based on high sampling rate is unlikely to be possible in free-living environments.

3 METHODS

The MESA Sleep dataset is introduced and described in Section 3.1 [14, 57]. All experiments reported here were conducted based in this dataset. In Section 3.2, we provide an overview of the data pre-processing and feature extraction method for modalities which consist of cardiac sensing (HR and HRV) and movement sensing (actigraphy). All tasks explored, including our *ensemble method*, are introduced in Section 3.3. In Section 3.4, we introduce the models used for our benchmark work. Section 3.5 describes how we designed our experiments. Finally, in Section 3.6, the metrics used to evaluate the classification models are discussed.

3.1 Dataset Description

The Multi-Ethnic Study of Atherosclerosis (MESA) dataset is a multi-centre longitudinal study designed to investigate the characteristics of sub-clinical cardiac disease. The study comprises 6814 asymptomatic men and women of black, white, Hispanic and Chinese-American ethnicity, of which 2,237 were also enrolled in the MESA Sleep Study. As part of the MESA Sleep Study, all participants wore an actigraphy device for one week and underwent concurrent PSG for one night. Data for this study was acquired in six different centers across the US and followed the appropriate Institutional Review Board approvals and written informed consent for participant data acquisition [14, 57].

The MESA Sleep Study was conducted using a Compumedics Somte System for PSG, which includes the ECG signals here used to derive HR and HRV and their associated features, alongside an Actiwatch Spectrum from Philips Respironics to record actigraphy data. This device captures measurements of movements defined as "activity counts"[3] and aggregates them into 30 second epochs. The Actiwatch was securely fastened to participant's non-dominant wrist. These actigraphy signals and their associated features can be derived in most research-grade wearable devices. The sensors for the Compumedics PSG comprised: cortical EEG, bilateral EOG, chin EMG, abdominal and thoracic respiratory inductance plethysmography, airflow, ECG, leg movement sensor and finger

[3] https://www.salusa.se/Filer/Produktinfo/Aktivitet/TheActiwatchUserManualV7.2.pdf

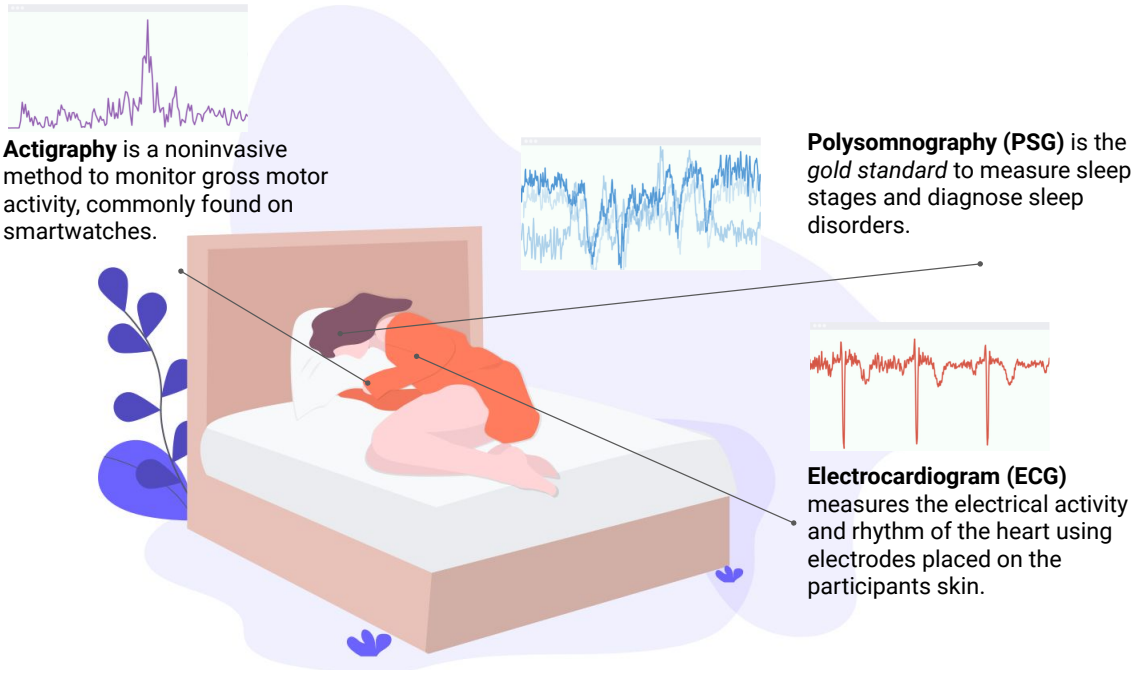


Fig. 1. Experimental setup and tasks: Our models are trained using a combined-sensing, multimodal approach which incorporates two time-series signals: actigraphy and ECG derived HR and HRV and uses Gold-Standard PSG for the training labels

pulse oximetry. These sensors collected three types of signals: bioelectrical potentials (EEG, EOG, EMG, ECG), waveforms received from transducers (thermistors on the airflow devices, inductance respiratory bands, piezo leg sensors and position sensors from the leg device) and auxiliary devices (oximetry measures of oxyhemoglobin saturation and nasal pressure records). Full details of the setup, protocol and sampling rates are available [4,5]. All participants included in our study had at least one full night of PSG recording with concurrent actigraphy and ECG. An illustration of the experimental set up is provided in Figure 1. All nocturnal recordings were transmitted to a centralized reading center at the Brigham and Women's Hospital (Boston, MA, USA) and data was scored by trained technicians using AASM guidelines. For our training labels, we used the expert scoring and epoch staging annotations on PSG data provided by Bild et al [8]. Note that the MESA Sleep dataset is the **only large open-access dataset** combining gold-standard measures of sleep through PSG with wearable sensor data from actigraphy as well as ECG (HR/HRV) and thus the only existing dataset appropriate for our purposes.

Table 1 summarizes the main demographic characteristics of the participants by training and test splits.

3.2 Data Pre-processing and Feature Extraction

In this work, we synchronized PSG, ECG and actigraphy records into 30-second sleep epochs for 1,743 of the 2,237 participants included in the study. A total of 494 participants were excluded on the basis of: (1) lack of concurrent PSG, ECG and actigraphy data; (2) lack of enough quality standard data (< 1.5 h of usable data

[4]https://sleepdata.org/datasets/mesa/pages/equipment/montage-and-sampling-rate-information.md
[5]https://sleepdata.org/datasets/mesa/files/documentation

Table 1. Breakdown of population based on sex, age and demographic characteristics, by dataset (training or test).

Dataset	Total	Female	Male	Black	Chinese-American	White	Hispanic	Age ($\mu \pm \sigma$)	Min Age	Max Age
Training	1395	752 (54%)	643 (46%)	383 (28%)	153 (11%)	511 (37%)	348 (25%)	69.29 ± 8.73	54	94
Test	348	198 (57%)	150 (43%)	103 (30%)	39 (11%)	128 (37%)	78 (22%)	68.52 ± 9.21	55	89

Numbers are N, N(%) or mean (SD). Age is given in years.

Table 2. Sleep statistics of participants in the study.

Dataset	Total Sleep Time (TST)	Total Time in Bed (TIB)	Sleep Efficiency (%)	Wake After Sleep Onset (WASO)	N1	N2	N3	REM
All	359.0±80.7	475.0±85.3	76±13.1	90.3±62.7	49.42±30.9	207.0±60.1	40.0±33.4	66.9±28.9
Training	357.5±80.2	473.5±84.9	75.9±13.1	90.5±62.7	49.3±31.1	206.6±60.4	39.6±33.3	66.6±29.3
Test	365.1±82.5	480.9±87.0	76.4±12.8	89.2±62.6	49.8±30.4	208.5±58.9	41.68±33.7	68.3±27.4

Numbers are minutes except sleep efficiency measured in percentage(mean ± SD)

from the concurrent three sensing methods); or (3) lack of data integrity or misalignment of data, we removed actigraphy outlier epochs based on human expert annotations. These outliers are either non-wearing periods or equipment failure periods. For actigraphy epochs labeled as outliers, their corresponding HR/HRV epochs were also removed [54].

For generalisability purposes, we included diseased participants in our analysis; full details are presented in Supplementary Table S1. Similarly, we did not exclude a total of 30 subjects (about 2% of the total cohort) who do not have any REM epochs at all, although we do understand that these sleep patterns are physiologically very unlikely. The sleep stages for subjects in this dataset were scored by individual sleep technicians, blind to the disease status of the participants, into five classes (wake, N1, N2, N3, REM) according to AASM guidelines [8].

For the ECG signal, we derived features that are only based on RR intervals instead of using the raw ECG signal. The rationale behind this was to make our work as transferable as possible to data collected from research-grade devices such as miniaturized ECGs or wrist wearables that incorporate PPG sensors (i.e., the Empatica E4 wristband). Participants whose ECG records did not include a full night of sleep, or whose data was corrupted were excluded from further analysis.

QRS complexes (R-points) were detected using Compumedics Somte (Abbotsford, VIC, Australia) software Version 2.10 (Builds 99 to 101). The R-points were classified as normal sinus, supraventricular premature complex or ventricular premature complex. Data cleaning, filtering and noise removal took place during this step of the process using the Python package HRV-analysis[6]. First, RR interval outlier data was filtered using a threshold method with a range between 300 to 2000 ms following the method previously described by Tanaka et al. [49], then the ectopic beats were removed by through the methods described in Malik et al. [29]. Second, we linearly interpolated the removed R-points. We grouped the RR intervals into 30 seconds to match the time interval of actigraphy data. Recalling the description in Section 2, the HRV describes the physiological variation of the beat-to-beat interval that can be extracted from the time-distance between adjacent R wave peaks. Thus, we calculated 30 cardiac features from each 30-second window that matches the epoch of the actigraphy data. Following the approach used by Radha et al. [37], we extracted features in four domains (time, geometrical, frequency and non-linear domains). Table 4 details the full set of cardiac features used in this work.

We adopted two strategies for extracting actigraphy-related features. For DL approaches, which can automatically extract high-level features, we used activity counts as input, which can be directly extracted from the device (at a sampling rate of 1/30 Hz). For the other ML models, a total of 370 handcrafted time-series features were extracted, as described below. These features have been commonly used in the literature (i.e., [26, 32, 51]).

[6]https://pypi.org/project/hrv-analysis/

Table 3. Full set of features extracted from the actigraphy signal.

Feature name	Description
Activity Count 🏃	*Raw activity count from the actigraphy device*
Log Activity Count 🏃	*Natural Logarithm of the activity count*
Mean Activity 🏃	*Mean value for the window of activity of size N.* $1 \leq N < 20$
Median Activity 🏃	*Median value for the window of activity of size N.* $1 \leq N < 20$
Std Activity 🏃	*Standard deviation value for the window of activity of size N.* $1 \leq N < 20$
Variance Activity 🏃	*Variance value for the window of activity of size N.* $1 \leq N < 20$
Minimum Activity 🏃	*Minimum value for the window of activity of size N.* $1 \leq N < 20$
Maximum Activity 🏃	*Maximum value for the window of activity of size N.* $1 \leq N < 20$
NAT Activity 🏃	*Number of epochs, in a window of size N, which the value for the activity count is larger than 50 and lower than 100. Devised from [40].* $1 \leq N < 20$
Any Activity 🏃	*Number of epochs that contain any activity in the window of size N.* $1 \leq N < 20$
Skewness of Activity 🏃	*Skewness for the window of activity of size N.* $4 \leq N < 20$
Kurtosis of Activity 🏃	*Kurtosis for the window of activity of size N.* $4 \leq N < 20$

For each sleep epoch T, we calculated the statistics (i.e., mean, variance, median, kurtosis) for actigraphy data that consider both centered and non-centered sliding windows of N sleep epochs (with $N = \{1, 2, ..., 19\}$), where each sleep epoch contains a scalar value. We also calculated other commonly used metrics, such as the raw and natural logarithm values of the activity counts for each epoch T. These features are listed in Table 3.

Our full feature set (i.e., both activity and cardiac features) were normalized using the z-score method. A summary of the pipeline used in this work is shown in Figure 2.

3.3 Tasks

We structured our work on 5 different tasks that allowed us to explore the objectives of this paper and test several hypotheses based on multimodal fusion and new model development.

Our first task, **Task 1**, aims to establish benchmarks for sleep-wake (binary) classification using single modality (either actigraphy or HR/HRV) and multimodality approaches (combining both modalities). In doing so, we compare conventional statistical learning methods and simple neural network methods across modalities. This task is the most explored one among the research community in this area [12, 26, 33, 39, 40] and we also aimed to augment the benchmarks previously reported by Palotti et al. [32].

Task 2 consisted of the same systematic evaluation, but this time, the simplest sleep staging paradigm was introduced (Wake, NREM, REM). Here, the AASM scores provided in the MESA dataset are simplified and collapsed into a simpler representation of sleep staging. Wake and REM remain the same, but N1, N2 and N3 are grouped together to become NREM sleep as an entity. The feasibility of this task has also been tested by other studies [37, 56].

Taking a step further in the level of granularity, **Task 3** classifies the data into Wake, REM, Light Sleep and Deep Sleep. Here, light sleep captured both N1 and N2 which is often considered a transition state between light and deep sleep and usually takes up the largest percentage of time during a full sleep cycle [4]. Given the heterogeneity and prevalence of N2, the difficulty of the task has risen significantly. The models are expected to perform worse than they did on previous tasks.

Task 4 explored the classification of sleep stages based on AASM rules (Wake, REM, N1, N2, N3). This task has the highest level of granularity, and it is, in fact, a task in which even the current state of the art, DL approaches on gold-standard PSG recordings often do not achieve satisfactory performance [47]. This task faces two challenges.

Table 4. Full set of cardiovascular related features grouped by domain.

Time Domain Features	
Mean HR ♥	*Mean heart rate for that window*
Maximum HR ♥	*Maximum heart rate for that window*
Minimum HR ♥	*Minimum heart rate for that window*
Std HR ♥	*Standard deviation for the heart rate for that window*
SDNN ♥	*Standard deviation of Normal-to-Normal interval (NNi)*
SDSD ♥	*Standard deviation of NNi differences*
NN50 ♥	*Number of NNi differences greater 50ms*
pNN50 ♥	*Ratio between NN50 and total number of NNi*
NN20 ♥	*Number of NNi differences greater 20ms*
pNN20 ♥	*Ratio between NN20 and total number of NNi*
RMSSD ♥	*Root mean of squared NNi differences*
Median NNi ♥	*Median of NNis*
Range NNi ♥	*Range between smallest RR intervals to largest RR intervals*
CVSD ♥	The coefficient of variation of successive differences , the RMSSD divided by mean NNi
Coeff. of Variation of NNI ♥	*The Coefficient of Variation of NNi, i.e. the ratio of sdNN divided by mean NNi*
Geometrical Domain Features	
Triangular Index ♥	*The HRV triangular index measurement is the integral of the density distribution (that is, the number of all RR intervals) divided by the maximum of the density distribution (class width of 8ms)*
Frequency Domain Features	
Low Frequency ♥	*Low Frequency is the variance (i.e., power) in HRV in the Low Frequency (.04 to .15 Hz). Reflects a mixture of sympathetic and parasympathetic activity*
High Frequency ♥	*High Frequency is the variance (i.e., power) in HRV in the High Frequency (.15 to .40 Hz). Reflects fast changes in beat-to-beat variability due to parasympathetic (vagal) activity*
Variance in Low Freq. ♥	*VLF is the variance (i.e., power) in HRV in the Very Low Frequency (.003 to .04 Hz). Reflect an intrinsic rhythm produced by the heart which is modulated by primarily by sympathetic activity*
Low/High Freq. Ratio ♥	*The LF/HF ratio is sometimes used by some investigators as a quantitative mirror of the sympathy/vagal balance*
Norm. Low Freq. Ratio ♥	*Normalized low frequency ratio calculated from the raw values of low frequency band (LF or HF) divided by the total spectral power*
Norm. High Freq. Ratio ♥	*Normalized high frequency ratio calculated from the raw values of high frequency band (LF or HF) divided by the total spectral power*
Mean NNi ♥	*Mean over the RR intervals*
Total Power ♥	*Total power of the density spectral*
Non-Linear Domain Features	
Cardiac Sympathetic IdNx ♥	*Cardiac Sympathetic Index [36]*
Mod. Cardiac Symp. IdNx ♥	*A modified cardiac sympathetic index calculated by* $\frac{SD2^2}{SD1}$
Cardiac Vagal IndeNx ♥	*Cardiac Vagal IndeNx [36]*
SD1 ♥	*Poincaré plot standard deviation perpendicular the line of identity*
SD2 ♥	*Poincaré plot standard deviation along the line of identity*
SD1/SD2 Ratio ♥	*Ratio of SD1 to SD2*

The first being the class imbalance, as N1 and N3 sleep epochs account for only 11% and 7% of the data respectively. The second challenge is the nature of our modalities that do not capture direct cortical signals, compromising the performance in more granular classification tasks.

Finally, we introduced an *ensemble method* which aims to combine the unique *perspectives* and *capabilities* of DL classifiers with different window sizes containing discriminant power from different temporal dependencies that could be characteristic of different sleep stages.

3.4 Models and Settings

Conventional heuristic approaches have been readily used in the past 30 years for Task 1 (binary sleep-wake classification). It has recently been shown that feature-based ML and DL approaches greatly outperform all these methods [32].

