Belun Ring (Belun Sleep System BLS-100): Deep learning-facilitated wearable enables obstructive sleep apnea detection, apnea severity categorization, and sleep stage classification in patients suspected of obstructive sleep apnea

Zachary Strumpf, MD a, Wenbo Gu, MPhil b,c, Chih-Wei Tsai, PhD b, Pai-Lien Chen, PhD d, Eric Yeh, MD a,e, Lydia Leung, PhD b, Cynthia Cheung, PhD b, I-Chen Wu, PhD c, Kingman P. Strohl, MD a,c,f, Tiffany Tsai g, Rodney J. Folz, MD, PhD h, Ambrose A. Chiang, MD a,e,f,⁎

a Division of Pulmonary, Critical Care, and Sleep Medicine, University Hospitals Cleveland Medical Center, Cleveland, OH, USA
b Belun Technology Company Limited, Hong Kong
c Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
d FHI360, Durham, NC, USA
e Department of Medicine, Case Western Reserve University, Cleveland, OH, USA
f Division of Sleep Medicine, Louis Stokes Cleveland VA Medical Center, Cleveland, OH, USA
g Division of Pulmonary, Critical Care, and Sleep Medicine, Houston Methodist Hospital, Houston, TX, USA

Article Info

Article history:
Received 8 November 2022
Received in revised form 25 March 2023
Accepted 3 May 2023

Keywords:
Obstructive sleep apnea
Photoplethysmography
Peripheral arterial tonometry
Home sleep apnea testing
Sleep technology
Artificial intelligence
Apnea-hypopnea index
Digital health
Validation

Abstract

Goal and aims: Our objective was to evaluate the performance of Belun Ring with second-generation deep learning algorithms in obstructive sleep apnea (OSA) detection, OSA severity categorization, and sleep stage classification.

Focus technology: Belun Ring with second-generation deep learning algorithms

Reference technology: In-lab polysomnography (PSG)

Sample: Eighty-four subjects (M: F = 1:1) referred for an overnight sleep study were eligible. Of these, 26% had PSG-AHI < 5; 24% had PSG-AHI 5–15; 23% had PSG-AHI 15–30; 27% had PSG-AHI ≥ 30.

Design: Rigorous performance evaluation by comparing Belun Ring to concurrent in-lab PSG using the 4% rule.

Core analytics: Pearson’s correlation coefficient, Student’s paired t-test, diagnostic accuracy, specificity, positive predictive value, negative predictive value, positive likelihood ratio, negative likelihood ratio, Cohen’s kappa coefficient (kappa), Bland-Altman plots with bias and limits of agreement, receiver operating characteristics curves with area under the curve, and confusion matrix.

Core outcomes: The accuracy, sensitivity, specificity, and kappa in categorizing AHI ≥ 5 were 0.85, 0.92, 0.64, and 0.58, respectively. The accuracy, sensitivity, specificity, and Kappa in categorizing AHI ≥ 15 were 0.89, 0.91, 0.88, and 0.79, respectively. The accuracy, sensitivity, specificity, and Kappa in categorizing AHI ≥ 30 were 0.91, 0.83, 0.93, and 0.76, respectively. BSP2 also achieved an accuracy of 0.88 in detecting wake, 0.82 in detecting NREM, and 0.90 in detecting REM sleep.

Core conclusion: Belun Ring with second-generation algorithms detected OSA with good accuracy and demonstrated a moderate-to-substantial agreement in categorizing OSA severity and classifying sleep stages.

Published by Elsevier Inc. on behalf of National Sleep Foundation. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

⁎ Corresponding author: Ambrose A. Chiang, MD, Division of Sleep Medicine, Louis Stokes Cleveland VA Medical Center, 10701 East Blvd, Suite 2B-129, Cleveland, OH 44106, USA.
E-mail address: Ambrose.chiang@va.gov (A.A. Chiang).

https://doi.org/10.1016/j.sleh.2023.05.001
2352-7218/Published by Elsevier Inc. on behalf of National Sleep Foundation. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Introduction

Untreated sleep-disordered breathing remains a serious concern at individual and global sleep health levels as it is associated with significant cardiovascular and metabolic comorbidities as well as mortality. In 2016, Frost & Sullivan calculated that the annual economic burden of undiagnosed sleep apnea among U.S. adults was approximately $149.6 billion when indirect costs were included. 

Benjafield recently estimated that one-seventh of the world’s adult population, or approximately one billion people, may have sleep apnea. In light of the high economic burden and continually increasing prevalence due to aging and the obesity epidemic, there is a genuine need for efficacious, pragmatic, and inexpensive methods to meet the growing demands for obstructive sleep apnea (OSA) diagnosis.

Historically, the diagnosis of sleep-disordered breathing relies on either polysomnography (PSG) or home sleep apnea testing (HSAT). However, they are not without pitfalls. PSG, the gold standard, often suffers low availability, high cost, and long wait time and is less likely to foster awareness of sleep health from a public education standpoint. On the other hand, HSAT, though less cumbersome and more readily available, continues to pose interpretation challenges due to false-negative results and the lack of neurophysiological signals. In recent years, there has been an extraordinary proliferation of innovative sleep technologies. These rapidly evolving technologies allow sleep physiology data acquisition outside sleep labs, permit longitudinal follow-ups, and empower population-based research.

