Non-contact Pressure-based Sleep/Wake Discrimination

Lorcan Walsh, Member, IEEE, Seán McLoone, Senior Member, IEEE, Joseph Ronda, Jeanne F. Duffy, Charles A. Czeisler

Abstract—Poor sleep is increasingly being recognised as an important prognostic parameter of health. For those with suspected sleep disorders, patients are referred to sleep clinics which guide treatment. However, sleep clinics are not always a viable option due to their high cost, a lack of experienced practitioners, lengthy waiting lists and an unrepresentative sleeping environment. A home-based non-contact sleep/wake monitoring system may be used as a guide for treatment potentially stratifying patients by clinical need or highlighting longitudinal changes in sleep and nocturnal patterns. This paper presents the evaluation of an under-mattress sleep monitoring system for non-contact sleep/wake discrimination. A large dataset of sensor data with concomitant sleep/wake state was collected from both younger and older adults participating in a circadian sleep study. A thorough training/testing/validation procedure was configured and optimised feature extraction and sleep/wake discrimination algorithms evaluated both within and across the two cohorts. An accuracy, sensitivity and specificity of 74.3%, 95.5%, and 53.2% is reported over all subjects using an external validation dataset (71.9%, 87.9% and 56%, and 77.5%, 98% and 57% is reported for younger and older subjects respectively). These results compare favourably with similar research, however this system provides an ambient alternative suitable for long term continuous sleep monitoring, particularly amongst vulnerable populations.

Keywords-Sleep, polysomnography, actigraphy, ambient intelligence, feature extraction, machine learning.

I. INTRODUCTION

SLEEP is a fundamental physiological process with important restorative functions. It occurs in all living mammals and generally over a significant portion of each day [1]. Sleep problems have been shown to be detrimental to human health. In humans, short (seven hours or less) and long (nine hours or more) durations of sleep have been shown to be significant predictors of death in prospective population studies [2]. Sleep disturbances may be indicative of poor health and functional deficits, especially in older adults [3], [4]. Total sleep time is reduced in the elderly and this is not due to a reduced need for sleep, but in a diminished ability to sleep [5]. Sleep complaints are commonly reported by over 50% of those aged 65 and older [4]. These complaints include getting less sleep, frequent awakenings, waking up too early, excessive daytime sleepiness, and napping during the day. Decreased quality of life, and higher rates of depression and anxiety are reported in patients with sleeping difficulties [6]. High incidences of balance, ambulatory and visual difficulties (after controlling for medication use) have been reported in older adults with sleep problems [7]. Furthermore, decreased total sleep time (TST), an increased sleep latency (SL), defined as the time taken to fall asleep in bed, and a poor sleep efficiency (SE), defined as the percentage of TST over total time in bed (TIB), are linked to a greater risk for mortality (even when controlling for related covariates) [8]. Additionally, the symptoms of various chronic conditions continue into the night and result in a disturbed sleep; these include movement disorders, neuromuscular diseases, depression, dementia, epilepsy, obesity and circadian rhythm disorders [9].

The gold standard sleep assessment technology is polysomnography (PSG) which records multiple physiological signals (including brain activity, muscle tone, eye movements, heart rate and respiration) during sleep. This is generally performed in a sleep clinic and a trained sleep scorer uses a strictly defined set of rules to classify each 30 second epoch into either wake or a variety of sleep stages including rapid eye movement (REM) and non-REM (NREM) sleep [10]. However the application of these rules is subjective and an inter-rater agreement rate of 82% has been reported using data from multiple subjects and across separate sleep laboratories [11]. Additionally, PSG is intrusive, costly, time consuming and often alienates the patient. Wrist actigraphy is the current ambulatory gold standard sleep monitoring device. It consists of a two axis accelerometer which records the rest/activity patterns of the wearer and converts an activity metric to sleep/wake estimates using a thresholding algorithm and some additional logic. Wrist actigraphy has been shown to estimate nocturnal sleep duration and sleep-wake patterns reliably where PSG is not a suitable alternative [12]. However, a low wake detection capacity is often reported with this device as the device cannot discriminate between quiescent wake and sleep [13]. Wrist actigraphy is dependent on the adherence (and active participation) of the wearer. Sleep diaries are also used (often concomitantly with wrist actigraphy) to estimate sleep duration in normal, institutionalised and pediatric populations. However their validity relies upon the attentiveness of the individual filling the diaries out (in cases where the diaries are filled out by the individual). A trade-off exists for these technologies between accuracy and suitability for long term deployment. This paper proposes an ambient sleep/wake and bed-movement sensor suitable for extended home deployments to investigate long-term changes in sleeping patterns, especially...
amongst vulnerable populations such as those with memory impairment. Existing systems require either a worn sensor or a sensing device visible within the bedroom which may result in stigma due to the need for such technology.

In this contribution, a non-contact unobtrusive under-mattress bed sensor (UMBS) is proposed as a means of: (1) quantifying bed-based movements, and; (2) discriminating between sleep and wake. This contribution builds on previous research where novel algorithms were found to reliably measure breathing rate and body movement from UMBS data [14], and a preliminary study of a cohort of community based older adults [15] that showed high correlation between wrist actigraphy and motion metrics derived from UMBS data. This paper initially details a number of movement features which may be extracted during sleep building on previous research [16]. Subsequently, a classification setup is detailed which optimises the reliability and accuracy of the discrimination of sleep and wake using a large and rigorous research clinical-based dataset. It is proposed that the combination of these data provide a holistic description of sleep using a non-invasive sensor which can be deployed over extended periods in the community.