ML and DL techniques are increasingly used in medical sciences [47, 56]. Here, we use supervised learning techniques on time-series data. This entails generating models that learn mappings between input and output

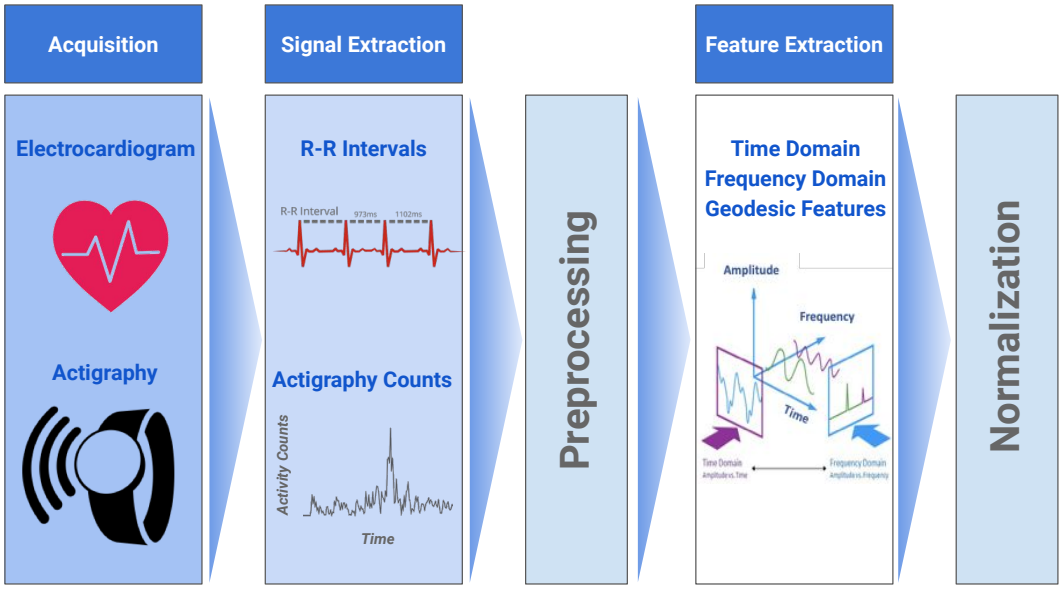


Fig. 2. Multimodal data processing pipeline: after removing low quality data, the signals from the actigraphy device and ECG are synchronized and features are extracted and normalized.

spaces. For instance, Random Forest (RF) approaches have shown strong performance on activity recognition tasks [5]. Similarly, Radu et al. [38] showed promising results using DL approaches on multimodal sensor data for activity and context recognition tasks. Indeed, wearable sensors exploiting multimodal approaches have shown the advantages of this methods over single modality approaches for human activity recognition tasks [19]. Going beyond traditional activity recognition tasks, ML and DL models have been shown to outperform conventional heuristic approaches for actigraphy sleep-wake classification [32, 51]. DL models have also shown great promise in the automatic classification of sleep stages using EEG or multimodal sensor data [35]. Here, we expand that work to multimodal wearable and minimally obtrusive sensors by systematically evaluating how the most well-established ML and DL models perform when using combined sensing.

For all included tasks and modalities, we explored the most common shallow ML and DL architectures. These comprise linear support vector machines, logistic regression, random forest, perceptrons, convolutional neural networks (CNN) and long-short term memory networks (LSTM). We hypothesised that given the large amounts of data DL models would be better suited. Details on the ML and DL classifier settings can be found in Table 13. Deeper architectures (more layers) are explored in the Appendix C, Tables 8 and 9, but for comparison purposes we only employed single layer architectures in the main results section.

3.5 Experimental Design

Once the feature sets were built for our two input streams, we randomly split the dataset into training and test sets following an 80/20 split where 80% (1,395 subjects) went to the training set and 20% (348 subjects) went to

Table 5. Experiment settings based on input modalities, where l is the window length of the input ($l = \{21, 51, 101\}$), the inputs are for each sleep epoch

Algorithm Type	Modality	Input Dimension	Features Used (Full list on Tables 3 and 4)
ML	*[Actigraphy]*	$x \in \mathbb{R}^{370}$	370 features were derived from Activity Counts
	[HR/HRV]	$x \in \mathbb{R}^{30}$	30 features were derived from RR intervals
	[HR/HRV, Actigraphy]	$x \in \mathbb{R}^{400}$	Concatenation of the two modalities above
DL	*[Actigraphy]*	$x \in \mathbb{R}^{l}$	Activity Counts
	[HR/HRV]	$X \in \mathbb{R}^{l \times 8}$	8 features were derived from RR intervals: Mean NNi, Standard Derivation of RR interval (SDNN), RR interval differences (SDSD), Very Low Frequency, Low Frequency, High Frequency Bands, Low Frequency to High Frequency Ratio and Total Power.
	[HR/HRV, Actigraphy]	$X \in \mathbb{R}^{l \times 9}$	Concatenation of the two modalities above

the test set. More details including demographic information can be found in Table 1 and a summary of sleep statistics is introduced in Table 2. .

The inputs to our single modality and multimodal experiments can be found in Table 5. When using multimodal approaches, we used a *channel-wise* stacking approach prior to inputting the resulting matrix into our models. These methods were adopted across all benchmarks for our tasks. Following the method used in [32], our hyperparameter search is described below:

- **ML hyperparameter search:** we employed 5-fold cross-validation on the training set.
- **DL hyperparameter search:** we employed a hold-out method to randomly split the training dataset into a validation set of 279 subjects (20%) and a training set of 1,116 (80%).

The full detailed list for our hyperparameter tuning can be found in Table 13. Furthermore, the hyper-parameter tuning results of CNNs and LSTMs can be found in Figure 8 and Figure 9 in Appendix C. We used Scikit-learn[7], Keras[8] and Tensorflow[9] to implement our models. For our feature set, we emulated previously used approaches [32, 37] for movement and cardiac sensor feature extraction in traditional ML and DL setups, which were mentioned in section 3.2. In our ML experiments, each input vector contains 400 features that combined 370 statistical features extracted from actigraphy and 30 HR/HRV features as we describe in the feature engineering section. As such, the single modality approaches for actigraphy input 370 features whereas for HR/HRV 30 features for each sleep epoch were included for each input vector. These were used as inputs for our feature-based ML benchmarks.

In our DL experiments, given that we wanted our work to be as transferable and device-agnostic as possible, we decided not to use the raw ECG signal. ECG signals are expensive and are not currently available in most of the wearable sensors. Thus, instead, we used an 8-dimensional HR/HRV feature set (see Table 5) that can be derived from many wearable cardiac sensors, such as Epatica[10], or ActiHeart[11]. For movement data, we simply use the activity counts that can be acquired directly from the wrist-worn actigraphy device. For PSG annotation in clinical settings, sleep technicians and physicians often look at *adjacent* information as well as contextual

[7]https://scikit-learn.org
[8]https://keras.io
[9]https://tensorflow.org
[10]https://www.empatica.com
[11]www.camntech.com

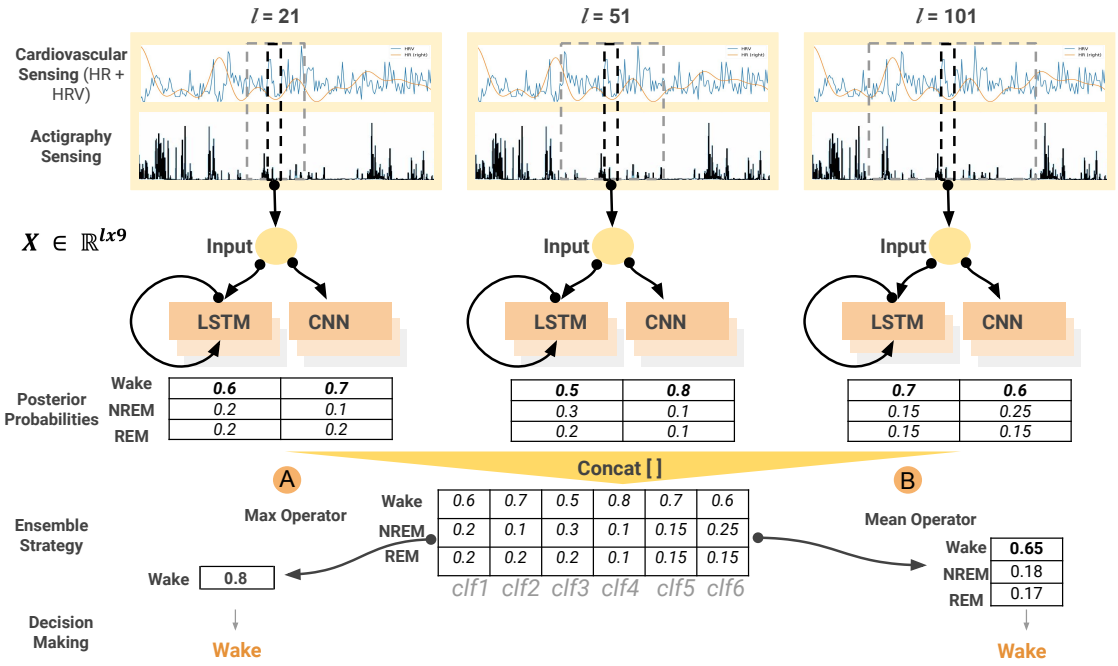


Fig. 3. Ensemble model. The model starts by taking inputs from different window lengths (l) from the multimodal sensors. A total of six different classifiers are used, combining a mixture of CNNs and LSTMs and exploiting their individual strengths. This results on posterior probability confusion matrix that is then combined through concatenation as part of the ensemble architecture. Finally, the decision making layer takes place by either (a) using a maximum operator approach or (b) a mean operator across all classifiers

temporal information to inform their decisions in scoring a particular epoch. They may look at information and trends within a 30 minute or 1 hour period as well as contextual information regarding the distribution of previous sleep stages to reach a decision for the final sleep scoring.

Motivated by this, in this work our ensemble model combines DL classifiers (a combination of CNNs and LSTMs) with various sliding window lengths (21, 51 and 101 sleep epochs, approximately equating to 10, 25, and 50 minutes, respectively). Following the previously described ensemble model pipelines, we explore two score-level fusion methods are *model-averaging* and *maximum posterior selection*.

The sleep stage ensemble classification model is based on standard single-layer CNN and LSTM networks. Figure 3 illustrates the structure for individual classifiers and their score level fusion mechanism. At the training stage, each classifier is trained independently given the hypothesis that data from sliding windows of different lengths carry different discriminative information for each sleep-stage class.

We used highly overlapping sliding windows with sleep classified at each sleep epoch (i.e., sample-wise classification). Assuming at timestamp (i.e., sleep epoch) t, the m^{th} classifier's output is a K-dimensional probability vector $\mathbf{p}_t^m \in \mathbb{R}^K$, where K is the total sleep class number (for a certain task). Probability vectors from all $M = 6$ models can then be combined using the two different fusing strategies. For *model-averaging*, the fused score can

be calculated via:

$$\mathbf{p}_t^{fusion} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{p}_t^m$$

and label $\hat{k}_t$ can be assigned to the class with the highest probability, i.e.,

$$\hat{k}_t = \underset{k}{\operatorname{argmax}}\,\mathbf{p}_t^{fusion}.$$

The second strategy *maximum posterior selection* simply assigns labels $\hat{k}_t$ to the class with the largest probability among all the M classifiers:

$$\hat{k}_t = \underset{k}{\operatorname{argmax}}\,\mathbf{P}_t, \quad where\ \mathbf{P}_t = [\mathbf{p}_t^1, \mathbf{p}_t^2, ...\mathbf{p}_t^M].$$

An example of these two fusing strategies can be found in Figure 3.

3.6 Evaluation Metrics

We adopted commonly used metrics in machine learning and medical sciences to evaluate the performance of the different classification algorithms based on task and modality combination. All of our performance metrics were derived at both a *subject level* (first derived on an individual by individual basis and then averaged across the population) and a *group level*.

To assess class imbalance and evaluate performance, we adopt several popular metrics based on True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) classifications. These evaluation metrics can be summarized as follows:

- **Accuracy** counts the number of correctly classified sleep epochs, normalized over the total number of sleep epochs (Acc = $\frac{TP+TN}{TP+TN+FP+FN}$).
- **Recall** measures the proportion of positives that are correctly identified as the given stage ($R = \frac{TP}{TP+FN}$).
- **Specificity**, also known as true negative rate, measures the proportion of negatives that are correctly identified as the given stage ($S = \frac{TN}{FP+TN}$).
- **Precision** is the fraction of correct classified instances among the overall positive predictions ($P = \frac{TP}{TP+FP}$).
- **F_1 score (F_1)** conveys the balance, with the harmonic mean, between precision and recall ($F_1 = 2 \times \frac{P \times R}{P+R}$).
- **Cohen's Kappa (κ)** measures inter-rater reliability/agreement, comparing observed accuracy with an expected accuracy ($\kappa = \frac{P_o - P_e}{1-P_e}$, where P_o the observed proportional agreement and and P_e the expected proportion of agreement). In this context, Cohen's κ factors out agreement by chance arising from the class imbalance of different sleep stages throughout the night.

With the exception of Task 1 (binary sleep-wake classification), all other tasks are multi-class classifications. We calculated the performance metrics in a class-wise manner, and reported the mean values. We also obtained the confusion matrices (corresponding to the best classifiers) for each task to further understand error types, and computed Cohen's Kappa to evaluate the agreement across the whole population. Two-tailed t-tests were used to calculated statistical significance. In this work, we also propose a measure, namely **Time Deviation**, to intuitively understand how long (in minutes), a classifier is either under or over estimating a certain sleep stage across the whole population.

Given N participants, for sleep class k the time deviation TD_k can be expressed as:

$$TD_k = \frac{1}{N}\sum_{i=1}^{N}(Pred_k^i - GT_k^i),$$

Table 6. Number of 30-seconds sleep epochs for each of the four tasks studied in this work. The numbers in parentheses were obtained within sleep period time which measured from the first to the last non-wake detected sleep epoch.

Task 1			**Task 2**			**Task 3**			**Task 4**		
Sleep Stages	# Epochs	%	Sleep Stages	# Epochs	%	Sleep Stages	# Epochs	%	Sleep Stages	# Epochs	%
Wake	652,509(314,784)	34%(20%)	Wake	652,509(314,784)	34%(20%)	Wake	652,509(314,784)	34%(20%)	Wake	652,509(314,784)	34%(20%)
Sleep	1,251,391	66%(80%)	NREM	1,022,346	54%(65%)	Light	893,472	47%(57%)	N1	171,027	9%(11%)
			REM	229,045	12%(15%)	Deep	128,874	7%(8%)	N2	722,445	38%(46%)
						REM	229,045	12%(15%)	N3	128,874	7%(8%)
									REM	229,045	12%(15%)
Total	1,903,900(1,566,175)	100%	Total	1,903,900(1,566,175)	100%	Total	1,903,900(1,566,175)	100%	Total	1,903,900(1,566,175)	100%

where $Pred_k$ is the classifier's prediction and GT_k refers to the ground truth for sleep class k. Both $Pred_k$ and GT_k were measured in minutes. This metric can help to better understand the classifier's performance/bias for a certain sleep class at the population level.

Many sleep classification studies used accuracy or F_1 to measure their model's performance, yet these are high-level metrics which do not consider class-wise performance. Confusion matrices, on the other hand, provide class-wise predictions and corresponding error types. However, for clinicians and other health practitioners, these matrices are not the most obvious way to represent the time deviation of sleep stages, as they include too many low-level details. Our proposed **Time Deviation**, is a mid-level metric, which summarizes the class-wise performance in an intuitive manner. As such, it can be used as a complementary metric to what is offered by traditional metrics (confusion matrix, accuracy, and F_1), allowing healthcare practitioners to gain an intuitive understanding towards a classifier's reliability.

4 RESULTS

For our experiments, we used a total of 1,743 nights of sleep, representing 1,903,900 sleep epochs of 30 seconds. The prevalence of sleep stages (AASM convention used) within these epochs is reported in Table 6.For consistency, all of our architectures and models were evaluated during sleep recording period across tasks, with performances reported in Table 7 for binary classification and Table 8 for multistage classification. However, we also present results derived only for the sleep period, which can be found in the Appendix B, Table 12. Within each table, results were sorted by mean accuracy in descending order. The performance of benchmark study during sleep period for all tasks can be found in Table 12 of Appendix B. For each task, a full breakdown of all classifiers is presented in the supplementary materials. In this section, we only show the top three DL classifiers alongside the best classifier from ML. Table 10 provides a summary of sleep measured by PSG and the time spent in different sleep stages for the best classifier in each task.

4.1 Task 1: Sleep-wake Classification

The best performing algorithms for Task 1 are presented in Table 7, and a full breakdown of all classifiers is presented in the supplementary tables for this task. We used baseline approaches, namely, *Always Sleep* and *Always Wake*, which showed that 66.5% of the epochs are sleep. Given the fact that for the purpose of this work, we only explored classification during the night period, we established that 66.5% was the minimum accuracy threshold for our models. Furthermore, although not reported on this work, we tested several of the well-established heuristic algorithms such as Cole-Kripke [12] and Sadeh [40] on our single-modality actigraphy data. Our results agree upon what was reported by Palotti et al. [32]. All of these approaches were outperformed by both the feature-based ML and DL models explored in our work.