Currently, two clinically available photoplethysmography (PPG)-based devices and one software as a medical device (SaMD) are cleared by the US Food and Drug Administration (FDA) for OSA diagnosis. WatchPAT (ZOLL-Itamar, Atlanta, GA) and NightOwl (ResMed, San Diego, CA) use signal conditioning methods to derive peripheral arterial tonometry (PAT) signals from PPG. They differ in their mechanisms in deriving PAT signals with WatchPAT using a pneumo-optic probe (hardware) and NightOwl using a software algorithm. SleepImage (SleepImage, Denver, CO), a SaMD, uses an electrocardiogram (ECG) or PPG to calculate cardiopulmonary coupling to derive the apnea-hypopnea index (AHI).

Belun Ring (formally, Belun Sleep System BL5-100, Belun Technology Company Limited, Hong Kong) is a deep learning (DL)-powered wearable system recently cleared by the US FDA as a class II home sleep testing device for the diagnosis of moderate to severe OSA and sleep stage classification (K222579). The key hardware is the Belun Ring sensor (Fig. 1), an FDA Class II reflectance pulse oximeter (K211407). Its first-generation artificial intelligence algorithm (BSP1), has demonstrated good overall accuracy in identifying moderate to severe OSA but had poor specificity at the AHI cutoff of 5 events/hour. BSP1 also could not classify sleep stages and tended to overestimate AHI when Belun-AHI (bAHI) was under 15 events/hour. To remedy these deficiencies, the BSP1 underwent several iterations of re-training and fine-tuning. The Belun Sleep Platform Generation 2 (BSP2), consisting of 2 proprietary neural network models (one for sleep detection and another for respiratory event detection), can automatically learn features and patterns from various data outputs to estimate the sleep/wake epochs, sleep stages, and AHI.

We hypothesized that BSP2 is superior to BSP1 in diagnosing OSA at the low AHI range and will allow accurate sleep stage classification. We rigorously assessed the performance of BSP2 in OSA detection, sleep apnea severity diagnostic concordance, sleep-wake estimation, and 3-stage model (Wake, REM, and non-rapid eye movement (NREM)) classification by directly comparing BSP2 to concurrent in-lab PSG in adult subjects suspected of OSA.

Methods

Sample

To be eligible to participate in this study, subjects must meet all of the following criteria: (1) sign and date informed consent; (2) age 18-80; and (3) clinically assessed and suspicious for OSA with a STOP-Bang score ≥3. Individuals who met any of the following criteria were excluded from participation in this study: undergoing a full-night positive airway pressure (PAP) titration study; on pacemaker/defibrillator, left ventricular assist device, home O2, noninvasive ventilator, diaphragmatic pacing, or any form of nerve stimulator; having atrial fibrillation-flutter, left ventricular ejection fraction (LVEF) < 55%, or status post-cardiac transplantation; taking narcotics; recent hospitalization or recent surgery in the past 30 days; unstable cardiopulmonary status on the night of the study judged to be unsafe for study by the sleep tech or the on-call sleep physician. Individuals who met the eligibility criteria but failed to have at least 4 hours of technically valid sleep based on BSP in a diagnostic study or at least 3 hours of technically valid sleep during the diagnostic portion of a split-night study were excluded from statistical analyses.

Based on our prior BSP trials, the mean and standard deviation of differences in AHI between BSP1 and PSG were set at 3 events/hour and 9 events/hour. We estimated that a sample size of 79 should be sufficient. The significance level and power were set at 95% and 0.8, respectively. The mean and standard deviation of the differences in AHI between BSP1 and PSG were assumed to be 3 events/hour and 4 events/hour, respectively.

Fig. 1. Belun Ring sitting on the charging cradle (A); with 7 adjustable arms for different finger sizes (B); and worn on the proximal phalanx of the index finger (C).
adequately provide 85% power to detect a maximum allowed difference of 26 events/hour with a significant level of 0.05. In the end, we successfully recruited 84 eligible subjects with a male: female ratio of 1:1 for final analysis. The study protocol was approved by the University Hospitals Institutional Review Board (STUDY20201259) and registered as ClinicalTrials.org (NCT04885062).

**Focus method/technology**

| Focus method/technology                                                                                       |
|                                                                                                                |

Table 1 shows the specifications of BSP1 and BSP2. Fig. 2 illustrates the architecture schematic representation of the BSP2 respiratory event detection model developed based on the ResNeXt architecture, and squeeze-and-excitation network, an architecture to improve the representational power of the neural network. BSP2 was trained using the domain-adversarial method to estimate AHI by assessing SpO2 and pulse rate inputs. This model was rebuilt using 2 independent datasets, including the sizable Sleep Heart Health Study (SHHS) open-source dataset containing 5,804 subjects and 8,434 studies for training, and 128 studies from the BSP Colorado study (N = 50) and the prior BSP University Hospitals Cleveland Medical Center study (N = 78) for tuning (Table 1).