The remainder of this paper is organised as follows. Section II provides a literature review. Section III introduces the UMBS methodology, and techniques used to extract sleep/wake state from the UMBS data is given in Section IV. Results are described in Section V. A discussion and conclusion are given in Sections VI and VII, respectively.

II. LITERATURE REVIEW

Proposed approaches for home-based and/or long term monitoring range from ambient (that is, embedded in the environment), worn, and/or image-based solutions. There is a wide disparity between how these systems are validated including methodological differences, the formulation of the gold standard comparison dataset, the potential use of an external test dataset, and also the inconsistent and often non-overlapping performance measures reported across these systems.

Video-based sleep monitoring solutions have been proposed [17], [18] and while their utility is impressive participants are often uncomfortable with the presence of video recording equipment in the home and especially in the bedroom [19]. Privacy concerns were diminished only when unfavourable alternatives (such as nursing homes) were considered [19]. However, other technologies may provide sufficient utility while retaining privacy.

Passive infra-red (PIR) based monitoring systems [20] have been developed and a high accuracy reported, but a number of potential usability issues remain including the varying location of heaters between environments and the presence and types of bed sheets shielding sensors from the minute movements of the individual.

Radar based technologies [21] report high accuracies in detecting movement compared to wrist actigraphy. Further work reported a high accuracy in sleep/wake discrimination (87% sensitivity, 50% specificity) across a test dataset of 113 nights [22]. Load-cell movement detection sensors [23], [24] have been shown to have a high capacity to detect respiration and movement although this may depend on the orientation of the individual.

Bed monitoring systems using pressure pads placed on top of the mattress, underneath the bed sheets, have been developed which detect the ballistocardiogram and detect the heart rate and respiratory rate accurately [25], [26]. Under mattress sensors [27]–[31] have been shown to measure respiration or both respiration and heart rate effectively. A pressure sensitive e-textile bed sheet reported high accuracy in sleep/wake discrimination (70% precision, 71% recall) with 7 nights of data [32]. Cardiorespiratory and movement-based signals from a respiratory inductance plethysmography band, electrocardiogram and a pressure sensitive bed were used to discriminate between sleep and wake: 92% accuracy, wake/REM/NREM: 81% and all sleep stages: 69% using 85 nights of sleep [33]. Actigraphic and respiratory effort-derived features and a linear discriminant classifier reported an accuracy of 95.7% using 15 nights of data [34]. Activity and cardiovascular data were used to generate a sleep/wake accuracy of 78.3% using 20 nights of data and a support vector machine and hidden markov model-based system [35].

Pillow based sensors [36] have been proposed as solutions to domestic non-contact sleep monitoring by measuring heart and respiration rate. These systems are particularly suited to non-contact long term sleep monitoring as they do not require specialist expertise to install.

Home-based sleep monitoring has not been forgotten in the burgeoning wave of Internet of Things and Quantified Self devices including the Withings Aura [37], Fitbit [38], Xiaomi Mi Band [39] and smartphone applications like Sleep Cycle Alarm Clock [40]. A comparison of these and similar commercial personal health monitoring devices against global sleep measures (via PSG) in healthy populations has found no significant differences between them in terms of their ability to estimate total sleep time [41]. However, it is reasonable to assume that a simple in/out of bed sensor may provide similar accuracy levels in terms of total sleep time. An epoch-by-epoch comparison is necessary to provide a more rigorous evaluation of the performance of these systems. Furthermore, healthy adult users were found to be unreliable for any devices which required the user to initiate a sleep tracking setting [41]. Therefore, when moving to long-term sleep monitoring, particularly in populations with potential memory issues, an ambient technology requiring no user interaction is desirable.

III. UNDER MATTRESS BED SENSOR

The UMBS system is composed of an under mattress pressure sensing grid, a data collection system and algorithms to extract respiration and body movement information, or features. The process of collecting and extracting these features is described below.
The evolution of the pressure signal for three unsaturated tactels is given in Figure 2. Tactel values are initially low (less than 500 units) and raise upon a bed entry event. Small cyclic variations in pressure can be observed in two of the three tactels which correspond to respiration and in some cases heart rate related movement [14]. Over extended periods of time, body posture changes, or macro-movements, can also be identified.

For a more thorough description of the UMBS, including a description of the extraction of respiration rate and body movement, the reader is referred to previous work [14], [16]. Other work elsewhere is examining the use of the UMBS for in-bed mobility monitoring during postural transitions [42].

B. Feature Extraction Algorithms

Four types of movement features were derived from the UMBS data: respiration rate, statistical, temporal and spatial (as described in Table I). A brief description of the feature extraction algorithms is given below:

1) Respiration: A method of extracting respiration rates using the UMBS has been developed and validated against a strain gauge [14]. In this work, a custom respiration rate detection algorithm yielded a mean difference of 0.12 breaths per five minutes and a mean percentage error (MPE) of 0.16% when the sensor was placed beneath the mattress. The algorithm removes the effect of slow postural changes and unwanted spectral noise on the UMBS using filtering, fuses the data from active signals and employs an optimised peak detection algorithm to detect respiration rate. Results are reported on the basis of a 5 minute window to ensure a relatively high number of respiration cycles occur.