All traditional ML modalities showed similar performance. Corroborating what had been shown in the related work that, when these algorithms are applied to actigraphy data, they result in high sensitivity but poor specificity [32]. Interestingly, for this task, adding HR and HRV to actigraphy for a combined sensing modality

Table 7. Sleep wake classification results (mean ± standard error at 95% confidence interval) and predicted minutes by multimodal and single modality approaches (full recording period); Actigraphy modality: 🏃, HR/HRV modality: ♥; (*Full Table available on supplementary, **Average time deviation from ground truth across all subjects ± standard error)

Sleep-Wake Classification Benchmarks*									
Method Specifics			**Performance Metrics**						**Time Deviation****
Modality	Sensors	Top 3 Classifiers	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Sleep (mins)
Multimodality	♥ 🏃 *[HR/HRV, Actigraphy]*	CNN (101)	**84.4 ± 1.0**	**67.9 ± 2.0**	**84.8 ± 1.3**	92.4 ± 1.2	87.6 ± 1.1	**62.0 ± 2.0**	36.2 ± 7.3
		LSTM (101)	**84.4 ± 1.0**	67.4 ± 1.9	84.7 ± 1.2	92.5 ± 1.1	**87.8 ± 1.0**	61.6 ± 2.1	**36.0 ± 6.7**
		CNN (51)	84.3 ± 1.0	67.3 ± 2.0	84.5 ± 1.3	**92.7 ± 1.2**	87.6 ± 1.1	61.7 ± 2.1	39.0 ± 7.2
		Random Forest (300)	82.3 ± 1.0	65.7 ± 2.1	83.7 ± 1.3	90.6 ± 1.1	57.6 ± 2.1	57.1 ± 2.1	32.9 ± 7.3
Single Modality	♥ *[HR/HRV]*	LSTM (101)	**79.5 ± 1.2**	**62.2 ± 2.1**	**81.8 ± 1.4**	88.9 ± 1.3	**84.1 ± 1.1**	**51.5 ± 2.2**	35.3 ± 4.4
		CNN (101)	79.1 ± 1.2	57.0 ± 2.1	79.9 ± 1.5	**91.0 ± 1.4**	83.9 ± 1.3	49.8 ± 2.1	54.4 ± 4.7
		LSTM (51)	78.6 ± 1.2	61.1 ± 2.0	81.2 ± 1.4	88.2 ± 1.3	83.4 ± 1.2	49.5 ± 2.1	**34.5 ± 4.6**
		Random Forest (300)	70.3 ± 1.2	39.2 ± 2.3	73.6 ± 1.4	86.7 ± 1.9	77.6 ± 1.5	27.1 ± 1.7	70.4 ± 12.5
	🏃 *[Actigraphy]*	CNN (101)	**84.9 ± 1.0**	67.1 ± 2.0	84.7 ± 1.3	**93.8 ± 1.0**	**88.3 ± 1.0**	**63.0 ± 2.0**	43.0 ± 6.9
		CNN (51)	84.4 ± 1.0	67.6 ± 2.0	84.6 ± 1.3	92.9 ± 1.1	87.8 ± 1.1	62.2 ± 2.1	39.0 ± 7.1
		LSTM (101)	84.3 ± 1.0	**69.7 ± 1.8**	**85.5 ± 1.2**	91.2 ± 1.1	87.6 ± 1.0	62.0 ± 2.0	**26.5 ± 6.6**
		Random Forest (300)	81.2 ± 1.0	63.4 ± 2.0	82.9 ± 1.3	89.7 ± 1.1	85.4 ± 1.0	54.1 ± 2.1	32.6 ± 7.2

on the top classifier of CNN (101) did not significantly improve F_1 ($p = 0.347$), accuracy ($p = 0.499$) or Cohen's κ ($p = 0.506$). As expected, HR/HRV alone did not yield comparable performance to actigraphy alone or the combined sensing approach.

4.2 Task 2: Wake, Non-REM sleep, REM Sleep Classification

Task 2 evaluated sleep stages from a low granularity perspective by aggregating the different partitions of NREM. As observed in the supplementary table for this task, although some ML models had reasonable performance, the DL approaches were superior. It is important to note that at this level of granularity, NREM is overestimated while REM is underestimated for almost all models except for CNN (51) and CNN (101). In contrast to what was observed in Task 1, all models explored have a higher specificity than sensitivity and accuracy higher than F_1 score due to the imbalanced dataset.

As reflected in the top part of Table 8, the best classifiers for this task were all DL models for all sensor modalities. These models were significantly better than the best traditional ML model (Random Forest), with $p < 0.001$ for all metrics evaluated. The best DL algorithm with respect to accuracy was LSTM (51) which was also statistically better than CNN (21) ($p < 0.001$) achieving an accuracy of 76.2%. However, it was not significantly better than CNN (101) in terms of F_1, sensitivity and specificity ($p = 0.364, p = 0.138, p = 0.063$ and $p = 0.399$). Nevertheless, CNN (101) achieved the lowest mean time deviation with a 2.5-minute overestimation of REM sleep. Interestingly, for this task, most algorithms' specificity significantly improved (e.g. LSTM (51) $p < 0.001$, reaching a specificity of 86%) when compared to Task 1 with the exception of the perceptron model.

In this task, it becomes apparent that multimodality is required for better performance at multistage classification, with the single modality approaches being significantly ($p < 0.001$) outperformed in all performance metrics and yielding much larger time deviations.

4.3 Task 3: Wake, Light Sleep, Deep Sleep and REM-sleep Classification

Task 3 explored sleep staging at a higher level of granularity than Task 2, with class imbalances being perhaps more apparent, as shown in Table 6. Here, DL approaches continued to outperform all feature-based ML models except for the Random Forest, which was not significantly worse than the CNN (21). The full results are available in supplementary tables. The best performing model was LSTM (51), although it was closely followed by LSTM (101) and CNN (101). Multimodal approaches were significantly better ($p < 0.001$) across all metrics upon

Table 8. Sleep stage classification results (mean ± standard error at 95% confidence interval and predicted minutes by multimodal and single modality approaches (full recording period); Actigraphy modality: 🏃, HR/HRV modality: ♥; Three different tasks: Task 2: 3 stages, Task 3: 4 stages, Task 4: 5 Stages (*Full Table available on supplementary, **Average time deviation from ground truth across all subjects ± standard error)

Task 2: Wake, NREM, REM

Method Specifics			Performance Metrics						Time Deviation*		
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	NREM
Multimod.	♥ 🏃	LSTM (51)	**76.2 ± 1.0**	**85.6 ± 0.5**	**72.2 ± 1.3**	68.8 ± 1.2	67.9 ± 1.3	58.4 ± 1.8	-13.2 ± 6.8	-10.7 ± 3.8	**23.9 ± 7.1**
		LSTM (101)	76.1 ± 0.9	85.1 ± 0.5	71.9 ± 1.4	66.8 ± 1.2	66.4 ± 1.3	57.4 ± 1.9	**-3.2 ± 6.8**	-23.3 ± 3.4	26.5 ± 7.0
		CNN (101)	76.0 ± 1.0	**85.6 ± 0.6**	**72.2 ± 1.2**	**69.7 ± 1.3**	**68.1 ± 1.3**	**58.6 ± 1.9**	-32.7 ± 7.2	**2.5 ± 4.5**	30.2 ± 7.7
		Random Forest (300)	70.5 ± 0.9	79.9 ± 0.5	59.2 ± 1.5	53.0 ± 0.7	50.3 ± 0.7	47.6 ± 1.7	-20.0 ± 7.6	-63.6 ± 2.9	83.5± 7.8
Single Modality	♥	LSTM (101)	**73.8 ± 1.2**	**84.3 ± 0.6**	**69.8 ± 1.5**	**66.1 ± 1.3**	**64.9 ± 1.5**	**50.0 ± 2.2**	-27.8 ± 8.5	-8.6 ± 4.2	36.4 ± 8.1
		LSTM (51)	72.9 ± 1.1	83.8 ± 0.6	67.9 ± 1.4	64.1 ± 1.3	62.9 ± 1.4	45.5 ± 2.1	-17.9 ± 8.5	-16.2 ± 4.3	34.1 ± 8.3
		CNN (101)	71.0 ± 1.2	83.6 ± 0.6	66.3 ± 1.4	65.4 ± 1.4	62.7 ± 1.4	46.1 ± 2.0	**-12.9 ± 9.4**	**2.3 ± 5.0**	**10.6 ± 9.1**
		Random Forest (300)	59.6 ± 1.0	73.8 ± 0.4	48.4 ± 1.4	43.4 ± 0.6	39.2 ± 0.8	19.7 ± 1.4	-26.8 ± 13.5	-65.3 ± 3.1	92.0 ± 13.2
	🏃	LSTM (101)	**71.4 ± 0.9**	**80.1 ± 0.6**	**51.7 ± 1.1**	**52.9 ± 0.7**	**49.8 ± 0.8**	**49.7 ± 1.7**	**-16.8 ± 7.3**	**-67.0 ± 3.0**	**83.8 ± 7.7**
		CNN (101)	71.0 ± 1.0	79.5 ± 0.6	50.1 ± 0.8	52.1 ± 0.8	49.1 ± 0.8	48.0 ± 1.8	-34.9 ± 7.5	-67.6 ± 3.0	102.5 ± 7.9
		LSTM (51)	70.9 ± 0.9	79.7 ± 0.6	49.0 ± 0.8	52.4 ± 0.7	49.2 ± 0.8	48.3 ± 1.7	-20.3 ± 7.4	-67.6 ± 3.0	87.9 ± 7.8
		Random Forest (300)	68.6 ± 0.9	78.8 ± 0.5	53.4 ± 1.1	51.0 ± 0.7	48.5 ± 0.8	44.3 ± 1.7	-20.5 ± 7.2	-60.7 ± 2.9	81.2 ± 7.4

Task 3: Wake, Light Sleep, Deep Sleep, REM

Method Specifics			Performance Metrics						Time Deviation*			
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	Deep Sleep	Light Sleep
Multimod.	♥ 🏃	LSTM (51)	**70.3 ± 1.0**	**87.4 ± 0.4**	57.9 ± 1.3	**54.0 ± 1.0**	**51.9 ± 1.0**	**53.8 ± 1.9**	**-1.0 ± 6.9**	-5.6 ± 4.0	-36.2 ± 3.5	**42.8 ± 7.4**
		LSTM (101)	70.2 ± 1.0	86.9 ± 0.4	**59.9 ± 1.5**	52.4 ± 1.0	51.3 ± 1.1	51.7 ± 1.8	-18.9 ± 6.6	-24.7 ± 3.7	**-32.4 ± 3.5**	76.0 ± 7.3
		CNN (101)	69.0 ± 1.0	87.0 ± 0.4	58.0 ± 1.4	53.7 ± 1.0	51.2 ± 1.1	51.6 ± 1.8	-15.9 ± 7.5	**4.4 ± 4.8**	-34.5 ± 3.5	46.1 ± 8.1
		Random Forest (300)	63.6 ± 1.0	83.3 ± 0.4	44.7 ± 1.3	40.1 ± 0.6	36.7 ± 0.6	34.4 ± 1.3	-15.2 ± 7.6	-61.3 ± 2.9	-38.9 ± 3.6	115.3 ± 8.3
Single Modality	♥	LSTM (101)	**67.4 ± 1.2**	**86.2 ± 0.4**	**56.2 ± 1.6**	**51.3 ± 1.1**	**49.5 ± 1.2**	**44.6 ± 2.2**	**-13.1 ± 8.4**	-13.5 ± 3.8	**-33.7 ± 3.5**	60.4 ± 8.1
		LSTM (51)	66.2 ± 1.1	85.6 ± 0.4	54.4 ± 1.5	49.5 ± 1.1	47.4 ± 1.1	41.2 ± 2.1	-15.0 ± 8.1	-14.6 ± 4.1	-36.4 ± 3.5	65.9 ± 7.9
		CNN (101)	64.3 ± 1.1	85.3 ± 0.4	54.4 ± 1.6	50.2 ± 1.1	47.1 ± 1.1	40.9 ± 2.1	-23.0 ± 9.5	**8.2 ± 5.1**	-34.8 ± 3.5	**49.6 ± 8.9**
		Random Forest (300)	53.3 ± 1.0	79.2 ± 0.4	35.5 ± 1.1	33.2 ± 0.5	28.6 ± 0.6	12.6 ± 1.1	-4.3 ± 14.0	-64.7 ±3.0	-39.0 ± 3.6	-39.0 ± 3.6
	🏃	LSTM (101)	**64.1 ± 1.0**	82.9 ± 0.5	35.6 ± 0.7	**39.6 ± 0.7**	**35.8 ± 0.7**	**33.5 ± 1.4**	-32.5 ± 7.4	-67.6 ± 3.0	**-39.3 ± 3.6**	139.4 ± 8.5
		CNN (101)	63.9 ± 1.0	**83.0 ± 0.4**	**36.3 ± 0.9**	**39.6 ± 0.7**	35.7 ± 0.7	**33.5 ± 1.4**	**-26.4 ± 7.6**	-67.5 ± 3.0	**-39.3 ± 3.6**	**133.2 ± 8.7**
		LSTM (51)	63.6 ± 1.0	82.7 ± 0.4	35.6 ± 0.8	39.3 ± 0.7	35.5 ± 0.7	33.0 ± 1.4	-36.3 ± 7.1	**-67.3 ± 3.0**	**-39.3 ± 3.6**	143.0 ± 8.2
		Random Forest (300)	61.4 ± 1.0	82.6 ± 0.4	39.6 ± 0.9	38.1 ± 0.5	35.0 ± 0.6	31.2 ± 1.3	-15.6 ± 7.3	-59.3 ± 2.9	-37.1 ± 3.6	112.1 ± 8.1

Task 4: Wake, REM, N1,N2,N3

Method Specifics			Performance Metrics						Time Deviation **				
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	N3 Sleep	N2 Sleep	N1 Sleep
Multimod.	♥ 🏃	LSTM (51)	**63.7 ± 1.0**	88.7 ± 0.3	47.1 ± 1.4	43.0 ± 0.8	39.9 ± 0.8	56.3 ± 1.8	22.2 ± 7.1	-12.9 ± 3.9	-35.2 ± 3.5	**71.9 ± 7.5**	-46.0 ± 3.0
		LSTM (101)	63.6 ± 1.0	88.7 ± 0.3	47.8 ± 1.3	43.3 ± 0.8	40.5 ± 0.9	**57.0 ± 1.8**	**-3.3 ± 6.8**	-15.9 ± 3.9	**-32.3 ± 3.5**	97.7 ± 7.5	-46.2 ± 3.0
		CNN (101)	63.1 ± 1.1	**88.8 ± 0.3**	**51.5 ± 1.4**	**44.7 ± 0.9**	**41.9 ± 0.9**	56.2 ± 1.8	-26.2 ± 7.1	**8.2 ± 5.0**	-34.3 ± 3.5	92.4 ± 8.0	**-40.2 ± 3.1**
		Random Forest (300)	56.9 ± 1.0	86.2 ± 0.3	36.4 ± 1.2	33.1 ± 0.5	28.8 ± 0.5	46.3 ± 1.6	18.6 ± 8.1	-54.9 ± 3.1	-38.7 ± 3.6	123.6 ± 8.4	-48.6 ± 3.2
Single Modality	♥	CNN (21)	**55.6 ± 1.1**	86.4 ± 0.3	40.4 ± 1.2	37.3 ± 0.8	33.6 ± 0.9	36.2 ± 1.8	-15.6 ± 10.1	**-3.9 ± 5.8**	-39.1 ± 3.6	103.0 ± 9.8	-44.4 ± 3.0
		CNN (101)	**55.6 ± 1.1**	**86.7 ± 0.3**	**44.9 ± 1.4**	**38.9 ± 0.9**	**35.9 ± 1.0**	**37.1 ± 1.8**	**1.1 ± 10.7**	-12.0 ± 4.8	**-29.5 ± 3.6**	**81.2 ± 9.4**	**-40.8 ± 3.1**
		CNN (51)	54.2 ± 1.1	86.0 ± 0.3	41.2 ± 1.3	35.6 ± 0.8	32.1 ± 1.0	32.3 ± 1.9	35.7 ± 12.1	-28.2 ± 5.1	-36.4 ± 3.5	69.5 ± 11.0	-40.6 ± 3.1
		Random Forest (300)	46.6 ± 1.0	83.1 ± 0.3	29.9 ± 1.0	27.1 ± 0.4	22.3 ± 0.5	17.6 ± 1.4	48.5 ± 14.3	-61.3 ±3.1	-38.9 ± 3.6	97.8 ± 13.7	-46.2 ± 3.2
	🏃	LSTM (51)	**56.9 ± 1.0**	**85.7 ± 0.4**	26.1 ± 0.8	32.2 ± 0.6	27.1 ± 0.7	46.9 ± 1.7	-12.9 ± 7.5	-67.6 ± 3.0	**-39.3 ± 3.6**	169.2 ± 8.5	**-49.4 ± 3.2**
		LSTM (101)	**56.9 ± 1.0**	**85.7 ± 0.4**	25.3 ± 0.7	**32.3 ± 0.6**	27.1 ± 0.7	**47.1 ± 1.7**	**-3.3 ± 7.5**	-67.6 ± 3.0	**-39.3 ± 3.6**	159.7 ± 8.7	**-49.4 ± 3.2**
		CNN (101)	56.8 ± 1.1	85.8 ± 0.3	**27.7 ± 0.9**	32.2 ± 0.5	**27.2 ± 0.6**	46.9 ± 1.7	9.6 ± 8.3	**-65.6 ± 3.0**	**-39.3 ± 3.6**	**144.7 ± 9.1**	**-49.4 ± 3.2**
		Random Forest (300)	54.4 ± 1.0	85.6 ± 0.3	31.7 ± 0.8	31.1 ± 0.4	27.2 ± 1.6	42.7 ± 1.6	16.2 ± 7.6	-56.0 ±2.8	-36.7 ± 3.5	120.8 ± 8.0	-44.3 ± 3.3

comparison of the best classifiers for each category explored, depicting the value of these combined sensing approaches for multistage classification. Across all sensing modalities and all algorithms, deep and REM sleep were underestimated, with the exception of CNN (101) in the multimodal setup. In contrast, light sleep was overestimated, with wake being slightly underestimated across all setups, due to the class imbalance, except for LSTM (51).