The BSP2 sleep stage model was built to learn features and patterns from SpO2, pulse rate, and accelerometer signals to classify the sleep stage in 30-second epochs. The structure of this sleep stage model consists of two parts: (1) signals of every 30-second segment passed a 5-layer convolutional neural network to extract the primary features automatically, and (2) the primary features from consecutive segments passed a Transformer structure, an attention-based recurrent neural network architecture, to obtain the results of classification. This sleep stage model was also trained using the SHHS dataset and tuned on the independent BSP Colorado and BSP Cleveland Medical Center datasets using 5-fold validation.

**Reference method/technology**

Participants willing to consent underwent nocturnal in-lab sleep studies using the SleepWorks sleep software (Natus, Pleasanton, CA). The PSG montage includes electroencephalogram (EEG) leads (O1M2, O2M1, C1M2, C2M1, F1M2, F2M1), right electrooculogram, left electrooculogram, chin electromyogram, nasal pressure airflow, thoracic airflow, chest and abdominal respiratory efforts, pulse oximetry, left leg electromyogram, right leg electromyogram, and ECG. Continuous positive airway pressure (CPAP) flow was monitored for split-night studies. All studies were manually scored in 30-second epochs according to the American Academy of Sleep Medicine (AASM) Scoring Manual, version 2.4 by a seasoned scoring technician and reviewed by one of our investigating sleep physicians. An obstructive apnea event is defined as a decrease in the thermistor airflow to < 10% of the baseline for ≥10 seconds with continued respiratory effort. A hypopnea event is defined as a decrease in nasal pressure signal excursions by 30%-90% of the baseline for ≥10 seconds.

*Fig. 2. Architecture schematic representation of BSP2 respiratory event detection model. BSP2, second-generation Belun Sleep Platform artificial intelligence algorithms; SE, squeeze-and-excitation network; 1D, one-dimensional; PR, pulse rate.*

**Table 1**

BSP1 versus BSP2 CNN Model comparison

<table>
<thead>
<tr>
<th></th>
<th>BSP1</th>
<th>BSP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI model structure</td>
<td>3 fully-connected layers (160, 80, 20) neurons for respiratory event detection (160, 80, 5) neurons for sleep time estimation</td>
<td>ResNeXt with SE modules for respiratory event detection Transformer with CNN for sleep stage classification</td>
</tr>
<tr>
<td>Inputs</td>
<td>74 features extracted from: SpO2 Pulse Rate Pulse Rate Variability PPG Accelerometry signal</td>
<td>Automatic feature extraction from: SpO2 Pulse Rate Accelerometry signal</td>
</tr>
<tr>
<td>Outputs</td>
<td>bAHI &amp; bTST every 5 min</td>
<td>bAHI every 5 min bTST, bWake, Belun REM sleep (bREM), and Belun non-REM sleep (bNREM) every 30 s</td>
</tr>
<tr>
<td>Training dataset</td>
<td>SHHS using 8,417 PSGs</td>
<td>Gu et al. (2020) 50 subjects Yeh et al. (2021) 78 subjects</td>
</tr>
<tr>
<td>Tuning dataset</td>
<td>No tuning/internal validation</td>
<td>SHHS using 8,434 PSGs</td>
</tr>
<tr>
<td>External validation dataset</td>
<td>Gu et al. (2020) 50 subjects Yeh et al. (2021) 78 subjects</td>
<td>SHHS, Sleep Heart Health Study</td>
</tr>
</tbody>
</table>

BSP1, first-generation Belun Sleep Platform artificial intelligence algorithms; BSP2, second-generation Belun Sleep Platform artificial intelligence algorithms; bAHI, Belun apnea-hypopnea index; bTST, Belun total sleep time; bWake, Belun wake; bNREM, Belun NREM; bREM, Belun REM; ResNeXt, a specific CNN model architecture; SE, squeeze-and-excitation network; SHHS, Sleep Heart Health Study.
accompanied by oxygen desaturation ≥4%. The overall AHI was calculated from the summation of AHI events divided by the total sleep time (TST). The Belun Technology Company investigators were blinded to the in-lab PSG results.

**Design, study setting, and procedures**

Fig. 3 shows the CONSORT flow diagram of this study. Five hundred fifty-five consecutive individuals (N = 555) referred to the University Hospitals Cleveland Medical Center Bolwell and Beachwood sleep labs were prescreened for eligibility. Of these, 104 did not meet the eligibility criteria. Seventy-one subjects refused to participate in the study (N = 71), and two hundred forty-four individuals were either no-shows for the study or refused to consent (N = 244). In the end, 136 subjects met the eligibility criteria and underwent BSP2 testing with a concurrent overnight in-lab sleep study (either diagnostic or split-night study). Thirty of the 136 individuals who failed to have adequate BSP2 recording time were excluded. As the protocol demands a 1:1 male: female ratio, only the first 42 males and 42 females were included in the statistical analysis.