2) Statistical Movement Features: Defining the difference between successive tactel values as

\[ \delta_{ij} = |x_{ij} - x_{(i-1)j}| \quad \forall 2 \leq i \leq N, 1 \leq j \leq 24 \]  

(1)
where $N$ is the number of samples in an epoch, $x_{ij}$ is the value of the $j^{th}$ tactel at the $i^{th}$ sample instant, and $\delta_{ij} = 0$
for $i=1$ and $1 \leq j \leq 24$, an overall UMBS motion metric can be computed as
\[
Act = \sum_{j=1}^{24} \left( \sum_{i=2}^{N} \delta_{ij} \right)
\]
(2)

This metric, which is similar in form to the activity motion count calculated by wrist actigraphy, is computed over a predefined epoch length of 60 seconds.

3) Temporal Movement Features: The mean of $\delta_{ij}$ over all tactels for each time instant $i$, denoted $TMF(i)$, was used as a temporal movement feature to track the magnitude of motion registered by each tactel over time. TMF is a representation of the total applied force without consideration of the number of tactels experiencing motion. Hence,
\[
TMF(i) = \frac{1}{24} \sum_{j=1}^{24} |\delta_{ij}|
\]
(3)

Based on an empirical evaluation, determined by experimentally gathered movement and non-movement data, the minimal threshold for movement detection, as measured using TMF, was determined to be 4. In this investigation, background non-movement activity (such as breathing, etc.) generally occurred at TMF less than 1 (where 94% of values below a TMF of 4 occurred during non-movement times). An illustrative example of movement and non-movement using both the UMBS data and the TMF metric is included as supplementary material.

4) Spatial Movement Features: A continuous description of spatial displacement (resultant from postural changes) in bed was derived by examining the pressure applied, and changes therein, to the individual UMBS tactels. This spatial movement feature (SMF) measures the change in position of the subject over time, but does not measure movement where no change in position occurs (TMF should reflect this). Tactels recording instantaneous body posture were defined as those registering a value of greater than 500 s tactels registered less than 500 when in their resting state (i.e. when unloaded). A centre of pressure (COP) metric was derived as the central point in the lateral direction through which pressure was applied. In order to increase the accuracy of COP measurement, linear interpolation was employed between neighbouring real tactels and virtual tactels created to yield a total of 71 equally spaced measurement points across the mat. The centre of pressure metric was then computed as
\[
COP_x = \frac{\sum_{x=1}^{71} \sum_{y=1}^{3} (Pos_x \times status_{x,y})}{\sum_{x=1}^{71} \sum_{y=1}^{3} status_{x,y}}
\]
(4)

where $status_{x,y}$ is a boolean value indicating whether or not pressure is being applied at point $(x, y)$, that is
\[
status_{x,y} = \begin{cases} 
0 & x_{x,y} < 500 \\
1 & Otherwise
\end{cases}
\]
and $Pos_x$ is a x-coordinate of the current measurement point $(x, y)$. COP$_x$ was not deemed to provide useful information when examined independently, however changes in COP$_x$ over time provide a spatial measurement of in-bed movement. The first derivative of the COP$_x$ data was used to quantify such movement, that is:
\[
SMF_1 = COP_x(i) - COP_x(i-1) \quad 2 \leq i \leq N
\]
(6)

where $SMF_1$ is 0.

IV. METHODS

This section describes how data was collected during the clinical studies, details the features extracted from the data, gives an overview of the classifiers used and how they were applied, and how the performance of the system was measured.

A. Study Protocol

Data were collected as part of two research protocols carried out by the Division of Sleep and Circadian Disorders in the Intensive Physiological Monitoring Unit of the Center for Clinical Investigation at the Brigham and Women’s Hospital, Boston, MA, USA. Full ethical approval was granted by the Partners Health Care Human Research Committee. UMBS data were collected using the system described in Section III. PSG data was collected for each scheduled sleep episode, including EEG (C3, C4), EOG, EMG (submental), and ECG using a Viataport Digital Sleep Recorder (Temec Instruments, Kerkrade, Netherlands) and scored according to the Rechtschaffen and Kales guidelines [10] in 30s epochs. Signals were sampled at 256 Hz, low-pass filtered at 30Hz and high-pass filtered at 0.159Hz (a time constant of 1.0 seconds), and stored at 128 Hz. Two 30 second PSG epochs were concatenated and taken as one 60 second epoch (where there was no change between sleep and wake) for the remainder of this analysis. A consistent lab time was used to ensure concomitant PSG and UMBS time. The electrodes were applied approximately two hours prior to the scheduled lights out and recording began approximately one hour prior to the scheduled lights out. Data during the lights out period was used in this analysis. The PSG data were scored according to the standard criterion [10] into one of seven states: wake, Non-REM sleep (stage 1, stage 2, stage 3 and stage 4), REM sleep or movement. All sleep stages were defined as sleep in this analysis. Movement epochs related to periods in which motion artefacts obscured more than half ($\geq 15$ sec) of the epoch and was re-scored as wake in later analysis. Thus, all epochs analysed in this study were either scored as sleep or wake.