4.4 Task 4: Wake, N1, N2, N3, REM Sleep Slassification

Finally, Task 4 aimed to classify sleep stages following AASM rules (N1, N2, N3, REM and Wake). This task is the most complex due to its level of granularity and high-class imbalance, and as expected, the models performed

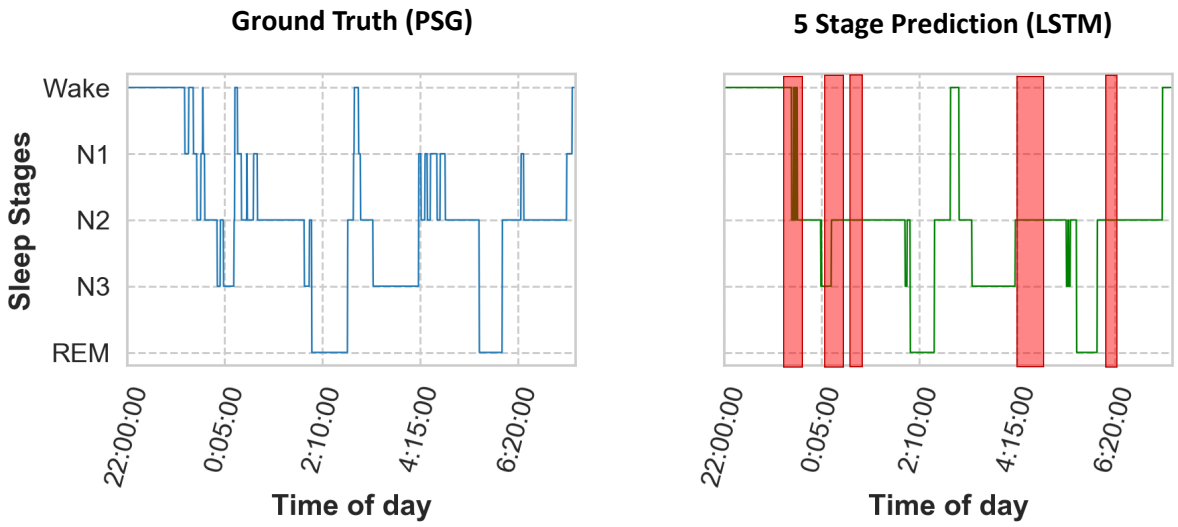


Fig. 4. Classification performance for multimodal, 5 stage classification using LSTM. On the top, the ground truth PSG, at the bottom, the predicted stages by the model. Highlighted in red are areas where the model does poorly.

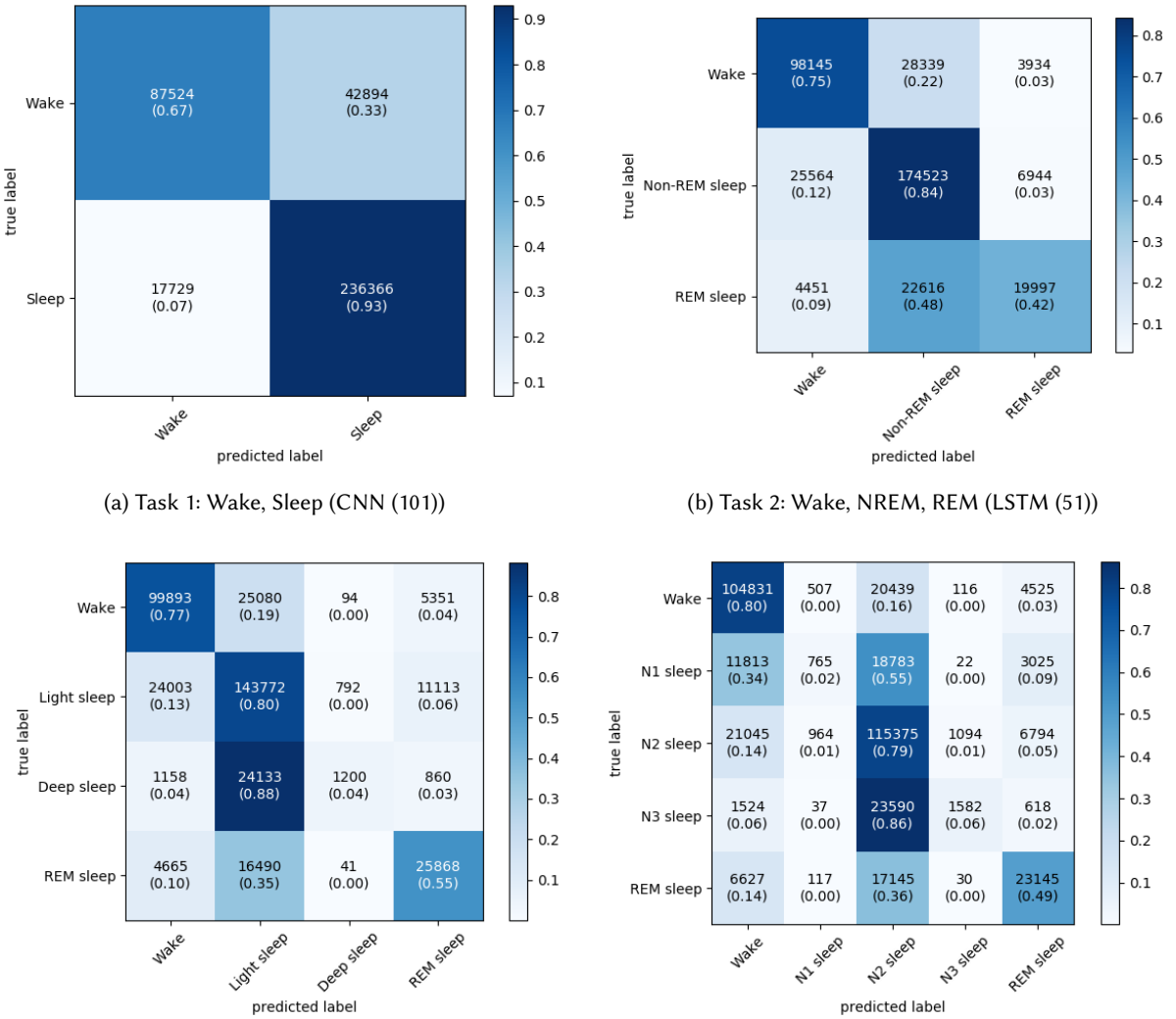
worse here than in the previous tasks. An example of the best performing model LSTM (101) and the *mistakes* it makes is highlighted in Figure 4.

Like in previous tasks, the performance of DL algorithms was significantly better than feature-based ML algorithms as depicted in Table 8. The three best performing DL algorithms were not significantly different from each other with respect to accuracy($p > 0.05$) and F1 scores ($p > 0.05$). The best performing algorithm was LSTM (51), with an accuracy of 63.7% and an F_1 score of 39.9%. Even in the best performing multimodal approach, N2 tended to be severely overestimated (71 minutes more on average across the population). Nevertheless, the multimodal and HR/HRV approaches were good at classifying wake and REM, with only moderate deviations in time for those classes.

It is important to note that although the performance in terms of accuracy for the single modality approaches was comparable, each method struggled or had strengths at very different things. For instance, HR/HRV was significantly better at classifying REM sleep in this modality than actigraphy. Similarly, upon evaluation of the algorithms only during the sleep period only (Table 12, multimodal approaches were significantly better at detecting awakenings, yielding a more accurate wake after sleep onset (WASO) metric.

Figure 5 shows the confusion matrix for the best classifiers per task, allowing us to observe how models have an *easier* time classifying REM and wake and struggle to classify N1 and N3 (NREM). The observed time deviation in minutes substantiates this finding.

Finally, we evaluated the performance of different ensemble methods for each task. To validate the performance of our ensemble model, we conducted t-test based on both subject level as well as the group of subjects level. The difference between the two experiments lies in that for the second approach, we randomly divide all test



(a) Task 1: Wake, Sleep (CNN (101))

(b) Task 2: Wake, NREM, REM (LSTM (51))

(c) Task 3: Wake, Light Sleep, Deep Sleep, REM (LSTM (51))

(d) Task 4: Wake, N1, N2, N3, REM (LSTM (51))

Fig. 5. Confusion matrix for the best classifier per Task

subjects into 29 groups, each group containing 12 individuals. The purpose is to test whether the benefits of using ensemble methods are due to random chance.

The results of the two ensemble architecture models explored (based on different score-level fusion approaches) are shown in Table 9. We found no significant differences between the two ensemble models for all performance metrics assessed. However, they achieve better accuracy than single classifier approaches for all tasks and are significantly better on several performance metrics.

In Task 2, the ensemble approaches significantly outperformed LSTM (51) in terms of accuracy ($p < 0.05$), F_1 score ($p < 0.05$) and Cohen's κ ($p < 0.05$) and CNN (101) in terms of accuracy ($p < 0.05$) and Cohen's κ ($p < 0.05$)

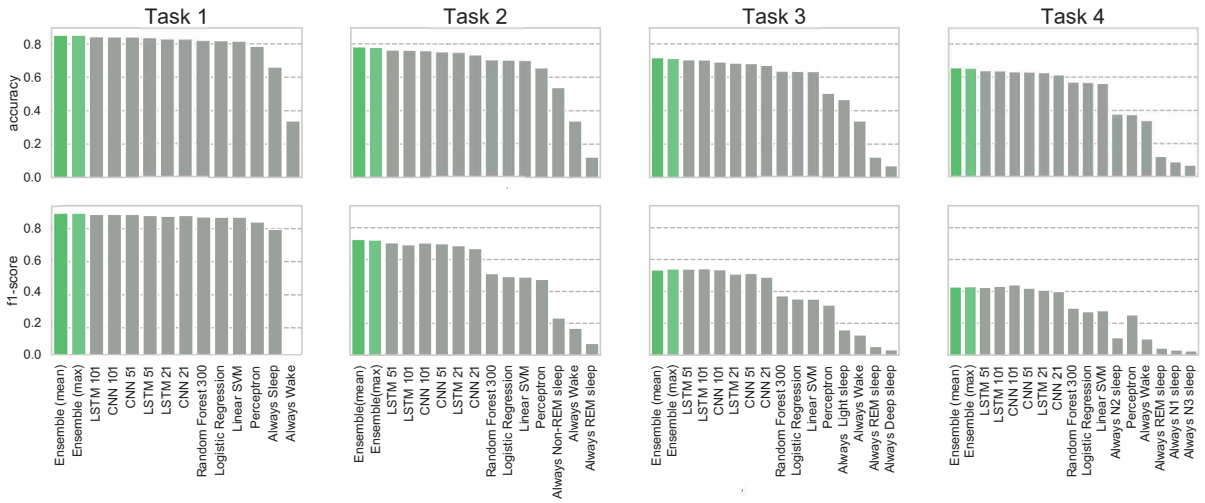


Fig. 6. Performance (accuracy, F_1) per Task and model. Task 5 (ensemble architectures) are depicted against all benchmarks per each task on green

Table 9. Results (mean ± standard error at 95% confidence interval) of different ensemble methods for each task.(Mean over classifiers and Maximum selection are ensemble models)

Ensemble method	Accuracy	Cohen's κ	F_1	Precision	Recall	Specificity	Time Deviation (mins)				
Task 1 (2 Stages)							**Sleep**				
Maximum selection	85.3 ± 1.0	**64.4 ± 2.0**	88.4 ± 1.1	**85.7 ± 1.2**	92.8 ± 1.1	**70.1 ± 1.9**	**33.4 ± 6.7**				
Mean over classifiers	**85.4 ± 1.0**	64.3 ± 2.1	**88.5 ± 1.0**	85.4 ± 1.3	**93.4 ± 1.1**	69.1 ± 2.0	37.5 ± 6.8				
Task 2 (3 stages)							**Wake**	**REM sleep**	**NREM sleep**		
Maximum selection	77.9 ± 1.0	61.4 ± 1.8	69.6 ± 1.3	74.5 ± 1.3	70.6 ± 1.2	**86.5 ± 0.5**	**-16.1 ± 6.8**	**-11.1 ± 4.2**	**27.3 ± 7.5**		
Mean over classifiers	**78.2 ± 0.9**	**61.9 ± 1.8**	**69.8 ± 1.3**	**75.2 ± 1.3**	**70.7 ± 1.3**	**86.5 ± 0.5**	-21.7 ± 6.8	-13 ± 4.2	34.7 ± 7.5		
Task 3 (4 stages)							**Wake**	**REM sleep**	**Deep sleep**	**Light sleep**	
Maximum selection	71.1 ± 1.0	55.7 ± 1.8	**52.4 ± 1.0**	**58.3 ± 1.3**	**54.8 ± 1.0**	**87.7 ± 0.4**	**-11.1 ± 6.7**	**-3.5 ± 4.6**	**-37.4 ± 3.5**	**52.0 ± 7.8**	
Mean over classifiers	**71.6 ± 1.0**	**56.1 ± 1.8**	52.1 ± 1.0	57.1 ± 1.2	54.3 ± 1.0	87.6 ± 0.4	-17.8 ± 6.7	-8.9 ± 4.4	-38.5 ± 3.5	65.2 ± 7.8	
Task 4 (5 stages)							**Wake**	**REM sleep**	**N3 sleep**	**N2 sleep**	**N1 sleep**
Maximum selection	65.2 ± 1.0	59.6 ± 1.8	**41.4 ± 0.8**	**49.7 ± 1.4**	**45.2 ± 0.8**	**89.3 ± 0.3**	**3.0 ± 6.9**	4.7 ± 5.1	**-37.1 ± 3.5**	**76.4 ± 7.8**	-47 ± 3.1
Mean over classifiers	**65.4 ± 1.0**	**60.1 ± 1.8**	41.2 ± 0.8	48.6 ± 1.4	44.9 ± 0.8	89.2 ± 0.3	-5.1 ± 6.9	**-2.1 ± 4.8**	-38.4 ± 3.5	91.9 ± 7.8	**-46.3 ± 3.1**

based on both subject and group level t test. These models were the best two performers for that task prior to the introduction of the ensemble approach. Interestingly, on Task 4 (highest level of class granularity) the ensemble models only outperformed the best classifier (LSTM (51)) in terms of Cohen's κ and accuracy (for both subject and group level t test).

A summary of results per class and model are presented in Figure 6.