A sleep technician measures the proximal index finger of the nondominant hand using a paper ring selector. Notably, we streamlined the BSP2 testing procedure in this study by discarding the second step of the 2-step ring selection approach utilized in the two prior BSP1 studies, as our experience showed that the PPG signal quality is typically adequate, and checking the perfusion index for signal quality is not an indispensable step. After the appropriate ring size is determined, the ring is placed and left overnight. When the study concludes in the morning, the sleep technician removes the ring and places it back on the cradle connected to a laptop. The acquired data are uploaded to the Amazon cloud where the BSP2 AI algorithms automatically compute and generate a report that can be downloaded. The PSG scoring technicians and the interpreting physicians were blinded to the BSP results per protocol.

All statistical analyses were carried out in R 4.1.0. One-way ANOVA was used to compare the mean of age and body mass index (BMI) among OSA severity groups (i.e., no OSA, mild, moderate, and severe OSA).
severe OSA). Kruskal-Wallis test by ranks was used to compare median AHI among OSA severity groups. Mantel-Haenszel Chi-Squared test to compare frequencies in sex and race among OSA severity groups.

The correlations between bAHI and PSG-AHI, as well as between bTST and PSG-TST, were determined by Pearson’s correlation coefficient. Bland-Altman plots for AHI, TST, wake, NREM, and REM were constructed to illustrate the differences between BSP2 and PSG of all individuals. Receiver operating characteristics (ROC) curves with estimated area under the curve (AUC) were displayed with the true positive rate (y-axis) against the false positive rate (x-axis) for AHI cutoffs of 5, 15, and 30 events/hour. BSP2 measures were statistically compared with PSG data using the Student’s paired t-test. Performance metrics, including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+), negative likelihood ratio (LR−), and Cohen’s Kappa coefficient (Kappa), were computed at AHI cutoffs of 5, 15, and 30 events/hour.

The 30-second epoch-by-epoch (EBE) TST and 3-stage analysis (wake, NREM, and REM) were compared between BSP2 and PSG. Performance metrics (accuracy, sensitivity, specificity, PPV, NPV, LR+, LR−, and Kappa) were computed for each stage versus the combination of two other classifications. We assessed the BSP2 performance in classifying wake against the combination of NREM and REM (i.e., sleep). The BSP2 performance in classifying NREM and REM sleep was assessed against the combination of the other two stages (i.e., NREM vs. wake and REM, or REM vs. wake and NREM, respectively). As the EBE sleep stage confusion matrix provides crucial information regarding diagnostic concordance, a 3 × 3 confusion matrix showing the classification agreement in the wake (proportion of PSG wake epochs identified correctly as wake by BSP2), NREM (proportion of PSG NREM epochs identified correctly as NREM by BSP2), and REM (proportion of PSG REM epochs identified correctly as REM by BSP2) was also presented.

### Results

#### Core analytics and main outcome variables

**Baseline information**

Table 2 summarizes the baseline data of 84 subjects. The male:female ratio is 1:1 per protocol, with the mean standard deviation (STD) age 48.3 (14.4). African Americans comprise the majority population in this cohort, occupying 62%, and the rest are 29% Caucasian, 2% Asian, 1% Hispanic, 1% others, and 4% without ethnicity details. The distribution of OSA severities in the population is relatively even, 26%, 24%, 23%, and 27% for subjects with no OSA (AHI < 5 events/hour), mild OSA (AHI 5 to < 15 events/hour), moderate OSA (AHI 15 to < 30 events/hour), and severe OSA (AHI ≥ 30 events/hour), respectively. The mean (STD) BMI is 38.7 (9.3). This study has a prevalence of 74%, 50%, and 27% for AHI ≥ 5, 15, and 30 events/hour, respectively.

**BSP2 performance: OSA detection and diagnostic concordance**

The bAHI values derived from BSP2 were compared to PSG-AHI. Table 3 summarizes the performance metrics, including accuracy, sensitivity, specificity, PPV, NPV, LR+, LR−, and Kappa at AHI cutoffs of 5, 15, and 30 events/hour. There was a moderate to substantial agreement between bAHI and PSG-AHI for these OSA severity thresholds (Kappa = 0.58-0.79; accuracy = 0.85-0.91). The overall OSA severity diagnostic concordance rate was 0.67 in BSP2, which significantly improved from 0.36 in PSG! with an increase of specificity at the AHI cutoff of 5 (Table S1, Fig. S1). We also analyzed all subjects with bTST ≥ 1 hour (N = 124), which revealed consistent performance with no statistically significant differences in performance metrics.