The UMBS study was placed within a broader research program where sleep and waking were scheduled to occur in controlled laboratory conditions, in some cases on a non-24-hour schedule, in order to better understand circadian control of sleep and waking in healthy adults. Each participant in this protocol spent multiple days in the laboratory in which their sleep episodes were scheduled by the experimenters. The subject was instructed to sleep, or attempt to sleep, during lights off episodes, and was required to remain in bed in the dark until scheduled lights on. This process of scheduling the sleep and wake times was external to this analysis. An example
of the schedules used in the two experiments is shown in Figure 3. The subjects were continually monitored by a team of specialist healthcare clinicians and technicians throughout both sleep and wake episodes. Data was collected under two protocols. For both protocols, data from the initial 21 days were used for analysis, however the studies ran beyond this duration.

a) Protocol I: In this protocol, the natural nocturnal sleep length was extended to twelve hours during the initial three days of the study (see Figure 3). A midday nap was scheduled, however sleep during these naps were not included in the present analysis. The study began with extended sleep times and naps in order to ensure the subject had no short term sleep debt. Over the next three days, the scheduled sleep opportunity was ten hours. Over the remaining thirteen days the day length was extended to a 28 hour schedule, while the sleep length was reduced to six and a half hours per 28 hour day. This protocol meant that subjects progressively accumulated sleep loss. As such, the wake time during scheduled sleep episodes diminished as the protocol continued.

b) Protocol II: Baseline data were collected during the initial 3 8-hour nights of the study (see Figure 3). Subsequently the subjects were kept awake for a period of forty hours. Subjects were then scheduled to sleep for 9.33 hours. The day length was then modified to a 28 hour schedule, with the sleep length reduced to six and a half hours per 28 hour day. This protocol meant that subjects progressively accumulated sleep loss.

B. Subject Details

Data, collected from 30 people, were split into 5 datasets for this analysis as detailed in Table II. Subjects were both healthy younger and older adults and were split between protocol I and protocol II. Datasets A (8 younger adults) and B (7 older adults) contained at least 13 sleep records each. These data were used to train and validate the classification algorithms. The total distribution of sleep stages per subject over the sleep episodes is given in Figure 4. Datasets C (8 younger adults) and E (1 older adult) contained data from one and three sleep records respectively and were used to test the performance of the optimal classification algorithm on unseen subjects. The cohort in dataset D did not have any corresponding PSG data and was used to assess the correlation in the underlying data. The mean and standard deviation of the sleep efficiency over all sleep episodes where data was collected for each subject is also given in Table II

![Double raster plot for both protocols (black represents scheduled periods of sleep).](image)

![Distribution of sleep stages per subject over all sleep episodes.](image)

TABLE II. DATA SETS FOR CLASSIFIER TRAINING AND VALIDATION.

<table>
<thead>
<tr>
<th>UMBS Subdataset</th>
<th>Study</th>
<th>Age (yrs)</th>
<th>Sex</th>
<th>Records scored</th>
<th>Sleep efficiency mean ± std. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set A: Healthy Younger Adults; for classifier training and validation</td>
<td>BN101 Protocol II</td>
<td>19 M</td>
<td>13</td>
<td>13.14 ± 7.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN103 Protocol II</td>
<td>23 M</td>
<td>13</td>
<td>9.47 ± 6.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN104 Protocol II</td>
<td>24 M</td>
<td>13</td>
<td>14.75 ± 10.31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN105 Protocol II</td>
<td>19 F</td>
<td>13</td>
<td>10.09 ± 5.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN106 Protocol II</td>
<td>21 M</td>
<td>13</td>
<td>15.12 ± 3.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN107 Protocol I</td>
<td>20 F</td>
<td>13</td>
<td>9.51 ± 7.74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN108 Protocol I</td>
<td>24 F</td>
<td>13</td>
<td>3.93 ± 2.23</td>
<td></td>
</tr>
<tr>
<td>Data Set B: Healthy Older Adults; for classifier training and validation</td>
<td>BN109 Protocol I</td>
<td>56 F</td>
<td>14</td>
<td>22.02 ± 10.97</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN110 Protocol I</td>
<td>70 M</td>
<td>14</td>
<td>33.04 ± 13.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN111 Protocol I</td>
<td>56 M</td>
<td>14</td>
<td>21.0 ± 12.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN112 Protocol I</td>
<td>58 F</td>
<td>14</td>
<td>15.52 ± 8.89</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN114 Protocol I</td>
<td>64 M</td>
<td>14</td>
<td>5.63 ± 4.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN115 Protocol I</td>
<td>55 M</td>
<td>14</td>
<td>13.48 ± 17.47</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN116 Protocol I</td>
<td>60 F</td>
<td>14</td>
<td>25.29 ± 12.36</td>
<td></td>
</tr>
<tr>
<td>Data Set C: Healthy Younger Adults; for classifier testing</td>
<td>BN117 Protocol I</td>
<td>19 F</td>
<td>1</td>
<td>7.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN118 Protocol I</td>
<td>18 F</td>
<td>1</td>
<td>16.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN119 Protocol I</td>
<td>27 M</td>
<td>1</td>
<td>22.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN120 Protocol I</td>
<td>24 F</td>
<td>1</td>
<td>10.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN121 Protocol I</td>
<td>19 F</td>
<td>1</td>
<td>2.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN122 Protocol I</td>
<td>24 M</td>
<td>1</td>
<td>2.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN123 Protocol I</td>
<td>23 M</td>
<td>1</td>
<td>1.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN124 Protocol I</td>
<td>27 M</td>
<td>1</td>
<td>14.13</td>
<td></td>
</tr>
<tr>
<td>Data Set D: Healthy Younger Adults; used for investigative purposes</td>
<td>BN125 Protocol II</td>
<td>23 F</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN126 Protocol II</td>
<td>22 F</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN127 Protocol II</td>
<td>21 F</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN128 Protocol II</td>
<td>19 M</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN129 Protocol II</td>
<td>21 F</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN130 Protocol II</td>
<td>24 M</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Data Set E: Healthy Older adult; for classifier testing</td>
<td>BN113 Protocol I</td>
<td>55 F</td>
<td>3</td>
<td>41.56 ± 10.35</td>
<td></td>
</tr>
</tbody>
</table>