4.5 Feature Importance Analysis

To understand how different modalities contribute to the prediction of each sleep stage, models that rank feature importance can be implemented. Many traditional ML approaches can provide feature importance ranking, such as logistic regression, linear Support Vector Machine or Random Forests. Of those, Random Forest is one of the most powerful traditional ML models, and it can rank feature importance by calculating the mean Gini impurity

Table 10. Sleep parameters and predicted minutes of each sleep stage in the *test* dataset. Numbers are minutes except for the sleep efficiencies which are reported as percentages. Results are in mean +- SD/ and numbers in parentheses indicate the range in 95% CI (Mean over classifiers and Maximum selection are ensemble models)

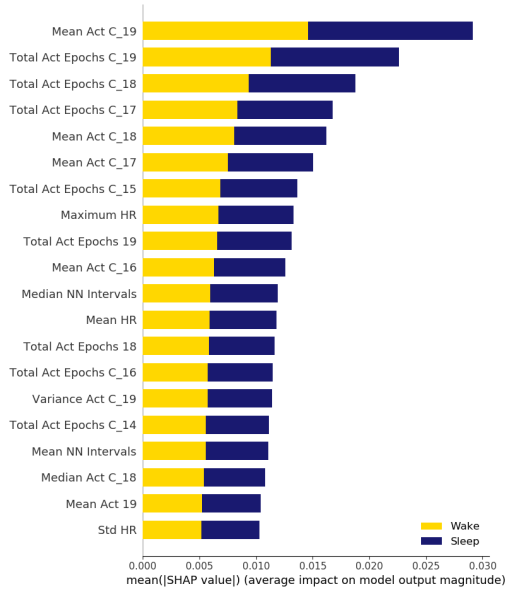
Minutes of Sleep Stages						
Task	**Methods**	**Wake**	**Sleep**			
1	Ground truth	187.8±81.6 (179.2-196.4)	365.7±81.8 (357.1-374.3)			
	Mean over classifiers	150.2±73.2 (142.5-157.9)	403.3±92.0 (393.6-413.0)			
		Wake	**REM**	**NREM**		
2	Ground truth	187.8±81.6 (179.2-196.4)	68.6±27.2 (65.7-71.5)	299.7±66.5 (292.7-306.7)		
	Mean over classifiers	166.1±77.2 (158.0-174.2)	57.4±39.0 (53.3-61.5)	332.7±87.8 (323.5-341.9)		
		Wake	**REM**	**Deep Sleep**	**Light Sleep**	
3	Ground truth	187.8±81.6 (179.2-196.4)	68.6±27.2 (65.7-71.5)	41.8±33.7 (38.3-45.3)	258.8±65.7 (251.9-265.7)	
	Maximum selection	176.7±78.3 (168.5-184.9)	65.4±43.1 (60.9-69.9)	5.2±6.2 (4.5-5.9)	310.6±85.4 (301.6-319.6)	
		Wake	**REM**	**N1**	**N2**	**N3**
4	Ground truth	187.8±81.6 (179.2-196.4)	68.6±27.2 (65.7-71.5)	50.1±30.5 (46.9-53.3)	209.1±58.7 (202.9-215.3)	41.8±33.7 (38.3-45.3)
	CNN (101)	161.6±79.1 (153.3-169.9)	76.7±46.8 (71.8-81.6)	9.6±14.6 (8.1-11.1)	301.5±87.0 (292.4-310.6)	7.7±8.7 (6.8-8.6)

Sleep Parameters						
Task	**Methods**	**Total Sleep Time**	**Wake After Sleep Onset**	**Sleep Period Duration**	**Sleep Efficiency (Recording Period)**	**Sleep Efficiency (Sleep Period)**
	Ground truth	365.7±81.8 (357.1-374.3)	89.4±62.5 (82.8-96.0)	455.1±90.0 (445.6-464.6)	66.5±12.9 (65.1-67.9)	80.9±11.8 (79.7-82.1)
1	Mean over classifiers	360.7±84.0 (351.9-369.5)	70.9±57.4 (64.9-76.9)	474.2±92.9 (464.4-484.0)	65.5±13.2 (64.1-66.9)	76.9±14.0 (75.4-78.4)
2	Mean over classifiers	359.5±84.3 (350.6-368.4)	81.7±60.1 (75.4-88.0)	469.1±93.0 (459.3-478.9)	65.3±13.2 (63.9-66.7)	77.4±13.9 (75.9-78.9)
3	Maximum selection	358.2±84.1 (349.4-367.0)	86.5±61.4 (80.0-93.0)	463.3±92.3 (453.6-473.0)	65.0±13.2 (63.6-66.4)	78.1±13.7 (76.7-79.5)
4	CNN (101)	361.1±83.3 (352.3-369.9)	94.6±68.6 (87.4-101.8)	486.5±90.7 (477.0-496.0)	65.6±13.1 (64.2-67.0)	75.0±14.2 (73.5-76.5)

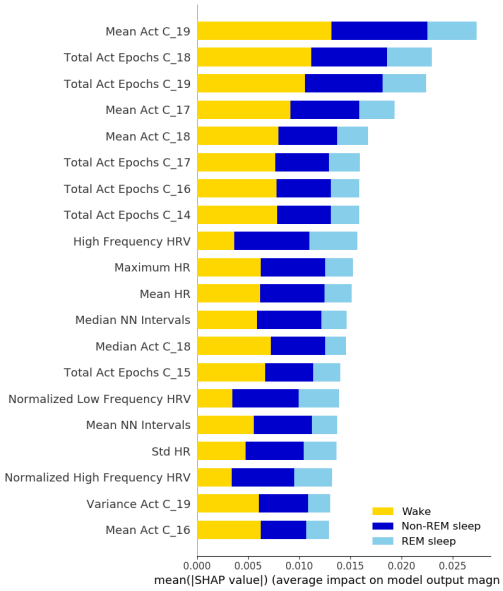
or mean information gain over all its decision trees. However, these approaches only yield features that are important to the holistic classification task, and do not provide information on how these features contribute to recognizing certain classes (e.g., a sleep stage like REM sleep). We used SHAP [28] with Random Forest, which can generate class-wise feature importance. More technical details of SHAP can be found in [28].

By using this SHAP implementation with Random Forest, we can calculate the most important features per class, as shown in Figure 7. We report the top 20 features on Tasks 1-4, respectively. It is interesting to see how the top ranked features differ from task to task, pointing towards what contributes to more granular levels of classification. For instance, in Task 1 (i.e., binary sleep/wake classification), the most informative features are from movement sensors (15 features out of 20), in contrast to those obtained from cardiac sensing (5 out of 20). However, cardiac features become more and more important as multi-stage classification tasks get more granular. In Tasks 2-4, with the increased class granularity, the most important features are cardiac features (8, 10, and 13 cardiac features, respectively) indicating the key role of cardiac sensing in distinguishing detailed sleep patterns.

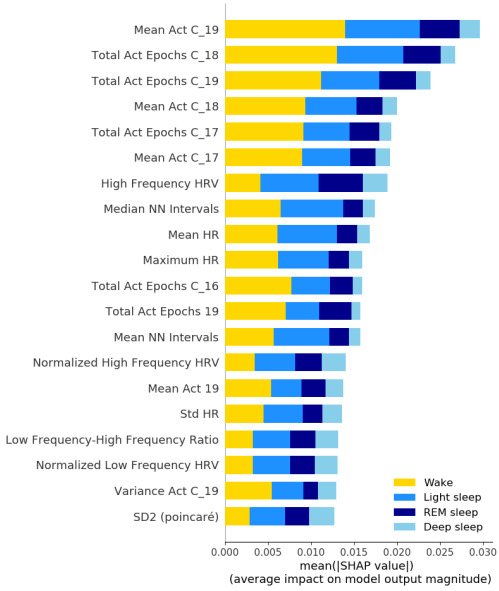
In multi-stage sleep classification (i.e., Tasks 2-4), it is also interesting to see feature importance associated to the different classes. Specifically, we observe high frequency HRV is the most discriminant feature in recognizing REM sleep, a finding that is consistent across all 3 multi-stage classification tasks. However, high frequency HRV is not as valuable in recognizing wake status. In Task 3 and 4, the non-linear HRV features such as SD2 which is the normalized Poincare plot parameter, SDNN, and coefficient of variation of NNI become more important than time domain features of HRV. Among these features, SD2 is ranked higher than many of the other HRV features except high frequency HRV in Task 4.



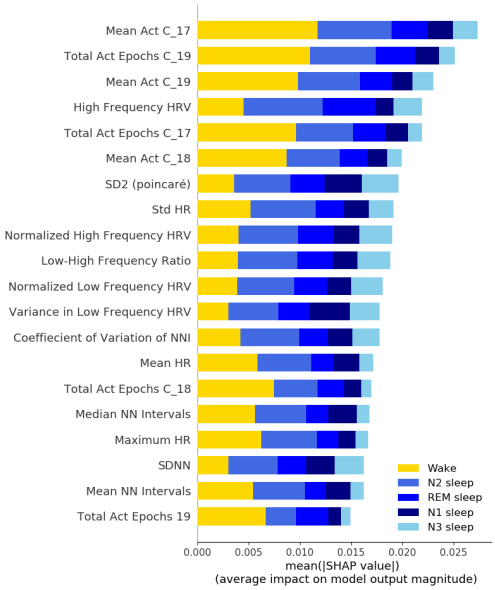
(a) SHAP for Task 1 : Wake, Sleep



(b) SHAP for Task 2 : Wake, NREM, REM



(c) SHAP for Task 3 : Wake, Light Sleep, Deep Sleep, REM



(d) SHAP for Task 4: Wake, N1, N2, N3, REM

Fig. 7. SHAP values (Random Forest) for class-wise feature importance ranking in Task 1-4

5 DISCUSSION

5.1 Summary

This work presents the first systematic analysis of sleep-wake and sleep-stage classification using multimodal sensor data in a large, diverse population of both healthy and sleep-disordered participants. The main aim of this work was to understand how different models performed based upon details of the task (sleep-wake or multistage sleep classification) and sensor combination. To achieve this, we applied a series of traditional ML and DL approaches to each individual modality (i.e., actigraphy, ECG) and sensor combination (multimodal sensor fusion). Furthermore, we ran four different tasks to gain a deeper understanding of the strengths and limitations of the different approaches.

These tasks include: sleep-wake (Task 1); wake, NREM and REM (Task 2); wake, light sleep, deep sleep and REM (Task 3); and wake, N1, N2, N3 and REM (Task 4). The framework and analysis we provide were based on sensor modalities and signals that can be obtained from research grade wearable devices. Hence, RR-based metrics were used instead of raw ECG signals. Unlike raw ECG, these metrics may be derived from commercial research-grade wearable devices and in the near future also from non-clinical smartwatches that use both actigraphy/accelerometers and photoplethysmogram (PPG) [31, 50]. With this work, we aim to provide with a set of benchmarks for commercial and research studies and to inspire others to create open-access large population repositories to study the role of sleep and other physical behaviors in health and disease.

Here, we systematically evaluated how sensor modality affects classification outcomes and how model choice leads to differences in performance. Yuda et al. also explored a multimodal approach for sleep classification. Although their work is strong methodologically, the cohort is much smaller than that presented here, almost 70% of their cohort is male and the majority of their subjects had sleep disorders, limiting the generalisability of their findings [56]. Furthermore, they only explored the classification of three sleep stages. Here, we show that although multimodal sensor approaches do not lead to great improvements in classification performance for sleep-wake classification tasks, they are essential to classify sleep stages. For instance, actigraphy by itself struggles to classify REM sleep in Tasks 2-4 (Table 8), but it's performance improves when combined with HR/HRV. To-date, conventional sleep-wake classification algorithms have mostly exploited count-based movement data [40, 41] and models that combined HR information have mostly been confined to commercial devices based on device-specific algorithms.

Furthermore, this work highlights the strengths and limitations of the different models used, for instance, while CNN models do well at classifying high frequency transitions, LSTMs excel at classifying smooth patterns. LSTMs outperformed all other classifiers at multistage classification, due to their deep temporal modelling characteristics which align well with the multi-class, time-series classification problem that sleep-stages introduce. Meanwhile, CNNs were the best performers for binary sleep-wake classification tasks due to their ability to track exponentially longer sequences, such as those used in this type of task where the objective is less granular and has lower transition frequencies than the multi-class scenario. As such, our ensemble approach aimed to exploit individual model *strengths* to achieve better performance.

In sum, this work presents the first systematic analysis of single modality (actigraphy, HR/HRV) and multimodal sensing approaches for sleep-wake and sleep-stage classification using the most common feature-based ML and DL frameworks. Furthermore, a new ensemble architecture is introduced, outperforming all other models.

5.2 Transparency in Algorithm Development in Machine Learning for Sleep Health

All our analyses were performed in MESA [14, 57], a publicly available dataset for which access can be requested through: https://sleepdata.org/datasets/mesa. Our code is also available on GitHub: https://github.com/bzhai/multimodal_sleep_stage_benchmark.git. This creates transparency and facilitates reproducibility in human sleep science. We encourage others to use this resource to develop novel, more accurate models that exploit multimodal

data. Here we provide an example of how the performance of well-established methods can be surpassed by an *ensemble architecture* of DL models of different window sizes. Similarly, we found that the performance of our models was influenced by optimal hyperparameter search, as reflected in the Appendix C. In particular, we observed that although there was no significant improvement in terms of accuracy and F1 on our search space, certain patterns did emerge. For most CNNs and LSTMs, increasing the length of the sliding window improved performance (except for Task 1, sleep-wake only). Different search spaces may be able to yield more significant results and shall be explored in future work.

In this work, we advocate for and demonstrate the value of including performance metrics beyond the conventional accuracy, specificity, precision, recall and F_1 scores. For instance, introducing time deviation metrics allowed us to understand what precisely each model over- or underestimated. This is of particular value for the translational applications of this work which may be implemented by HCI researchers, clinicians or epidemiologists and have an impact on the field of digital health. These metrics are more interpretable for non-machine learning experts who may seek to understand how certain inferences should be interpreted. Clear, interpretable measures that allow non-specialists to understand the limitations of our models is critical both to the development of better study cohorts and to understanding the inferences made by these models.

5.3 Sleep Classification Performance by Task

Binary sleep-wake classification (Task 1) using actigraphy had been previously explored by Palotti et al. in the same cohort [32]. Our results corroborate this study, with the CNN architecture narrowly outperforming the LSTM architecture. Interestingly, our multimodal approach did not add much to this binary classification task when exploring conventional metrics, but did yield a lower time deviation of overall sleep time than actigraphy alone. Models based only on cardiac signals had a slightly worse performance than both actigraphy alone and the multimodal approach, with accuracy estimates in the high 70s (79% for the LSTM (101)).

Task 2 consists of Wake, NREM and REM classification. This represents a valuable yet holistic overview of sleep stages and is what most free-living commercial devices aim to measure. Actigraphy and HR/HRV yielded an accuracy of around 74% and F_1 scores between 49% and 65%. Our time deviation metrics allowed us to observe that actigraphy cannot accurately determine REM sleep, whilst HR/HRV perform better. Both sensor modalities tend to overestimate time spent in NREM sleep. This finding also pertained to the multimodal approach, where NREM overestimation was the most error-prone estimate of the three states classified, at around 24 minutes mean deviation from gold-standard measures per participant on average when using the LSTM (101) and 24 minutes deviation when using the LSTM (51). NREM could be overestimated because it is the most common state among all participants on average, meaning that errors could be magnified. Accuracy estimates for our multimodal approach were in the high 70s for the majority of the classifiers, with LSTM (51) reaching 76% accuracy. These results are in line with the best performance previously reported in the literature. However, none of these previous studies had the scale and diversity that the MESA dataset offers [8].

Task 3 explored classification into Wake, REM, light sleep and deep sleep. In this classification, N1 and N2 were considered part of light sleep and N3 was classified as deep sleep. Actigraphy and HR/HRV reached accuracies of around 67% through LSTM (101). However, F_1 scores for the single modality approaches were between 35-50%. Similarly, both approaches overestimated time spent in light sleep and also struggled to pick up REM sleep. The multimodal approach outperformed the single modality approach with a higher accuracy of around 70% and an F_1 score of 52% for LSTM (51) (the highest performing model). Interestingly, LSTM (51) has a very strong performance at classifying wake and REM but struggles to discern light sleep and deep sleep, overestimating light sleep, probably given the transitional characteristics and high prevalence of the N2 stage nested in the light sleep class.

Task 4 aimed to evaluate classifiers that followed AASM scoring rules of Wake, REM, N1, N2 and N3. This task is the most complex of the four, given the high level of granularity required and the imbalance severity between sleep stages increased. Actigraphy and HR/HRV performances at this task were poor, with F_1 scores ranging from 27-36%. Both heavily overestimated the most prevalent state, N2. The multimodal approach struggled to discern among the different NREM stages and, again, overestimated time spent in N2. Its performance on Wake and REM was much better but, intriguingly, worse than what had been observed in Task 3. Accuracy did not exceed 64% and F_1 scores were between 40-42%. A visual illustration of this task is presented in Figure 4, where overestimation of N2 can be observed, alongside how the model struggles to discern transitions between N1, N2 and N3.

Across all multistage classification tasks LSTMs, outperformed every other modeling approach. This is most likely due to its ability to learn temporal dependencies from longer window sizes, contrasting with CNN models which focus on local dependencies. This makes LSTMs a particularly attractive candidate for multistage sleep classification given the intrinsic transitional nature of the task. The ensemble model approach serves as an example of how multistage classification benchmarks ought to be improved by new model architectures. The example approach is a rather simple one and thus only improves the performance marginally. Nevertheless, the results are promising. By incorporating different temporal domains and classifier types, these new models are able to pick up *nuances* that may have been tougher to identify by using a single convolutional or recurrent neural network.

Understanding sleep stage dynamics at a population level could be of value for digital health, epidemiology and clinical studies. We envision that research on behavioral change (i.e., [13]) can take advantage of ubiquitous systems for sleep stage classification to recommend changes for better sleep hygiene and healthier sleep architectures.

5.4 Physiological Underpinnings of Classifiers and Sensor Modality Contributions

Our classification tasks aimed to explore how the different modalities performed with regards to the level of granularity and detail generated. Physiologically, sleep stages are quite different and one objective of this work was to understand the individual contributions of each sensor, as well as model biases and preferences. Following the AASM staging convention:

(1) N1 (the first stage of NREM sleep) is the stage in which the change between wakefulness and sleep occurs. During this stage, heart rate, breathing and eye movement slow, with occasional muscle twitches. Similarly, slow-wave activity starts to appear on the PSG's EEG signal.
(2) N2 (the second stage of NREM sleep) is the transition period between light and deep sleep. Heartbeat and breathing slow, muscles relax even further and body temperature drops. This stage is the most repeated across all sleep cycles. Together with N1, it is often referred to as *light sleep*.
(3) N3 (the third stage of NREM sleep) is often referred to as *deep sleep*. Heart rate and breathing are at their lowest, muscles are very relaxed and it is rare for the person to awaken during this stage. These changes are also observed on the PSG's EEG signal, where the lowest frequency and highest amplitude waves can be found. Together with N1 and N2, this stage constitutes NREM sleep.
(4) Finally, as explored in the introduction, REM sleep occurs in a cyclical fashion, approximately every 90 minutes. Breathing is faster and irregular, heart rate increases and, in healthy people, the body is in a state of temporary paralysis that prevents sharp movements related to dreams.