The BSP2 bAHI was significantly correlated with PSG-AHI (r = 0.90, P < 0.001, Fig. 4A). The average difference (bias) between bAHI and PSG-AHI was −1.0 events with 1.96 standard deviations of

### Table 2

Summary of subject characteristics and PSG results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>All</th>
<th>Apnea Severity</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No OSA</td>
<td>Mild OSA</td>
<td>Moderate OSA</td>
</tr>
<tr>
<td>Subject (%)</td>
<td>84 (100%)</td>
<td>22 (26%)</td>
<td>20 (24%)</td>
</tr>
<tr>
<td>Sex (%)</td>
<td>Male</td>
<td>42 (50%)</td>
<td>7 (8%)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>42 (50%)</td>
<td>15 (18%)</td>
</tr>
<tr>
<td>Age (y)</td>
<td>48.3 ± 14.4</td>
<td>40.9 ± 8.3</td>
<td>51.9 ± 16.4</td>
</tr>
<tr>
<td>Race (%)</td>
<td>African American</td>
<td>52 (62%)</td>
<td>11 (13%)</td>
</tr>
<tr>
<td></td>
<td>Caucasian</td>
<td>24 (29%)</td>
<td>6 (7%)</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>1 (1%)</td>
<td>1 (1%)</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>2 (2%)</td>
<td>1 (1%)</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>1 (1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>4 (4%)</td>
<td>3 (3%)</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>38.7 ± 9.3</td>
<td>37.3 ± 10.8</td>
<td>35.5 ± 8.1</td>
</tr>
<tr>
<td>AHI (events/h)</td>
<td>15.8 (6.4–30.2)</td>
<td>3.7 (3.1–6.9)</td>
<td>8.3 (5.0–12.3)</td>
</tr>
</tbody>
</table>

AHI, apnea-hypopnea index; BMI, body mass index; OSA, obstructive sleep apnea; PSG, polysomnography. No OSA, AHI < 5; Mild OSA, AHI 5 to < 15; Moderate OSA, AHI 15 to < 30; Severe OSA, AHI ≥ 30.

Median (IQR) for age and BMI.

* Mantel-Haenszel Chi-Squared test compares frequencies in sex and race among No obstructive sleep apnea (OSA), Mild OSA, Moderate OSA, and Severe OSA groups.

** One-way ANOVA compares means of age and body mass index among No OSA, Mild OSA, Moderate OSA, and Severe OSA groups.

Kruskal-Wallis test by ranks compares medians of apnea-hypopnea index among No OSA, Mild OSA, Moderate OSA, and Severe OSA groups.
Table 3  
Performance (mean with 95% CI) of bAHI at PSG-AHI cutoffs of 5, 15, and 30

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>LR+</th>
<th>LR-</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.85</td>
<td>(0.75-0.92)</td>
<td>0.92</td>
<td>0.89</td>
<td>(0.77-0.95)</td>
<td>0.74</td>
<td>2.53</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.82-0.97)</td>
<td>0.64</td>
<td>0.88</td>
<td>(0.77-0.95)</td>
<td>0.74</td>
<td>2.53</td>
<td>0.13</td>
</tr>
<tr>
<td>15</td>
<td>0.89</td>
<td>(0.81-0.95)</td>
<td>0.91</td>
<td>0.88</td>
<td>(0.77-0.95)</td>
<td>0.74</td>
<td>2.53</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.77-0.97)</td>
<td>0.88</td>
<td>0.88</td>
<td>(0.77-0.95)</td>
<td>0.74</td>
<td>2.53</td>
<td>0.13</td>
</tr>
<tr>
<td>30</td>
<td>0.91</td>
<td>(0.82-0.96)</td>
<td>0.91</td>
<td>0.88</td>
<td>(0.77-0.95)</td>
<td>0.74</td>
<td>2.53</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.61-0.95)</td>
<td>0.83</td>
<td>0.88</td>
<td>(0.77-0.95)</td>
<td>0.74</td>
<td>2.53</td>
<td>0.13</td>
</tr>
</tbody>
</table>

bAHI, Belun apnea-hypopnea index; PSG-AHI, polysomnography apnea-hypopnea index; PPV, positive predictive value; NPV, negative predictive value; LR+, positive likelihood ratio; LR-, negative likelihood ratio; Kappa, Cohen’s Kappa coefficient.

Fig. 4. BSP2 OSA diagnostic performance. (A) Scatterplot comparing bAHI to PSG-AHI; (B) Bland-Altman plot for bAHI vs. PSG-AHI; (C) Receiver operating characteristics curve (ROC) with the area under each curve (AUC) with 95% CI in parenthesis for AHI cutoffs of 5, 15, and 30 events/hour; and (D) Confusion matrix of the OSA severity (percentage per PSG label with the number of participants in parenthesis). For (A) and (B), blue solid lines represent the best fit of the data with a 95% confidence interval. For (B), blue dashed lines indicate upper and lower limits of agreement (1.96 standard deviations) and bias (the average difference between BSP2 and PSG). bAHI, Belun apnea-hypopnea index; PSG-AHI, polysomnography apnea-hypopnea index.
17.7 events. The difference between bAHI and PSG-AHI was not statistically significant (t = 1.05, P = 0.30) (Fig. 4B). The ROC curve with the AUC plotted at AHI cutoffs of 5, 15, and 30 events/hour is shown in Fig. 4C. The confusion matrix of the OSA severity (percentage per PSG label with the number of participants in parenthesis) is shown in Fig. 4D. Only 1 of the 22 cases without OSA was classified as moderate OSA, and none of the 22 fell in the severe OSA category. No moderate or severe OSA cases were misclassified as no OSA (Fig. 4D).