1 Subjects had less than twenty 60 second epochs of wake.
2 Subjects sleep records were not scored, however their UMBS data were used to generate independent coefficients.

C. Features

13 UMBS features, taken from the temporal (TMF), spatial (SMF) and statistical (Act) metrics and the UMBS respiration rate estimate, were used in this analysis (as given in Table III). A large number of features were found to be correlated with each other (see Table IV) using dataset D.
TABLE III. UMBS-DERIVED FEATURES.

<table>
<thead>
<tr>
<th>No.</th>
<th>Feature Type</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Respiration</td>
<td>Number of Respiratory Peaks</td>
</tr>
<tr>
<td>2</td>
<td>Spatial Movement</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>3</td>
<td>SMF (Eqn. 6)</td>
<td>Maximum</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Time Greater Than 0</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Num. Distinct Movements</td>
</tr>
<tr>
<td>7</td>
<td>Temporal Movement</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>8</td>
<td>TMF₂ (Eqn. 3)</td>
<td>Maximum</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Median</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Time Greater Than 4</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Num. Distinct Movements</td>
</tr>
<tr>
<td>13</td>
<td>Statistical</td>
<td>Log of Act</td>
</tr>
</tbody>
</table>

D. Classifier

A number of classifiers were investigated for discriminating between sleep and wake using the features derived from the UMBS. In addition to classical Linear and Quadratic Discriminant Analysis (LDA and QDA), the non-parametric k-Nearest Neighbour (kNN) classifier, and the nonlinear Artificial Neural Network (ANN) and Support Vector Machine (SVM) classifiers were tested to see if they provided superior performance to LDA. These are now briefly introduced and details given on how their internal configurations were selected and optimised. The software implementation employed for each classifier was that provided in MATLAB version 7.12.0 (R2011a). Default parameters and training algorithms were used, unless otherwise specified.

a) Discriminant Classifiers: Discriminant classifiers, configured using training data, divide an n-dimensional input (or feature) space into subspaces which optimally discriminate between classes on the assumption that the conditional probability distribution for each class is multivariate normal. Decision boundaries between classes are linear if class covariance matrices are assumed to be the same and quadratic otherwise, yielding linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), respectively [45]. An attractive feature of LDA and QDA is that they have relatively low complexity and are parameter free.

b) k-Nearest Neighbour (kNN) classifier: The k-Nearest Neighbour classifier is a simple, non-linear classification method which generally results in relatively high accuracy. It is a non-parametric, memory-based algorithm which classifies new samples based on the classification of the k most similar samples in a reference training dataset. Here, the Euclidean distance between samples was used as a similarity metric and the majority class of the k nearest neighbours taken as the sample classification. The optimum number of nearest neighbours k was determined by cross-validation from a candidate set k = 1, 5, 9 based on a preliminary investigation that showed that the optimal k value was less than 10. The restricted search was undertaken to reduce the computational overhead.

c) Artificial Neural Network (ANN) Classifier: An ANN is an interconnected network of nodes, or neurons, loosely modelled on the interconnection and operation of neurons in the brain. By employing appropriate learning strategies to adjust network parameters ANNs can learn to approximate arbitrary non-linear input-output relationships between variables. For the classification task considered here, a single hidden layer multilayer perceptron (MLP) neural network with tanh activation functions was used to perform the classification task. Early stopping using cross-validation was employed to determine the optimum number of iterations in order to avoid over-fitting. To allow for the potential of getting stuck in a poor local minimum training was repeated 5 times with different random parameter initializations and the best performing network selected. There are several options for optimising the topology of a neural network, for example varying the number of hidden layer neurons or even the number of hidden layers. Here, a single hidden layer topology was used and the number of hidden layer units was selected as either 20 or 40. More extensive optimization of network topology was not undertaken due to the computational overhead involved, and instead early stopping was employed to ensure good generalization performance. For further information on ANNs the interested reader is referred to Bishop [46].

d) Support Vector Machines (SVM): Support Vector Machines generate boundaries that maximally separate classes. Boundaries can be linear or non-linear. Non-linear boundaries are achieved by employing the so called kernel trick, where the feature space is mapped into a high dimensional space using a kernel function, \( K(x; y) \) [47], [48]. Determining linear boundaries in this high dimensional space corresponds to defining non-linear boundaries in the original feature space. Many functions can be used for the mapping kernel. In this work the radial basis function (RBF) kernel, a popular choice for continuous feature spaces, is selected. Using this configuration SVMs have two hyperparameters, the kernel width, \( \sigma \), and the box constraint, \( C \). The box constraint is a parameter which controls the extent to which misclassifications are allowed to occur on the training data and effectively determines the smoothness of the boundaries between classes. The optimal hyperparameters for each classifier were determined by performing a grid search and selecting the values that yielded the best cross-validation error. Following an initial investigation, and mindful of the computational complexity of SVM training, the search grid range was set as 0.01 to 4 for \( \sigma \) with a grid resolution of 0.5 and 0.01 to 20 for \( C \) with a grid resolution of 1 unit.