Given the physiological differences between sleep stages, we hypothesised that depending on the sensors used and the measurements they enabled, performance at classifying certain stages may differ. This is of high importance when considering the deployment of these technologies in clinical settings or for the exploration of the association between sleep characteristics and disease end-points in population-based research. Understanding time

spent at different stages over a long-term period is of great importance for the greater sleep scientific community. For instance, during non-REM sleep, slow-wave activity has been shown to support memory consolidation [16] and reduce next-day anxiety [43]. In [16], for example, these slow-wave oscillations have been shown to affect the way the brain cerebrospinal fluid dynamics work, leading to oscillations in blood volume that draw this fluid across the blood-brain barrier.

We used SHAP to further understand how different sensor features contribute to the individual classification of sleep stages. Through this method, it became apparent that whilst actigraphy features were the most informative for sleep-wake classification, when moving to multistage classification tasks, HR/HRV features were also important. This is reflected in Figure 7 where the top features contribute more to the model than the bottom ones, indicating their higher predictive power. For example, frequency domain features were very informative for recognizing non-REM sleep. Similarly, the application of this method to the different tasks allows for the direct comparison of feature importance across different levels of sleep architecture granularity. When exploring SHAP results, for Task 1, we found that the most informative features came from actigraphy, although maximum HR and NN intervals were also notable contributors. Activity coming from the wrist actigraphy was particularly important for wake classification. This finding carried through all 4 tasks and makes sense given the considerably higher amount of movement present during wake than in any sleep stage unless a sleep disorder is present. For Tasks 2–4, SHAP results helped us understand why the multimodal approach performs significantly better ($p < 0.001$) than individual sensors at sleep-stage classification. We observed that while HRV features were not particularly informative for wake classification, they added a lot of value to NREM and REM predictions. Interestingly, we observed that frequency domain HRV features were amongst the most informative for light and REM sleep classification, confirming our initial hypothesis derived from previous clinical reports [22, 24]. These findings emphasise the importance of including HR/HRV measurements combined with movement for multistage classification tasks using wearable devices.

5.5 Future Work and Limitations

The strengths of this work derive from its novelty, population size and generalisability. It is the first systematic assessment of multistage sleep classification using non-obtrusive sensors. This makes an important contribution to the literature with potential applications for clinicians, researchers and the wellness industry. The population used is uniquely diverse, including a breadth of racial backgrounds, balanced sex and a representative sample of sleep-disordered participants. This enhances the generalization capabilities of the findings in contrast to previous studies [37, 56].

Here, we only explored benchmark models, but more complex networks can be designed, for example, by training deeper architectures, bi-directional architectures or using attention mechanisms. A natural next step for this line of research would be to incorporate multi-task learning approaches as well as models that combine a convolutional base with LSTMs/bi-directional LSTMs on deeper layers. Similarly, given the temporal dependencies of these tasks, attention mechanisms may be well suited to improve model performance [55]. Another architecture that may yield interesting results is the addition of a dense layer to merge representations learned by RNNs and CNNs, also exploiting the unique contributions of each classifier (temporal representations by the RNNs and spatial representations by CNNs). Similarly, one potential avenue that is an alternative to the current approach would be to keep the same window size while using different model stride sizes.

Given the scope of our work and objectives, we did not explore in detail how different models perform in diseased versus healthy populations, or highlight the differences between them. Similarly, our models did not exploit the well-known reciprocal interaction model first introduced by McCarley and Hobson [21], which describes ultradian periodicity, the approximately 90-min sleep cycle, which indicates that NREM-REM stage transitions are regulated by both cholinergic and monoaminergic neuronal structures. This inherent sleep

architecture shall be explored in future work to improve model performance. Furthermore, the ensemble model used in this paper is a just an example of how our benchmarks can be improved by using a novel approach tool, many other models can be used inspired by what has been already done in automatic sleep scoring in EEG, as the tasks will follow the same temporal dependencies. In our paper, we did not explore bespoke deep network architectures neither covered the various possible approaches for multi-modal fusion. Our future work may consider exploring these areas of research.

Furthermore, given the scope of this paper, we did not enforce strict quality control on the polysomnography data, which could have lead to the models to perform more poorly than if those practices and more stringent exclusion criteria had been applied. For instance, we found that a total of 30 subjects (about 2% of the total cohort) did not have any REM epochs at all. Similarly, on a small percentage of participants (less than 1%) accuracy scores were very low (< 45%). After post-hoc visual inspection of those cases, we found that their sleep patterns were abnormal and five of them had a reduced number of sleep transitions. However, for the purposes of this work, those participant results were included in the final performance metrics. To exclude non-wear time and activity measurement failure from actigraphy data, we used human noted tags as the selection criteria. Full processing pipelines shall integrate automated non-wear time and data corruption detection algorithms in the preprocessing phase.

One limitation of this work is that the MESA cohort only includes adult participants. Thus the results cannot be generalised to teenagers or children. Further, as with all studies in this area to-date, all inferences are derived in laboratory settings, whilst the potential applications are in a free-living environment. Good quality "ground truth" data collection in free-living environments is complex and expensive although it is interesting to explore the possibility of larger studies using ambulatory PSG or wireless EEG for this purpose. Similarly, commercial, non-research grade devices have been shown to be unreliable at collecting longitudinal sleep measures [6]. Thus, we encourage these companies to be more transparent about the way they collect data and the algorithms they use.

6 CONCLUSION

In conclusion, this work introduces a systematic benchmark approach to sleep-wake and sleep stage classification using ML and DL approaches in single modal and multimodal settings. This approach advocates for model transparency, alongside reproducibility by exploring these methods in the only open-access dataset, which includes diseased participants. The findings indicate that multimodal approaches combining movement with HR and HRV data were a valuable tool for the monitoring of sleep stages when those stages were aggregated to the level of NREM, REM, Wake. We further provide information regarding the performance of specific algorithms and guidance regarding algorithm selection depending on the classification tasks. Moreover, we introduce a deep ensemble model architecture which shows promising improvements in performance across the different multistage tasks explored. Overall, the findings highlight the promise of using wearable sensors as a low-burden, cheap and scalable approach for large, population-based studies. Future work should explore new model architectures to improve performance on more granular tasks, such as Tasks 3 and 4. These new architectures should aim to mimic the results obtained by EEG-based systems in both healthy and diseased populations as closely as possible. Furthermore, other auxiliary learning tasks and the inclusion of new, minimally obtrusive sensors may improve the performance of these models.

ACKNOWLEDGMENTS

REFERENCES

[1] Saeed Abdullah, Mark Matthews, Elizabeth L. Murnane, Geri Gay, and Tanzeem Choudhury. 2014. Towards Circadian Computing: "Early to Bed and Early to Rise" Makes Some of Us Unhealthy and Sleep Deprived. In *Proceedings of the 2014 ACM International Joint*

Conference on Pervasive and Ubiquitous Computing (UbiComp '14). Association for Computing Machinery, New York, NY, USA, 673–684. https://doi.org/10.1145/2632048.2632100

[2] M. Aktaruzzaman, M. Migliorini, M. Tenhunen, S. L. Himanen, A. M. Bianchi, and R. Sassi. 2015. The addition of entropy-based regularity parameters improves sleep stage classification based on heart rate variability. *Medical and Biological Engineering and Computing* 53, 5 (5 2015), 415–425. https://doi.org/10.1007/s11517-015-1249-z

[3] Emina Alickovic and Abdulhamit Subasi. 2018. Ensemble SVM Method for Automatic Sleep Stage Classification. *IEEE Transactions on Instrumentation and Measurement* 67, 6 (6 2018), 1258–1265. https://doi.org/10.1109/TIM.2018.2799059

[4] Bruce M Altevogt, Harvey R Colten, et al. 2006. *Sleep disorders and sleep deprivation: an unmet public health problem*. National Academies Press, Washington (DC). https://doi.org/10.17226/11617

[5] Salikh Bagaveyev and Diane J Cook. 2014. Designing and Evaluating Active Learning Methods for Activity Recognition. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 469–478. https://doi.org/10.1145/2638728.2641674

[6] Argelinda Baroni, Jean Marie Bruzzese, Christina A. Di Bartolo, and Jess P. Shatkin. 2016. Fitbit Flex: an unreliable device for longitudinal sleep measures in a non-clinical population. , 853–854 pages. https://doi.org/10.1007/s11325-015-1271-2

[7] Richard B. Berry, Rita Brooks, Charlene E. Gamaldo, Susan M. Harding, Robin M. Lloyd, Carole L. Marcus, and Bradley V. Vaughn. 2016. American Academy of Sleep Medicine. The AASM Manual for the Scoring of Sleep and Associated Events : Rules, Terminology, and Technical Specifications, Version 2.2. *American Academy of Sleep* 28, 3 (2016), 391–397. www.aasmnet.org.

[8] Diane E. Bild, David A Bluemke, Gregory L Burke, Robert Detrano, Ana V Diez Roux, Aaron R Folsom, Philip Greenland, David R. Jacobs, Richard Kronmal, Kiang Liu, Jennifer Clark Nelson, Daniel O'Leary, Mohammed F Saad, Steven Shea, Moyses Szklo, and Russell P Tracy. 2002. Multi-Ethnic Study of Atherosclerosis: Objectives and design. *American Journal of Epidemiology* 156, 9 (11 2002), 871–881. https://doi.org/10.1093/aje/kwf113

[9] M H Bonnet and D L Arand. 1997. Heart rate variability: sleep stage, time of night, and arousal influences. *Electroencephalography and Clinical Neurophysiology* 102, 5 (5 1997), 390–396. https://doi.org/10.1016/S0921-884X(96)96070-1

[10] Philippe Boudreau, Wei-Hsien Yeh, Guy A. Dumont, and Diane B. Boivin. 2013. Circadian Variation of Heart Rate Variability Across Sleep Stages. *Sleep* 36, 12 (12 2013), 1919–1928. https://doi.org/10.5665/sleep.3230

[11] Liqiong Chang, Jiaqi Lu, Ju Wang, Xiaojiang Chen, Dingyi Fang, Zhanyong Tang, Petteri Nurmi, and Zheng Wang. 2018. SleepGuard: Capturing Rich Sleep Information Using Smartwatch Sensing Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–34. https://doi.org/10.1145/3264908

[12] R J Cole, D F Kripke, W Gruen, D J Mullaney, and J C Gillin. 1992. Automatic sleep/wake identification from wrist activity. *Sleep* 15, 5 (10 1992), 461–9. http://www.ncbi.nlm.nih.gov/pubmed/1455130

[13] Nediyana Daskalova, Bongshin Lee, Jeff Huang, Chester Ni, and Jessica Lundin. 2018. Investigating the Effectiveness of Cohort-Based Sleep Recommendations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (9 2018), 1–19. https://doi.org/10.1145/3264911

[14] Dennis A. Dean, Ary L. Goldberger, Remo Mueller, Matthew Kim, Michael Rueschman, Daniel Mobley, Satya S. Sahoo, Catherine P. Jayapandian, Licong Cui, Michael G. Morrical, Susan Surovec, Guo-Qiang Zhang, and Susan Redline. 2016. Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. *Sleep* 39, 5 (5 2016), 1151–1164. https://doi.org/10.5665/sleep.5774

[15] Sigrid Elsenbruch, Michael J. Harnish, and William C. Orr. 1999. Heart Rate Variability During Waking and Sleep in Healthy Males and Females. *Sleep* 22, 8 (12 1999), 1067–1071. https://doi.org/10.1093/sleep/22.8.1067

[16] Nina E Fultz, Giorgio Bonmassar, Kawin Setsompop, Robert A Stickgold, Bruce R Rosen, Jonathan R Polimeni, and Laura D Lewis. 2019. Coupled electrophysiological, hemodynamic, and cerebrospinal fluid oscillations in human sleep. *Science (New York, N.Y.)* 366, 6465 (11 2019), 628–631. https://doi.org/10.1126/science.aax5440

[17] Jennifer Girschik, Lin Fritschi, Jane Heyworth, and Flavie Waters. 2012. Validation of self-reported sleep against actigraphy. *Journal of epidemiology* 22, 5 (2012), 462–8. https://doi.org/10.2188/jea.je20120012

[18] Yu Guan and Thomas Plötz. 2017. Ensembles of Deep LSTM Learners for Activity Recognition using Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (6 2017), 1–28. https://doi.org/10.1145/3090076

[19] Haodong Guo, Ling Chen, Liangying Peng, and Gencai Chen. 2016. Wearable Sensor Based Multimodal Human Activity Recognition Exploiting the Diversity of Classifier Ensemble. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 1112–1123. https://doi.org/10.1145/2971648.2971708

[20] Vincent T van Hees, Severine Sabia, Samuel E Jones, Andrew R Wood, Kirstie N Anderson, Mika Kivimaki, Timothy M Frayling, Allan I Pack, Maja Bucan, Diego R Mazzotti, Phil R Gehrman, Archana Singh-Manoux, and Michael N Weedon. 2018. Estimating sleep parameters using an accelerometer without sleep diary. *Scientific Reports* 8, 1 (2018), 12975. https://doi.org/10.1101/257972

[21] J. Allan Hobson, Robert W. McCarley, and Peter W. Wyzinski. 1975. Sleep Cycle Oscillation: Reciprocal Discharge by Two Brainstem Neuronal Groups. *Science* 189, 4196 (1975), 55–58. http://www.jstor.org/stable/1740806

[22] M Hornyak, M Cejnar, M Elam, M Matousek, and B G Wallin. 1991. Sympathetic muscle nerve activity during sleep in man. *Brain* 114 (Pt 3, 3 (6 1991), 1281–95. https://doi.org/10.1093/brain/114.3.1281

[23] Luca Imeri and Mark R Opp. 2009. How (and why) the immune system makes us sleep. , 199–210 pages. https://doi.org/10.1038/nrn2576

[24] D. A. Kirby and R. L. Verrier. 1989. Differential effects of sleep stage on coronary hemodynamic function. *American Journal of Physiology-Heart and Circulatory Physiology* 256, 5 (5 1989), H1378–H1383. https://doi.org/10.1152/ajpheart.1989.256.5.H1378

[25] B. Koley and D. Dey. 2012. An ensemble system for automatic sleep stage classification using single channel EEG signal. *Computers in Biology and Medicine* 42, 12 (12 2012), 1186–1195. https://doi.org/10.1016/j.compbiomed.2012.09.012

[26] Daniel F. Kripke, Elizabeth K. Hahn, Alexandra P. Grizas, Kep H. Wadiak, Richard T. Loving, J. Steven Poceta, Farhad F. Shadan, John W. Cronin, and Lawrence E. Kline. 2010. Wrist actigraphic scoring for sleep laboratory patients: Algorithm development. *Journal of Sleep Research* 19, 4 (12 2010), 612–619. https://doi.org/10.1111/j.1365-2869.2010.00835.x

[27] Jung-Min Lee, Wonwoo Byun, Alyssa Keill, Danae Dinkel, and Yaewon Seo. 2018. Comparison of Wearable Trackers' Ability to Estimate Sleep. *International journal of environmental research and public health* 15, 6 (2018), 1265. https://doi.org/10.3390/ijerph15061265

[28] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. 2018. Consistent Individualized Feature Attribution for Tree Ensembles. *ArXiv* abs/1802.03888 (2018), 1–9.

[29] Marek Malik. 1996. Heart Rate Variability. *Annals of Noninvasive Electrocardiology* 1, 2 (4 1996), 151–181. https://doi.org/10.1111/j.1542-474X.1996.tb00275.x

[30] Farid Melgani and Lorenzo Bruzzone. 2004. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing* 42, 8 (8 2004), 1778–1790. https://doi.org/10.1109/TGRS.2004.831865

[31] Luca Menghini, Evelyn Gianfranchi, Nicola Cellini, Elisabetta Patron, Mariaelena Tagliabue, and Michela Sarlo. 2019. Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions. *Psychophysiology* 56, 11 (2019), e13441. https://doi.org/10.1111/psyp.13441 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.13441

[32] Joao Palotti, Raghvendra Mall, Michael Aupetit, Michael Rueschman, Meghna Singh, Aarti Sathyanarayana, Shahrad Taheri, and Luis Fernandez-Luque. 2019. Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *npj Digital Medicine* 2, 1 (12 2019), 50. https://doi.org/10.1038/s41746-019-0126-9

[33] Sanjay R. Patel, Jia Weng, Michael Rueschman, Katherine A. Dudley, Jose S. Loredo, Yasmin Mossavar-Rahmani, Maricelle Ramirez, Alberto R. Ramos, Kathryn Reid, Ashley N. Seiger, Daniela Sotres-Alvarez, Phyllis C. Zee, and Rui Wang. 2015. Reproducibility of a Standardized Actigraphy Scoring Algorithm for Sleep in a US Hispanic/Latino Population. *Sleep* 38, 9 (9 2015), 1497–1503. https://doi.org/10.5665/sleep.4998

[34] Ignacio Perez-Pozuelo, Bing Zhai, Joao Palotti, Raghvendra Mall, Michaël Aupetit, Juan M Garcia-Gomez, Shahrad Taheri, Yu Guan, and Luis Fernandez-Luque. 2020. The future of sleep health: a data-driven revolution in sleep science and medicine. *NPJ digital medicine* 3, 1 (2020), 1–15.