BSP2 performance: sleep-wake outcomes and sleep stage classification

The BSP2 bTST correlated well with PSG-TST (r = 0.83, P < 0.001, Fig. 5A). The average difference between bTST and PSG-TST was 25.2 minutes with 1.96 standard deviations of 87.4 minutes (Fig. 5B).

For detecting stage wake, the BSP2 algorithm had a sensitivity of 0.58 and a specificity of 0.96 (Table 4). BSP2 achieved an accuracy of 0.88 in detecting wake, 0.82 in detecting NREM, and 0.90 in detecting REM with correspondingly adequate levels of Kappa (Kappa 0.60 for detecting wake; Kappa 0.60 for NREM; Kappa 0.63 for REM) (Table 4). Fig. 6 shows the Bland-Altman plots for stages wake, NREM, and REM. BSP2 overestimated NREM by an average (1.96 SD) of 23.4 (−76.1 to 122.9) minutes and underestimated wake by an average of 25.2 (−112.6 to 62.3) minutes. For REM, there were no significant differences between BSP2 and PSG with a bias of 1.8 (−70.6 to 74.1) minutes (t = 0.44, P = 0.66).

Table 4

<table>
<thead>
<tr>
<th>Number of Epochs</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>LR+</th>
<th>LR-</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wake</td>
<td>67,107</td>
<td>0.88</td>
<td>0.58</td>
<td>0.96</td>
<td>0.81</td>
<td>0.89</td>
<td>14.72</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>(0.87-0.88)</td>
<td>(0.57-0.59)</td>
<td>(0.96-0.96)</td>
<td>(0.80-0.81)</td>
<td>(0.89-0.89)</td>
<td>(14.08-15.39)</td>
<td>(0.43-0.45)</td>
<td>(0.59-0.61)</td>
</tr>
<tr>
<td>NREM</td>
<td>0.82</td>
<td>0.90</td>
<td>0.68</td>
<td>0.82</td>
<td>0.80</td>
<td>2.80</td>
<td>0.15</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>(0.81-0.82)</td>
<td>(0.90-0.90)</td>
<td>(0.67-0.68)</td>
<td>(0.82-0.83)</td>
<td>(0.80-0.81)</td>
<td>(2.75-2.85)</td>
<td>(0.14-0.15)</td>
<td>(0.59-0.60)</td>
</tr>
<tr>
<td>REM</td>
<td>0.90</td>
<td>0.70</td>
<td>0.94</td>
<td>0.68</td>
<td>0.94</td>
<td>11.58</td>
<td>0.32</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>(0.90-0.90)</td>
<td>(0.69-0.71)</td>
<td>(0.94-0.94)</td>
<td>(0.67-0.69)</td>
<td>(0.94-0.95)</td>
<td>(11.18-11.99)</td>
<td>(0.31-0.33)</td>
<td>(0.62-0.64)</td>
</tr>
</tbody>
</table>

Table 4: Statistics performance (mean with 95% CI) of 3-stage classification of sleep (wake, NREM, and REM)

PPV, positive predictive value; NPV, negative predictive value; LR+, positive likelihood ratio; LR-, negative likelihood ratio; Kappa, Cohen’s Kappa coefficient.
In a study by Ioachimescu et al., concurrent PSG and WatchPAT data were examined in 500 veterans. The diagnostic concordance was 0.42, 0.41, and 0.83 for mild, moderate, and severe OSA, respectively. Iftkhar et al. recently performed a meta-analysis on 6 validation studies using WatchPAT, which showed a Kappa of 0.45, 0.29, 0.25, and 0.64 for classifying patients with no OSA, mild, moderate, and severe OSA severity, respectively. Al Ashry et al. recently investigated the performance of SleepImage SaMD using ECG and pulse oximetry tracings of PSGs from APPLES (Apnea Positive Pressure Long-term Efficacy Study). ROC curves demonstrated good agreement in all OSA severity categories. It is worth noting that the comparison with PSG in this study was “internal” between PSG-AHI and the SleepImage AHI derived from ECG and pulse oximetry tracings of the same PSGs. Notably, no rigorous performance evaluation study using the SleepImage ring has been published. Moreover, two NightOwl studies showed good overall diagnostic concordance of 0.74 and 0.78 using the 4% rule. Table 5 summarizes the key performance metrics of BSP2, WatchPAT-200, NightOwl, and SleepImage SaMD.

**Sleep assessment: comparison with other PPG-based devices/models**

BSP2 has a sensitivity of 0.96 for detecting sleep and a specificity of 0.58 for detecting wake, which on the surface, is comparable to the sensitivities and specificities of consumer wearable sleep trackers such as Fitbit Alta HR (0.95, 0.54), Fitbit Charge 2 (0.96, 0.61), Apple Watch (0.98, 0.60), or Oura Ring (0.96, 0.41). However, it should be pointed out that the BSP2 sleep-wake outcomes in this study were acquired in subjects suspected of OSA, not...
in healthy individuals. On average, BSP2 overestimated TST by 25.2 minutes in this population.