E. Performance Measures

The performance of a binary classifier can be quantified in terms of its ability to correctly predict the actual class (positive or negative) for all samples under investigation. When a class is incorrectly predicted it results in either a false negative (FN) or false positive (FP), while correctly classified samples are either true positives (TP) or true negatives (TN). Using these definitions the accuracy of a classifier can be defined as

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\] (7)

However, seeking to maximise this metric can be sub-optimal if the data be largely skewed toward either class. In order to
take account of this bias, the specificity and sensitivity (also referred to as recall) metrics, defined as:

$$\text{Specificity} = \frac{TN}{TN + FP}$$  \hspace{1cm} (8)

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN}$$  \hspace{1cm} (9)

are often reported instead. These metrics quantify the proportion of correctly classified negative and positive samples, respectively. Alternatively, the classifier performance can be expressed in terms of the F-Score,

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$  \hspace{1cm} (10)

which is the harmonic mean of the recall (eqt. 9) and the precision of the classifier, where precision is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$  \hspace{1cm} (11)

The use of the harmonic mean ensures that the F-score is heavily skewed toward the lower value and this addresses the situation where there is a large difference between precision and recall, which can arise with unbalanced class sizes.

F. Classifier Training and Validation Preprocessing Procedure

Datasets A and B were split into subsets of sleep and wake epochs on a per subject basis. 400 epochs of sleep and 400 epochs of wake, randomly chosen from each person, were grouped into a single training/validation dataset. For two of the younger subjects (BN101 and BN108) and one older adult (BN113), random sampling with replacement was used to select the wake epochs as there were less than 400 wake epochs available (95, 198 and 196 wake epochs available respectively as shown in Figure 4). These data were then grouped into three sets: younger adults, older adults and all subjects.

The number of samples available to train the classifiers for the younger, older and all subjects groups are given in Table V. For the SVM, a reduced number of samples was used for the training procedure (as per Table V) due to the significantly increased computational overhead. Due to the low number of data points available, Monte Carlo cross-validation was used to avoid data related bias and ensure consistency of results. The training/validation task, as described above, was repeated 5 times for each classifier with randomly chosen subsets of samples in the training and validation datasets.

G. Independent Classifier Testing Cohort

Each classifier was applied to respective independent data of 5 younger adults (1 night each), an older adult (3 nights) and a combination of these (datasets C and E as per Table II).

For the younger adult cohort, fifty samples of sleep and fifty samples of wake were randomly selected from each of the five subjects in dataset C and compiled into the younger adult testing dataset. Random sampling with replacement was used to select the fifty wake epochs due to the low number of wake epochs available in this dataset. A much lower number of samples was used here compared to the training and validation stage as the number of wake epochs was below 50 for subjects BN117 and BN120 (23 and 25 epochs respectively). The subjects were only awake for 13% of the recordings. Random sampling without replacement was used to select fifty of the sleep epochs in this dataset. The subsets of chosen sleep and wake epochs from all subjects were combined to create a set of 500 chosen epochs.

For the older adult cohort, 200 samples of sleep and 200 samples of wake were randomly selected for each subject.

---

**TABLE IV.** Correlation ($r$) Between the UMBS Features Averaged over the 6 Subjects in Data Set D.

<table>
<thead>
<tr>
<th></th>
<th>Resp. Peaks</th>
<th>SMF</th>
<th>TMT</th>
<th>NCM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Resp. St. Dev.</td>
<td>Max</td>
<td>Mean</td>
<td>NCM St. Dev.</td>
</tr>
<tr>
<td>SMF</td>
<td>-0.145</td>
<td>0.810</td>
<td>1.000</td>
<td>0.094</td>
</tr>
<tr>
<td>Max</td>
<td>-0.159</td>
<td>0.923</td>
<td>1.000</td>
<td>0.854</td>
</tr>
<tr>
<td>Mean</td>
<td>-0.107</td>
<td>0.519</td>
<td>0.942</td>
<td>0.509</td>
</tr>
<tr>
<td>TMT</td>
<td>-0.157</td>
<td>0.412</td>
<td>0.937</td>
<td>0.979</td>
</tr>
<tr>
<td>NCM</td>
<td>-0.093</td>
<td>0.353</td>
<td>0.170</td>
<td>0.137</td>
</tr>
</tbody>
</table>

**TABLE V.** Sample sizes for all classifiers.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Younger No. Samples</th>
<th>Older No. Samples</th>
<th>All No. Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tr</td>
<td>Va</td>
<td>Tr</td>
</tr>
<tr>
<td>SVM</td>
<td>320</td>
<td>6080</td>
<td>280</td>
</tr>
<tr>
<td>Others</td>
<td>4266</td>
<td>2134</td>
<td>3732</td>
</tr>
</tbody>
</table>

No. of Training Samples(Tr), No. of Validation Samples (Va)
over the three nights and compiled into the older adult testing dataset. Random sampling without replacement was used to select the sleep and wake epochs in this dataset as the dataset contained over 41% wake epochs. The subsets of chosen sleep and wake epochs from all sleeping episodes were combined to create a set of 400 chosen epochs.