[35] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chen, and Maarten De Vos. 2019. Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification. *IEEE Transactions on Biomedical Engineering* 66, 5 (5 2019), 1285–1296. https://doi.org/10.1109/TBME.2018.2872652

[36] Athi Ponnusamy, Jefferson L B Marques, and Markus Reuber. 2012. Comparison of heart rate variability parameters during complex partial seizures and psychogenic nonepileptic seizures. *Epilepsia* 53, 8 (8 2012), 1314–21. https://doi.org/10.1111/j.1528-1167.2012.03518.x

[37] Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M. Aarts. 2019. Sleep stage classification from heart-rate variability using long short-term memory neural networks. *Scientific Reports* 9, 1 (12 2019), 1–11. https://doi.org/10.1038/s41598-019-49703-y

[38] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K. Marina, and Fahim Kawsar. 2018. Multimodal Deep Learning for Activity and Context Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–27. https://doi.org/10.1145/3161174

[39] Avi Sadeh, P J Hauri, Daniel F Kripke, and P Lavie. 1995. The role of actigraphy in the evaluation of sleep disorders. *Sleep* 18, 4 (5 1995), 288–302. https://doi.org/10.1093/sleep/18.4.288

[40] Avi Sadeh, Katherine M Sharkey, and Mary A Carskadon. 1994. Activity-Based Sleep—Wake Identification: An Empirical Test of Methodological Issues. *Sleep* 17, 3 (1994), 201–207. https://doi.org/10.1093/sleep/17.3.201

[41] Edward Sazonov, Nadezhda Sazonova, Stephanie Schuckers, Michael Neuman, and CHIME Study Group. 2004. Activity-based sleep-wake identification in infants. *Physiological measurement* 25, 5 (10 2004), 1291–304. http://www.ncbi.nlm.nih.gov/pubmed/15535193

[42] Jonathan R L Schwartz and Thomas Roth. 2008. Neurophysiology of sleep and wakefulness: basic science and clinical implications. *Current neuropharmacology* 6, 4 (12 2008), 367–78. https://doi.org/10.2174/157015908787386050

[43] Eti Ben Simon, Aubrey Rossi, Allison G Harvey, and Matthew P Walker. 2020. Overanxious and underslept. *Nature Human Behaviour* 4, 1 (2020), 100–110.

[44] Urminder Singh, Sucheta Chauhan, A Krishnamachari, and Lovekesh Vig. 2015. Ensemble of deep long short term memory networks for labelling origin of replication sequences. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, Paris, France, 1–7. https://doi.org/10.1109/DSAA.2015.7344871

[45] Frederick Snyder, J. Allan Hobson, Donald F. Morrison, and Frederick Goldfrank. 1964. Changes in respiration, heart rate, and systolic blood pressure in human sleep. *Journal of Applied Physiology* 19, 3 (5 1964), 417–422. https://doi.org/10.1152/jappl.1964.19.3.417

[46] Virend K. Somers, Mark E. Dyken, Allyn L. Mark, and Francois M. Abboud. 1993. Sympathetic-Nerve Activity during Sleep in Normal Subjects. *New England Journal of Medicine* 328, 5 (2 1993), 303–307. https://doi.org/10.1056/NEJM199302043280502

[47] Jens B. Stephansen, Alexander N. Olesen, Mads Olsen, Aditya Ambati, Eileen B. Leary, Hyatt E. Moore, Oscar Carrillo, Ling Lin, Fang Han, Han Yan, Yun L. Sun, Yves Dauvilliers, Sabine Scholz, Lucie Barateau, Birgit Hogl, Ambra Stefani, Seung Chul Hong, Tae Won Kim, Fabio Pizza, Giuseppe Plazzi, Stefano Vandi, Elena Antelmi, Dimitri Perrin, Samuel T. Kuna, Paula K. Schweitzer, Clete Kushida, Paul E. Peppard, Helge B. D. Sorensen, Poul Jennum, and Emmanuel Mignot. 2018. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications* 9, 1 (12 2018), 5229. https://doi.org/10.1038/s41467-018-07229-3

[48] Xiao Sun, Li Qiu, Yibo Wu, Yeming Tang, and Guohong Cao. 2017. SleepMonitor: Monitoring Respiratory Rate and Body Position During Sleep Using Smartwatch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (9 2017), 1–22. https://doi.org/10.1145/3130969

[49] Hirofumi Tanaka, Kevin D. Monahan, and Douglas R. Seals. 2001. Age-predicted maximal heart rate revisited. *Journal of the American College of Cardiology* 37, 1 (2001), 153–156. https://doi.org/10.1016/S0735-1097(00)01054-8

[50] Elizabeth A. Thomson, Kayla Nuss, Ashley Comstock, Steven Reinwald, Sophie Blake, Richard E. Pimentel, Brian L. Tracy, and Kaigang Li. 2019. Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities. *Journal of Sports Sciences* 37, 12 (6 2019), 1411–1419. https://doi.org/10.1080/02640414.2018.1560644

[51] Joëlle Tilmanne, Jérôme Urbain, Mayuresh V Kothare, Alain Vande Wouwer, and Sanjeev V Kothare. 2009. Algorithms for sleep-wake identification using actigraphy: A comparative study and new results. *Journal of Sleep Research* 18, 1 (3 2009), 85–98. https://doi.org/10.1111/j.1365-2869.2008.00706.x

[52] Eleonora Tobaldini, Lino Nobili, Silvia Strada, Karina R. Casali, Alberto Braghiroli, and Nicola Montano. 2013. Heart rate variability in normal and pathological sleep. *Frontiers in Physiology* 4 (10 2013), 1–11. https://doi.org/10.3389/fphys.2013.00294

[53] Emilio Vanoli, Philip B. Adamson, Ba-Lin, Gian D. Pinna, Ralph Lazzara, and William C. Orr. 1995. Heart Rate Variability During Specific Sleep Stages. *Circulation* 91, 7 (4 1995), 1918–1922. https://doi.org/10.1161/01.CIR.91.7.1918

[54] A. Varri, Bob Kemp, Thomas Penzel, and A. Schlogl. 2001. Standards for biomedical signal databases. *IEEE Engineering in Medicine and Biology Magazine* 20, 3 (2001), 33–37. https://doi.org/10.1109/51.932722

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in neural information processing systems*. Curran Associates, Long Beach, CA, United States, 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[56] Emi Yuda, Yutaka Yoshida, Ryujiro Sasanabe, Haruhito Tanaka, Toshiaki Shiomi, Junichiro Hayano, Emi Yuda, Yutaka Yoshida, Ryujiro Sasanabe, Haruhito Tanaka, Toshiaki Shiomi, and Junichiro Hayano. 2017. Sleep Stage Classification by a Combination of Actigraphic and Heart Rate Signals. *Journal of Low Power Electronics and Applications* 7, 4 (11 2017), 28. https://doi.org/10.3390/jlpea7040028

[57] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. 2018. The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association* 25, 10 (10 2018), 1351–1358. https://doi.org/10.1093/jamia/ocy064

A EPOCH BY EPOCH PERFORMANCE METRICS

Table 11. Task 1-4 classification results by multimodal and single modality approaches, using epoch-by-epoch performance metrics. This table complements what was found on the main text and reported in Figures 6 and 7; Actigraphy modality: 🏃, HR/HRV modality: ♥; Three different tasks: Task 2: 3 stages, Task 3: 4 stages, Task 4: 5 Stages

Task 1: Wake, Sleep								
Method Specifics			**Performance Metrics**					
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's k
Multimodality	♥ 🏃 *[HR/HRV, Actigraphy]*	LSTM (51)	84.2	67.2	84.7	93.0	88.6	63.1
		LSTM (101)	84.2	67.1	84.6	93.0	88.6	63.1
		CNN (101)	84.1	66.2	84.3	93.3	88.6	62.7
		Maximum selection	**85.2**	68.2	85.2	**94.0**	**89.4**	65.3
		Mean over classifiers	85.2	**69.3**	**85.6**	93.4	89.3	**65.5**
Single Modality	♥ *[HR/HRV]*	LSTM (101)	**79.4**	**60.2**	**81.4**	89.2	**85.1**	**51.8**
		LSTM (51)	78.8	54.3	79.6	**91.4**	85.1	49.2
		CNN (101)	78.4	59.0	80.8	88.4	84.4	49.6
	🏃 *[Actigraphy]*	LSTM (101)	**84.7**	66.0	84.4	**94.3**	**89.1**	**63.9**
		CNN (101)	84.3	66.4	84.4	93.5	88.7	63.1
		LSTM (51)	84.3	**69.7**	**85.5**	91.7	88.5	63.6
Task 2: Wake, NREM, REM								
Method Specifics			**Performance Metrics**					
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's k
Multimodality	♥ 🏃 *[HR/HRV, Actigraphy]*	LSTM (51)	76.3	85.7	72.3	69.3	70.6	60.9
		LSTM (101)	76.1	85.4	72.9	67.3	69.3	60.2
		CNN (101)	75.9	85.6	71.5	70.0	70.4	61.0
		Mean over classifiers	**78.2**	**86.6**	**75.1**	**70.8**	**72.6**	**64.4**
		Maximum selection	77.9	86.6	74.2	70.8	72.2	63.9
Single Modality	♥ *[HR/HRV]*	LSTM (101)	**73.7**	**84.1**	**69.1**	**66.1**	**67.3**	**51.2**
		LSTM (51)	72.7	83.6	68.3	64.3	65.9	47.3
		CNN (101)	71.0	83.3	65.3	65.2	65.2	47.7
	🏃 *[Actigraphy]*	LSTM (101)	**71.3**	**80.8**	58.5	**52.9**	**50.5**	**52.5**
		CNN (101)	70.9	80.2	**76.5**	52.0	49.8	51.0
		LSTM (51)	70.8	80.4	48.4	52.3	49.8	51.0
Task 3: Wake, Light Sleep, Deep Sleep, REM								
Method Specifics			**Performance Metrics**					
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's k
Multimodality	♥ 🏃 *[HR/HRV, Actigraphy]*	LSTM (51)	70.4	87.7	65.5	54	54	56.8
		LSTM (101)	70.4	87.2	66	52.1	**54.2**	54.4
		CNN (101)	69.1	87.3	63.2	53.6	53.6	54.4
		Mean over classifiers	**71.7**	87.9	67.5	53.8	53.5	**58.8**
		Maximum selection	71.3	**88**	**68**	**54.4**	54.1	58.1
Single Modality	♥ *[HR/HRV]*	LSTM (101)	**67.4**	**86.3**	**62.9**	**50.9**	**52.2**	**46.9**
		LSTM (51)	66.2	85.8	61.9	48.9	49.4	43.4
		CNN (101)	64.4	85.5	59.2	49.8	49.6	43.4
	🏃 *[Actigraphy]*	LSTM (101)	**64.1**	**83.7**	34.4	**38.8**	**35.6**	**35.1**
		CNN (101)	64.0	83.7	**42.5**	38.8	35.6	35.1
		LSTM (51)	63.6	83.4	36.3	38.4	35.3	34.6
Task 4: Wake, REM, N1,N2,N3								
Method Specifics			**Performance Metrics**					
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's k
Multimodality	♥ 🏃 *[HR/HRV, Actigraphy]*	LSTM (51)	63.9	89.0	55.9	43.4	42.5	59.0
		LSTM (101)	63.8	89.0	55.6	43.3	43.2	59.7
		CNN (101)	63.2	89.0	55.6	44.8	**44.1**	58.4
		Mean over classifiers	**65.6**	**89.6**	**59.9**	44.7	42.8	**62.3**
		Maximum selection	65.3	89.6	58.4	**45.1**	43.0	61.7
Single Modality	♥ *[HR/HRV]*	CNN (101)	**55.6**	**86.7**	**48.3**	**38.4**	**38.8**	**38.1**
		CNN (21)	55.5	86.6	47.0	36.8	35.0	37.6
		CNN(50)	54.3	86.1	47.1	35.0	34.6	32.9
	🏃 *[Actigraphy]*	LSTM (101)	**57.0**	**86.4**	24.4	31.7	27.0	**50.0**
		CNN (101)	57.0	86.4	**32.1**	**31.8**	**27.2**	49.5
		LSTM (51)	57.0	86.3	29.8	31.6	27.0	49.9

B SLEEP STAGE CLASSIFICATION RESULTS MEASURED IN SLEEP PERIOD

Table 12. Sleep stage classification results during sleep period (mean ± standard error at 95% confidence interval and predicted minutes by multimodal and single modality approaches; Actigraphy modality: 🏃, HR/HRV modality: ♥; Three different tasks: Task 2: 3 stages, Task 3: 4 stages, Task 4: 5 Stages (*Average time deviation from ground truth across all subjects ± standard error)

Task 1: Wake, Sleep

Method Specifics			Performance Metrics						Time Deviation*	
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	Sleep
Multimodality	♥ 🏃	CNN (51)	**85.1 ± 1.1**	50.1 ± 2.5	88.7 ± 1.0	92.8 ± 1.2	90.0 ± 1.0	44.7 ± 2.2	-19.6 ± 5.8	19.6 ± 5.8
		CNN (101)	**85.1 ± 1.1**	**50.9 ± 2.5**	**88.9 ± 1.0**	92.5 ± 1.2	90.0 ± 1.0	**44.8 ± 2.2**	**-17.4 ± 5.9**	**17.4 ± 5.9**
		CNN (21)	84.9 ± 1.0	48.9 ± 2.2	88.3 ± 1.0	**93.1 ± 1.0**	**90.1 ± 0.9**	44.0 ± 2.0	-22.2 ± 5.4	22.2 ± 5.4
Single Modality	♥	CNN (51)	**81.7 ± 1.2**	43.0 ± 2.3	86.3 ± 1.2	**91.3 ± 1.3**	**87.9 ± 1.1**	35.3 ± 1.9	23.2 ± 7.3	-23.2 ± 7.3
		CNN (101)	81.7 ± 1.3	43.7 ± 2.3	86.5 ± 1.1	91.0 ± 1.4	87.8 ± 1.2	**35.7 ± 2.0**	21.4 ± 7.4	-21.4 ± 7.4
		LSTM (21)	80.9 ± 1.1	**48.4 ± 2.2**	**87.1 ± 1.2**	89.0 ± 1.2	87.3 ± 1.0	35.4 ± 1.8	**8.9 ± 6.8**	**-8.9 ± 6.8**
	🏃	CNN (101)	**85.5 ± 1.0**	46.9 ± 2.5	88.3 ± 1.0	**93.9 ± 0.9**	**90.5 ± 0.8**	43.5 ± 2.2	25.5 ± 5.5	-25.5 ± 5.5
		CNN (51)	85.3 ± 1.1	**50.7 ± 2.4**	**88.8 ± 1.0**	92.9 ± 1.1	90.2 ± 0.9	**45.4 ± 2.2**	19.7 ± 5.7	-19.7 ± 5.7
		LSTM (101)	83.9 ± 1.0	46.2 ± 2.5	88.5 ± 0.9	91.3 ± 1.1	89.4 ± 0.9	38.6 ± 2.3	**13.3 ± 5.4**	**-13.3 ± 5.4**

Task 2: Wake, NREM, REM

Method Specifics			Performance Metrics						Time Deviation*	
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	REM	NREM
Multimodality	♥ 🏃	CNN (101)	**74.9 ± 1.1**	**82.1 ± 0.8**	**68.7 ± 1.3**	64.5 ± 1.3	**62.9 ± 1.3**	**47.0 ± 2.0**	-14.4 ± 5.8	14.9 ± 6.7
		LSTM (51)	74.2 ± 1.1	81.7 ± 0.8	66.5 ± 1.4	62.8 ± 1.2	61.5 ± 1.3	45.1 ± 2.1	**0.3 ± 5.6**	12.2 ± 6.2
		CNN (51)	73.6 ± 1.2	82.5 ± 0.8	67.7 ± 1.3	**65.5 ± 1.3**	62.6 ± 1.4	47.0 ± 2.1	-4.7 ± 5.7	**-2.9 ± 7.5**
Single Modality	♥	LSTM (101)	**72.9 ± 1.3**	81.0 ± 0.8	**64.7 ± 1.5**	**60.1 ± 1.3**	**58.7 ± 1.5**	**37.2 ± 2.4**	-6.6 ± 7.0	**-11.8 ± 4.0**
		LSTM (51)	72.4 ± 1.2	**81.1 ± 0.8**	63.0 ± 1.5	59.4 ± 1.3	57.6 ± 1.4	32.9 ± 2.3	5.7 ± 6.9	-19.1 ± 4.1
		CNN (51)	71.1 ± 1.2	80.3 ± 0.8	63.6 ± 1.5	59.3 ± 1.3	56.6 ± 1.4	33.2 ± 2.1	**-5.5 ± 7.3**	-12.0 ± 5.5
	🏃	Linear SVM	**68.7 ± 1.0**	71.5 ± 0.7	**43.3 ± 1.1**	**43.9 ± 0.9**	**40.8 ± 0.9**	**25.1 ± 1.8**	-24.3 ± 6.1	-67.4 ± 3.0
		CNN (51)	68.4 ± 1.0	**71.6 ± 0.7**	42.2 ± 1.1	43.8 ± 0.9	40.4 ± 1.0	24.5 ± 2.0	**-17.4 ± 6.2**	**-67.3 ± 3.0**
		CNN (101)	68.4 ± 1.0	71.4 ± 0.7	42.5 ± 1.1	43.5 ± 0.9	40.1 ± 1.0	24.1 ± 2.0	-19.4 ± 6.3	**-67.3 ± 3.0**