Few studies have assessed the accuracy of PPG-based HSATs or wearables on the sleep-wake outcome and sleep stage classification in adults suspicious of OSA. Moreno-Pino et al. evaluated the performance of Fitbit devices (Charge 2 and Alta HR) against PSG in 65 participants suspected of OSA. Fitbit overestimated TST by 59.8 minutes with a specificity of 0.44 in detecting wake. In terms of medical-grade devices, two earlier studies that assessed WatchPAT demonstrated comparable outcomes to our results. Notably, SleepImage SaMD does not derive conventional sleep stages, which makes the comparison to in-lab PSG difficult.

There has been no published data on the accuracy of sleep stage classification on NightOwl. Recently, Korkalainen et al. utilized PPG signals directly from the diagnostic PSGs of 894 suspected OSA patients to develop a combined deep learning (DL) neural network. The 3-stage model achieved an overall EBE agreement of 0.80. This DL model for 3-stage classification achieved 0.72, 0.87, and 0.70 agreement in detecting wake, NREM, and REM sleep. In our study, the B2P2 EBE classification agreement for the wake, NREM, and REM stages was 0.58, 0.90, and 0.70, respectively. The overall B2P2 EBE classification agreement rate was 0.80, comparable to the Korkalainen study in subjects suspected of OSA.

Additional results and implications

Compared to B1, the overall B2P2 OSA severity concordance significantly improved due to an increase in specificity at the AHI cutoff of 5 (Table S1, Fig. S1). This implies that the robustness and generalizability of DL-enabled B2P2 algorithms can be

### Table 4
Comparison of the OSA diagnosis performance metrics among B2P2, WatchPAT-200, NightOwl, and SleepImage SaMD

<table>
<thead>
<tr>
<th>Device</th>
<th>B2P2</th>
<th>WatchPAT-200</th>
<th>NightOwl</th>
<th>SleepImage SaMD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Device</strong></td>
<td>N = 84</td>
<td>N = 500</td>
<td>N = 167</td>
<td>N = 833</td>
</tr>
<tr>
<td><strong>Signal</strong></td>
<td>PPG</td>
<td>PAT</td>
<td>PAT</td>
<td>ECG and pulse oximetry from the same PSGs (not with an independent sensor)</td>
</tr>
<tr>
<td><strong>Algorithm</strong></td>
<td>DL-enabled</td>
<td>Rule-based</td>
<td>Rule-based</td>
<td>Rule-based</td>
</tr>
<tr>
<td><strong>OSA severity distribution</strong></td>
<td>No OSA = 26%</td>
<td>No OSA = 15%</td>
<td>No OSA = 13%</td>
<td>No OSA = 4%</td>
</tr>
<tr>
<td><strong>Moderate</strong></td>
<td>23%</td>
<td>27%</td>
<td>27%</td>
<td>25%</td>
</tr>
<tr>
<td><strong>Severe</strong></td>
<td>27%</td>
<td>31%</td>
<td>45%</td>
<td>51%</td>
</tr>
<tr>
<td><strong>PSG Scoring Method and criteria</strong></td>
<td>Manual scoring using 4% desaturation hypopnea criteria</td>
<td>Manual scoring using 3% and 4% desaturation hypopnea criteria</td>
<td>Both 3% and 4% rules. Computer-aided sleep scoring service (Cerebral Medical, CM, Canada), including a proprietary automated Michele Sleep Scoring System followed by manual scoring</td>
<td>Manual scoring with hypopnea was defined as &gt; 50% airflow reduction alone or less airflow reduction associated with &gt; 3 desaturation or arousal</td>
</tr>
<tr>
<td><strong>OSA Detection Performance Metrics</strong>: means (95% confidence interval provided when data are available)</td>
<td>ACC = 0.85 (0.75-0.92)</td>
<td>ACC = 0.87</td>
<td>ACC = 0.91 (0.87-0.95)</td>
<td>ACC = N/A</td>
</tr>
<tr>
<td><strong>AHI ≥ 5 events/h</strong></td>
<td>Sen = 0.92 (0.82-0.97)</td>
<td>Sen = 0.97</td>
<td>Sen = 0.95 (0.91-0.99)</td>
<td>Sen = N/A</td>
</tr>
<tr>
<td><strong>AHI ≥ 15 events/h</strong></td>
<td>Spec = 0.64 (0.41-0.83)</td>
<td>Spec = 0.35</td>
<td>Spec = 0.80 (0.68-0.92)</td>
<td>Spec = N/A</td>
</tr>
<tr>
<td><strong>AHI ≥ 30 events/h</strong></td>
<td>PPV = 0.88 (0.77-0.95)</td>
<td>PPV = 0.89</td>
<td>PPV = 0.93 (0.88-0.97)</td>
<td>PPV = N/A</td>
</tr>
<tr>
<td><strong>Sleep Stage Classification Statistics (95% confidence interval provided when data are available)</strong></td>
<td>ACC = 0.89 (0.81-0.95)</td>
<td>ACC = 0.78</td>
<td>Acc = 0.93 (0.89-0.97)</td>
<td>Acc = N/A</td>
</tr>
<tr>
<td><strong>Wake</strong></td>
<td>ACC = 0.91 (0.82-0.96)</td>
<td>ACC = 0.80</td>
<td>Acc = 0.93 (0.89-0.97)</td>
<td>Acc = N/A</td>
</tr>
<tr>
<td><strong>NREM</strong></td>
<td>ACC = 0.83 (0.61-0.95)</td>
<td>ACC = 0.83</td>
<td>Acc = 0.86 (0.77-0.95)</td>
<td>Acc = N/A</td>
</tr>
<tr>
<td><strong>REM</strong></td>
<td>ACC = 0.93 (0.84-0.98)</td>
<td>ACC = 0.79</td>
<td>Acc = 0.94 (0.90-0.99)</td>
<td>Acc = N/A</td>
</tr>
</tbody>
</table>