The data from one sleeping episode from one younger subject (BN119) and one older subject (BN113) were compiled into a testing set for older and younger subjects. The data from the younger subject contained 113 wake epochs and 472 sleep epochs. The data from the older subject contained 264 wake epochs and 233 sleep epochs. 200 samples were randomly selected, without replacement where possible, to create a dataset of 400 sleep epochs and 400 wake epochs. This dataset contained an equal contribution of data from younger and older adults.

V. UMBS Sleep/Wake Classification Results

In this section, results from applying each classifier to the features with the aim of discriminating sleep from wake epochs to training/validation UMBS data from each cohort, using Monte Carlo cross-validation, are given. Finally, the optimally trained classifier is selected and applied to an independent testing dataset to provide impartial results.

For the training/validation procedure, data from 8 younger subjects over 13 scheduled sleeps each and 7 older subjects over 14 scheduled sleeps each were collected. 400 epochs of sleep and 400 epochs of wake were randomly selected from each subject and evenly split to create training and validation data sets for a younger, older and all subjects cohorts. For the testing procedure, data from 5 younger subjects over 1 schedule sleep each and 1 older subject over 3 sleep episodes were used to create an independent test data set of younger cohort and older cohort respectively. 50 epochs of sleep and 50 epochs of wake were selected from each sleep episode from the younger data set. 200 epochs of sleep and 200 epochs of wake was extracted over the 3 sleep episodes from the older subject. An additional all subjects cohort was extracted using data from one sleep episode from one younger and one older subject to ensure the data was not biased to a certain cohort.

The mean performance of the classifiers (F-score) over all repetitions on the validation data is reported in Table VI. Low standard deviations (less than 0.025) were found in the F-score, sensitivity and specificity values.

<table>
<thead>
<tr>
<th>Cohort/Classifier</th>
<th>F-score</th>
<th>Sens.</th>
<th>Spec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Younger Adult</td>
<td>0.758</td>
<td>0.879</td>
<td>0.560</td>
</tr>
<tr>
<td>Older Adult</td>
<td>0.813</td>
<td>0.980</td>
<td>0.570</td>
</tr>
<tr>
<td>All Subjects</td>
<td>0.788</td>
<td>0.955</td>
<td>0.532</td>
</tr>
</tbody>
</table>

The optimal classifier was a neural network, however the performance of the LDA classifier (see Table VII) was only slightly lower (by less than 0.006) and within one standard deviation of the Neural Network. The LDA classifier is characterised by very high sensitivity values and adequate specificity values. Additionally, due to the simpler implementation in low-level hardware of an LDA classifier, it was chosen as the optimal classifier for sleep/wake classification. Low standard deviations (less than 0.02) were found in the F-score, sensitivity and specificity values.

Performance results for the optimal classifiers, as applied to their cohorts independent unseen data is given in Table VIII.

VI. Discussion

A systematic and robust study has been presented demonstrating that a non-contact sleep monitoring system offers similar performance to existing non-clinical/home-based validated sleep monitoring solutions in a manner conducive to long term monitoring, especially amongst vulnerable populations such as those with memory impairment. The system was trained and validated using a large dataset and a varied cohort, especially given the constraints of collecting sleep related data. The underpinning feature extraction was optimized using a cross-validation procedure with an independent unseen testing cohort to ensure validity across new data sets when presented to the system.

Results derived on a per-night basis using an independent test data set are given in Table IX. The performance levels of state-of-the-art commercially available regulated sleep monitors are also reported in the table for comparison purposes. As can be seen, the UMBS system provides comparable results. Wrist actigraphy reports slightly higher specificity levels, however this technology requires subjects to wear a watch for the duration of data collection. The UMBS is particularly suitable for sensitive populations, such as people with dementia, for whom unobtrusive ambient technologies are preferred or where wearable options have low adherence rates due to extended data collection periods. Overall, the UMBS can provide an accurate and reliable estimation of clinically important sleep metrics over extended periods in the real-world.

Within this study, some constraints must be noted. Firstly, the subjects partaking in this research were screened for health
and potential sleep disorders and therefore they may not be representative of a population requiring sleep monitoring technologies. Additional research evaluating this technology on a clinical population should be carried out. In particular, its suitability for heavier and lighter populations and those suffering with insomnia and breathing problems needs to be assessed. It is noted that the system was designed based on changes in pressure distribution rather than absolute weight, and as such performance is expected to be robust to variation in weight. However, in the case of sleep conditions such as sleep apnea and insomnia performance is likely to degrade in line with other movement based sleep monitoring devices due to the abnormal sleep movement profiles associated with these conditions. Secondly, the sleep schedules in these studies are atypical of natural sleep timing patterns. While the extended day length and sleep restriction were artificially enforced, this should not have any bearing on the relationship between UMBS-derived data and PSG-derived sleep/wake state. The range of sleep timing imposed in the studies also produced sleep that varied in quality from night-to-night, and while artificial, resulted in a wide range of sleep qualities that allowed for more robust testing. As such, this data is adequate for the investigation of the classification of wake and sleep stages from UMBS data. Additionally, sleep efficiencies over all nights were within the expected range for healthy younger and older adults (mean of 9.7 % and 19.4% respectively).