Task 3: Wake, Light Sleep, Deep Sleep, REM

Method Specifics			Performance Metrics						Time Deviation*		
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	REM	Deep Sleep	Light Sleep
Multimodality	♥ 🏃	LSTM (51)	**66.5 ± 1.1**	**84.1 ± 0.7**	53.8 ± 1.4	50.2 ± 1.0	**47.4 ± 1.1**	**44.8 ± 2.2**	10.5 ± 5.8	-7.5 ± 3.9	-36.2 ± 3.5	**33.2 ± 6.7**
		LSTM (101)	**66.5 ± 1.1**	82.9 ± 0.6	**55.5 ± 1.6**	47.5 ± 1.1	46.2 ± 1.2	40.6 ± 2.1	-7.3 ± 5.5	-25.7 ± 3.6	**-32.5 ± 3.5**	65.5 ± 6.8
		CNN (101)	65.6 ± 1.2	83.7 ± 0.7	54.8 ± 1.5	**49.8 ± 1.1**	47.1 ± 1.1	42.6 ± 2.0	**-1.7 ± 6.3**	**1.5 ± 4.7**	-34.6 ± 3.5	34.7 ± 7.5
Single Modality	♥	LSTM (101)	**64.5 ± 1.3**	**83.2 ± 0.6**	**51.8 ± 1.6**	**47.0 ± 1.1**	**44.7 ± 1.3**	**35.0 ± 2.3**	**4.7 ± 7.0**	**-16.1 ± 3.6**	**-33.7 ± 3.5**	45.2 ± 7.2
		LSTM (51)	64.3 ± 1.2	83.1 ± 0.6	51.0 ± 1.5	46.5 ± 1.0	43.7 ± 1.2	33.2 ± 2.3	8.9 ± 6.6	-18.1 ± 3.9	-36.3 ± 3.5	45.5 ± 6.9
		LSTM (21)	62.8 ± 1.2	82.7 ± 0.6	47.9 ± 1.4	45.5 ± 0.9	42.0 ± 1.0	30.5 ± 2.0	16.7 ± 6.9	-18.8 ± 4.0	-38.5 ± 3.6	**40.6 ± 7.2**
	ENMO	LSTM (101)	**60.1 ± 1.1**	**77.3 ± 0.6**	30.1 ± 1.0	33.5 ± 0.8	29.5 ± 0.9	15.9 ± 1.3	-21.6 ± 5.5	-67.3 ± 3.0	**-39.2 ± 3.6**	**128.1 ± 7.3**
		CNN (101)	60.0 ± 1.1	**77.3 ± 0.6**	29.9 ± 1.0	33.4 ± 0.8	29.3 ± 0.9	16.2 ± 1.4	**-18.2 ± 5.7**	**-67.2 ± 3.0**	**-39.2 ± 3.6**	124.6 ± 7.5
		LSTM (51)	60.0 ± 1.1	**77.3 ± 0.6**	**30.3 ± 1.0**	**33.8 ± 0.8**	**29.7 ± 0.9**	**16.7 ± 1.4**	-18.2 ± 6.2	-67.3 ± 3.0	**-39.2 ± 3.6**	124.7 ± 7.7

Task 4: Wake, REM, N1, N2, N3

Method Specifics			Performance Metrics						Time Deviation*			
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	N3 Sleep	N2 Sleep	N1 Sleep
Multimodality	♥ 🏃	CNN (101)	**59.2 ± 1.2**	**86.5 ± 0.6**	**49.7 ± 1.4**	**42.0 ± 1.0**	**39.1 ± 1.0**	**45.1 ± 1.9**	-8.6 ± 5.8	**5.0 ± 4.8**	-34.3 ± 3.5	78.2 ± 7.4	**-40.3 ± 3.1**
		CNN (51)	58.8 ± 1.2	86.4 ± 0.6	46.6 ± 1.4	41.5 ± 0.9	37.8 ± 1.0	44.7 ± 2.0	**-4.7 ± 5.7**	11.5 ± 5.7	-37.9 ± 3.5	**73.1 ± 7.9**	-42.0 ± 3.1
		LSTM (101)	58.4 ± 1.1	85.9 ± 0.6	44.3 ± 1.4	39.9 ± 0.9	36.7 ± 1.0	43.5 ± 2.0	7.7 ± 5.7	-17.5 ± 3.8	**-32.3 ± 3.5**	88.0 ± 7.1	-45.9 ± 3.0
Single Modality	♥	CNN (21)	**53.8 ± 1.3**	85.0 ± 0.6	38.9 ± 1.3	36.7 ± 0.9	32.1 ± 1.0	**28.7 ± 1.8**	**18.8 ± 8.2**	**-9.4 ± 5.4**	-39.0 ± 3.6	74.0 ± 8.7	-44.4 ± 3.0
		CNN (101)	52.4 ± 1.3	**85.1 ± 0.6**	**42.8 ± 1.4**	**37.6 ± 1.0**	**33.6 ± 1.1**	26.8 ± 1.9	28.9 ± 9.0	-15.7 ± 4.5	**-29.8 ± 3.5**	57.6 ± 8.4	-41.0 ± 3.1
		CNN (51)	50.4 ± 1.3	84.6 ± 0.6	39.2 ± 1.4	35.2 ± 0.9	30.0 ± 1.1	22.9 ± 1.9	62.0 ± 10.4	-30.8 ± 4.8	-36.4 ± 3.5	**46.2 ± 9.8**	**-40.9 ± 3.0**
	🏃	LSTM (51)	**51.3 ± 1.1**	**82.0 ± 0.6**	**21.8 ± 0.9**	28.2 ± 0.8	**22.7 ± 0.8**	27.8 ± 1.9	**2.3 ± 6.2**	-67.3 ± 3.0	-39.2 ± 3.6	153.3 ± 8.1	**-49.1 ± 3.2**
		CNN (21)	50.9 ± 1.2	**82.0 ± 0.6**	21.5 ± 0.9	28.1 ± 0.7	22.5 ± 0.8	27.4 ± 1.8	4.6 ± 6.5	**-67.1 ± 3.0**	-39.2 ± 3.6	150.7 ± 8.1	**-49.1 ± 3.2**
		LSTM (21)	50.9 ± 1.1	**82.0 ± 0.6**	21.1 ± 0.9	**28.3 ± 0.8**	**22.7 ± 0.8**	**27.8 ± 1.8**	5.1 ± 6.1	-67.4 ± 3.0	**-39.1 ± 3.6**	**150.5 ± 7.9**	**-49.1 ± 3.2**

C HYPERPARAMETERS TUNING AND RESULTS

Table 13. Hyper-parameters for ML and DL algorithms

Tree Approaches						
Algorithms	**Task**	**Number of trees**	**Number of Features For the Best Split**	**Criterion**	**Use Out-of-bag Samples**	
Random Forest	All tasks	100, 200, 300	20	Gini	No	
Best Hyperparameters						
Random Forest	All Tasks	300	20	Gini	No	
Shallow Machine Learning Approaches						
Algorithms	**Task**	**Alpha**	**Fit Intercept**	**Max Iteration**	**Classifier Penalty**	**Sliding Window Length**
Linear SVM Perceptron Logistic Regression	All tasks	1.e-01, 1.e-02, 1.e-03, 1.e-04, 1.e-05, 1.e-06	True, False	5, 10 , 20	L1, L2, Elastic net	Actigraphy = 20 sleep epochs HR/HRV = 1 sleep epoch
Best Hyperparameters Used In Study						
Linear SVM	All tasks	1.e-3	True	5	L2	Actigraphy = 20 sleep epochs HR/HRV = 1 sleep epoch
Perceptron	All tasks	1.e-1	False	5	L2	Actigraphy = 20 sleep epochs HR/HRV = 1 sleep epoch
Logistic Regression	All tasks	1.e-4	True	20	L2	Actigraphy = 20 sleep epochs HR/HRV = 1 sleep epoch
Deep Neural Network Approaches						
Algorithms	**Task**	**Number of Layers**	**Number of Kernels (CNN)** **Hidden Units (LSTM)**	**Kernel Length**	**Optimiser**	**Window Length**
CNN (1D-Conv)	All tasks	1, 2, 3	32, 64, 128	3, 5, 7	RMSprop	21, 51, 101
LSTM (Many-to-one)	All tasks	1, 2, 3	64, 128, 256	N/A	RMSprop	21, 51, 101
Best Hyperparameters						
CNN (1D-Conv)	Task1	1	128	7	RMSprop	101
LSTM (Many-to-one)		3	128	N/A	RMSprop	51
CNN (1D-Conv)	Task2	3	64	5	RMSprop	101
LSTM (Many-to-one)		3	64	N/A	RMSprop	101
CNN (1D-Conv)	Task3	3	64	3	RMSprop	101
LSTM (Many-to-one)		3	128	N/A	RMSprop	101
CNN (1D-Conv)	Task4	3	64	5	RMSprop	101
LSTM (Many-to-one)		3	64	N/A	RMSprop	101
Hyperparameters Used In Study						
CNN (1D-Conv)	All tasks	1	64	2	RMSprop	21, 51, 101
LSTM (Many-to-one)		1	32	N/A	RMSprop	21, 51, 101

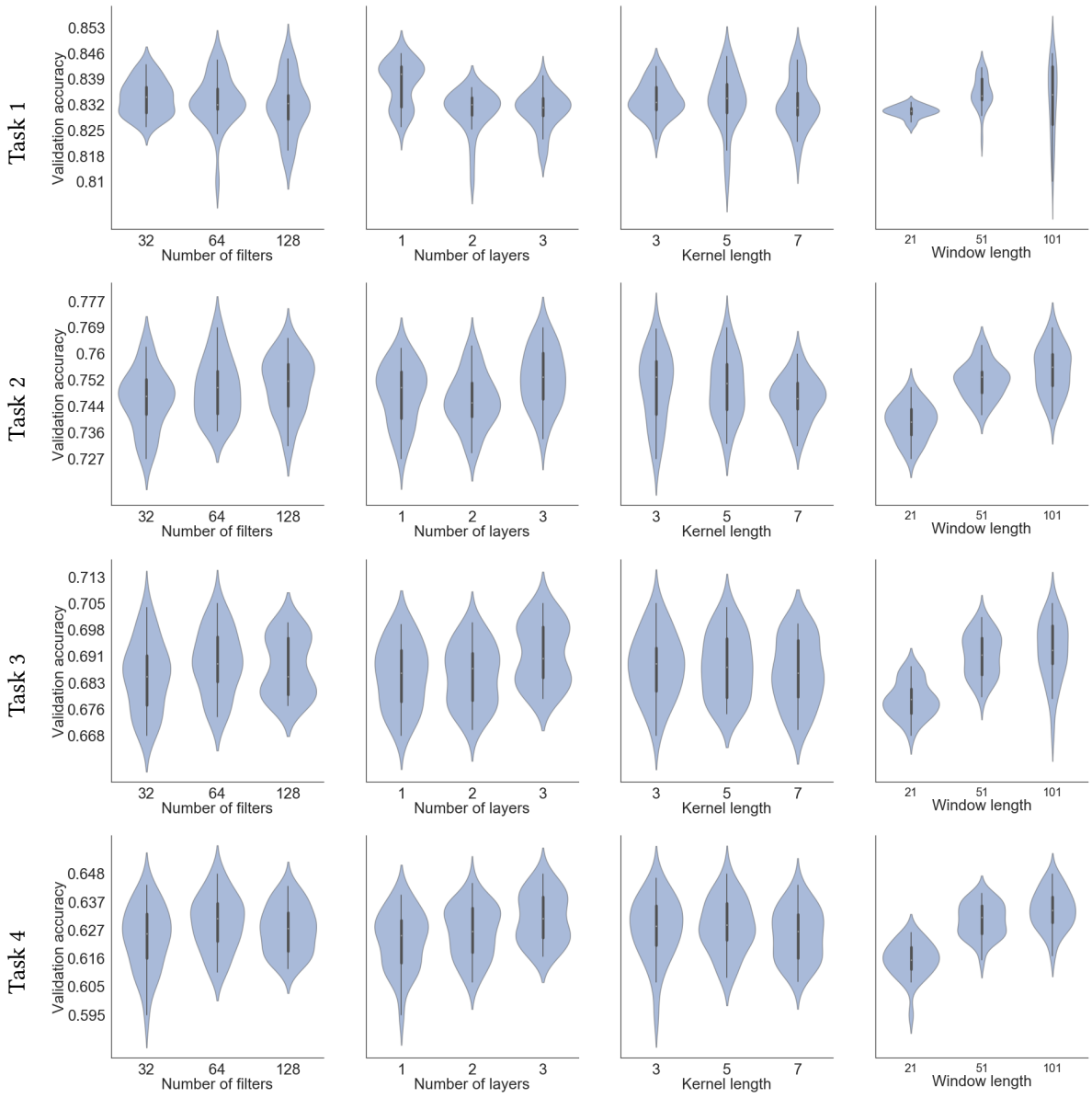


Fig. 8. CNN hyper-parameters tuning results

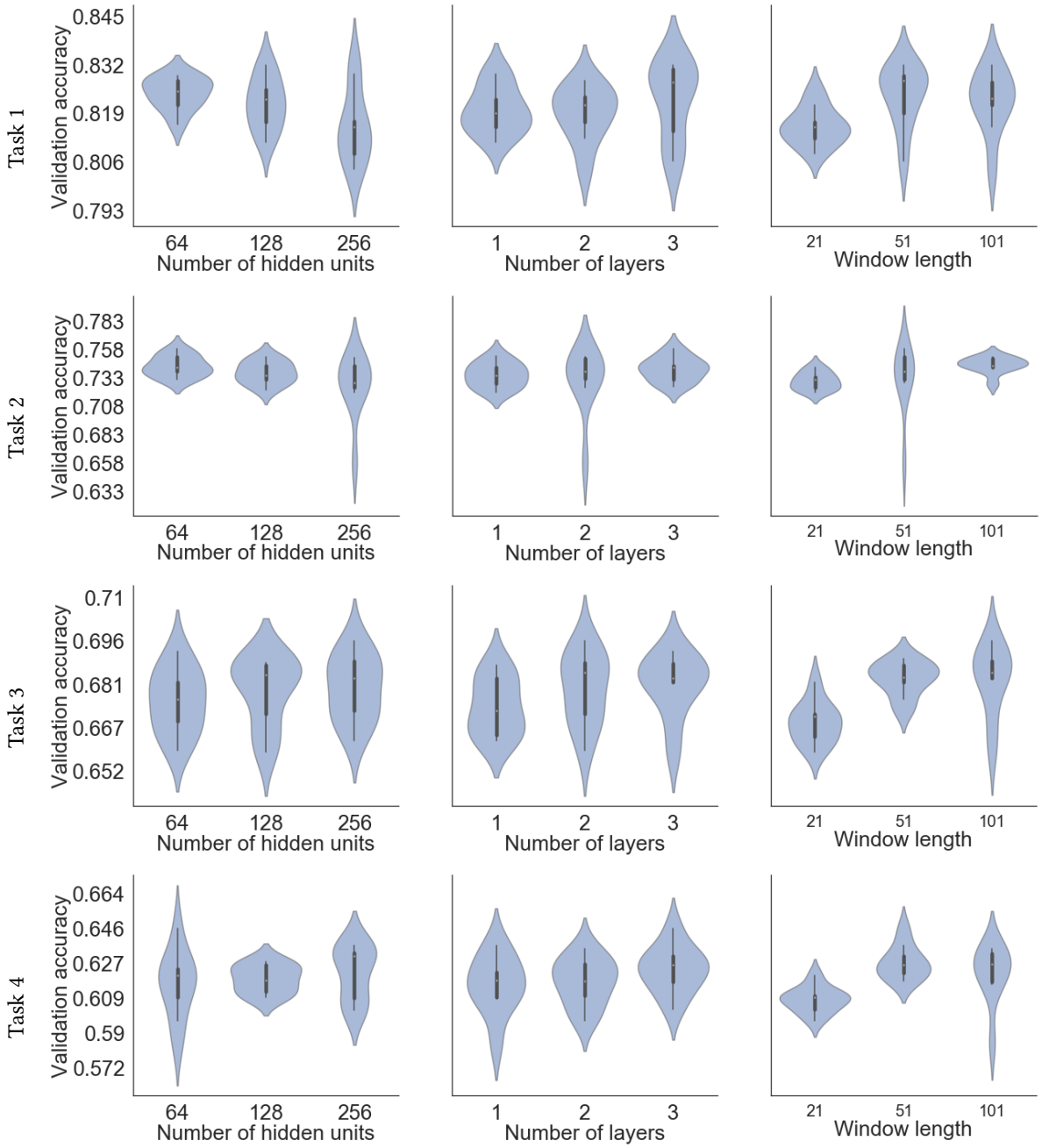


Fig. 9. LSTM hyper-parameters tuning results