**BSP2, Belun Sleep Platform with the second-generation AI algorithm; ACC, accuracy; Sen, sensitivity; Spec, specificity; PPV, positive predictive value; NPV, negative predictive value; Kappa, Cohen’s Kappa coefficient; CPC, cardiopulmonary coupling; N/A, not available; PAT, peripheral arterial tonometry.**

*Calculated from Table 2 of Hedner et al.* Please note that the original table is a 4-stage confusion matrix (wake, light sleep, deep sleep, and REM), and we computed the metrics based on 3-stage by combining light and deep sleep as NREM.
improved in a relatively short period (in this case, within a year) by adding unseen datasets for retraining and fine-tuning the model.

Limitations and future perspectives

A few strengths of BSP2 are worth noting. The first strength is using the SHHS dataset for training and building the BSP algorithms. SHHS’s extensive sleep database provides diverse, accurately labeled PSG data for model construction. Secondly, BSP2 algorithm training utilized only physiological parameters from PSG without including any demographic, anthropometric, or clinical data. This may help prevent dataset shifts when patient populations vary. Thirdly, BSP2 can calculate most time-domain and frequency-domain pulse rate variability metrics (performance evaluation manuscript in preparation). Fourthly, BSP2 allows sleep assessment for multiple nights, thus enabling longitudinal follow-up of a wealth of sleep physiology metrics. Consequently, BSP2 can be an invaluable tool in assessing sleep and may fill the gap between consumer-grade wearables and the standard sleep testing methods in clinical practice and research. One limitation of this study is that it was not performed in home settings. Another limitation is this sample population is skewed toward African Americans. But despite the concern that darker skin tones may lead to inaccuracy in SpO2 measurement, BSP2 performed well in all tested categories. It is worth mentioning that performance comparisons between devices using published historical data from various clinical trials are often challenging as these clinical cohorts may have diverse sample sizes, demographics, comorbidities, and OSA severity. Moving forward, a head-to-head cross-device comparison with multiple medical-grade OSA-detecting devices should be performed, as has been carried out for consumer-grade sleep technologies.14,47–49

Core conclusion

Belun Ring with BSP2 algorithms is the first DL-facilitated wearable capable of providing OSA diagnosis, severity categorization, sleep-wake estimation, and conventional 3-stage classification with reasonable accuracy in subjects suspected of OSA. Further clinical investigations on how best to employ BSP2 for assessing sleep disorders are warranted.

Data sharing statement

The data presented in this study are available by request through the corresponding author.

Acknowledgments

This work was performed at the University Hospitals Cleveland Medical Center Sleep Lab and the University Hospitals Beachwood Sleep Lab. We acknowledge Subhra Chakraborti, the University Hospitals Sleep Labs manager, and the sleep lab technicians for their expertise and assistance in data collection. This work was financially supported by the Belun Technology Company Limited under Grant UHCMC-2021-1.

Funding

This clinical research was supported by grant UHCMC-2021-1 from Belun Technology Limited, Hong Kong. The Belun Rings used in this study were provided by the company. Belun Technology agreed with the design of the study and has no role in the data collection. The sponsor did contribute to the data analysis, decision to publish, and manuscript preparation.

CRediT authorship contribution statement


Declaration of conflicts of interest

Wenbo Gu is a Ph.D. student at the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, and is also an engineer at Belun Technology Company. Drs. Lydia Leung, Chih-Wei Tsai, and Cynthia Cheung are Belun Technology employees. Dr. I-Chen Wu is a professor of Computer Science at National Yang Ming Chiao Tung University in Hsinchu, Taiwan. Dr. Wu has received a research grant from Belun for Belun Sleep Platform software algorithm development but has no other financial conflicts of interest. Dr. Ambrose Chiang has received two research grants from Belun Technology Company for conducting Belun Sleep Platform validation trials at University Hospitals Cleveland Medical Center but has no other financial conflicts of interest. The rest of the authors have no financial conflicts of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.sleep.2023.05.001.

References