Thirdly, the nature of clinical sleep research is artificial by definition and, as such, all data collected in a clinic is non-representative of real-world sleep. For example, participants may have increased wake time due to the PSG recording equipment, and may lie more passively than when in their own homes resulting in an increased durations or formations of quiescent wake. However, this is an inherent constraint for any type of clinical research.

Methods for reducing the dimensionality of the feature space (Principal component analysis [43] and feature subset selection [44]) were investigated, but these did not significantly increase the overall accuracy of the system. As such only results for classification using the full set of input features have been reported.

A variety of linear and non-linear classifiers were investigated for sleep/wake classification. On balance the LDA classifier is recommended as the preferred solution over all cohorts as although its performance is marginally inferior to other classifiers considered (it is within 1% of the best F-score achieved for each cohort), it is the least complex to implement and hence has a much lower computational overhead.

Sleep stage discrimination and wake/REM/NREM discrimination was also investigated, however the system did not perform to an acceptable level. A similar design to that described above with additional classes relating to each sleep stage was used. Similarly to the approach taken in this paper, these data were examined on a per-cohort basis, however using a per-subject design may achieve better performance but would require significant quantities of data. Integrating the timing of the sleep epoch within the sleep episode as a feature may improve the performance of this type of system as increased levels of wake occur at the beginning and end of sleep episodes. This was not considered in the current study, where the focus has been on relating UMBS data directly to sleep/wake state such that is provides a balanced identification of wake episodes occurring throughout a sleep episode.

Often in sleep research, sensitivity and specificity values are reported due to the inherently biased nature of sleep datasets. This paper proposes the use of a harmonic mean performance measure (F-score) which is insensitive to an uneven sample distribution.

Another consideration not generally taken into account in the design and development of home-based sleep/wake monitoring systems is the usability or acceptability of the proposed solution over long durations. Considerations like the perceived benefit of the system, the stigma in the systems use, physical space limitations, the aesthetic design and the visibility of the system to visitors may reduce the practical use/implementation of such a system.

**VII. Conclusion**

In this paper, a non-contact unobtrusive pressure based technology has been shown to offer similar performance to existing non-clinical/home-based validated sleep monitoring solutions and to also capture body movement information during sleep. The UMBS sleep scoring system is suitable for long term placement in domestic homes and is ideal for the non-intrusive collection of sleep data. The UMBS is especially suited to monitoring vulnerable populations where it is preferential to keep modifications to the sleeping environment at a minimum (such as those with mild cognitive impairment or dementia where sleeping difficulties pervade). The performance and reliability of the UMBS technology is comparable to the current ambulatory sleep/wake monitoring gold standard, wrist actigraphy [13], and an alternative non-contact sleep/wake monitor (BiancaMed SleepMinder [2], [22] (as given in Table IX and calculated over the data for each complete sleeping episode).

**Acknowledgment**

The authors wish to thank the study participants, the technical staff of the Division of Sleep and Circadian Disorders, and the technical and nursing staff of the Intensive Physiological Monitoring Unit. We are indebted to the Project Leaders of the studies, Drs. Anne-Marie Chang, Sean W. Cain, and Mirjam Y. Münch. The studies were supported by grants P01 AG09975 (to CAC) and R01 HL080978 (to JFD) from the US National Institutes of Health, and were carried out.
in the Brigham and Women’s Hospital Center for Clinical Investigation, part of the Harvard Catalyst — The Harvard Clinical and Translational Science Center (National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health Award ULI TR001102) and financial contributions from Harvard University and its affiliated academic healthcare centers. L. Walsh was supported by a Fulbright Award for Science and Technology.

REFERENCES


Lorcan Walsh Biography to be entered in final draft.

Sean McLoone Biography to be entered in final draft.

Jeanne Duffy Biography to be entered in final draft.

Joseph Ronda Biography to be entered in final draft.

Charles A. Czeisler Biography to be entered in final draft.
ADDENDUM I

An investigation into UMBS derived movement and non-movement data. Data were collected experimentally to investigate the effect of movement and non-movement on UMBS sensor data. A healthy adult was asked to shift between 4 postures: lying on back (B), lying on front (F), lying on left side (L) and lying on right side (R) under three conditions: 1) rapid transition without lateral movement, 2) rapid transition with lateral movement, and 3) slow rolling movement with lateral movement. UMBS data relating to the various bed entry, bed exits and postural changes are shown in Figure 1A below. The corresponding TMF data is presented in Figure 1B (with the range of value given an upper limit of 5 for this example). It can be determined empirically that all bed entry, bed exit and postural changes can be discriminated from non-movement activity (e.g. micro-movements from breathing) using a TMF threshold of 4.

(a) UMBS data collected while the participant was asked to assume and shift between four typical sleeping postures: lying on their back (B); left side (L); right side (R); and front (F). The subject was asked to change postures under three conditions 1) rapid transition without lateral displacement, 2) rapid transition with lateral displacement and 3) slow rolling transitions (inherently including lateral displacement).

(b) Corresponding temporal movement features with movement threshold (time domain).

Figure 5. UMBS derived temporal feature with movement threshold of 4 from a participant shifting between postures on the UMBS.