NSU Florida
NOVA SOUTHEASTERN UNIVERSITY

NeuroSports

Volume 1 | Issue 1 Article 11

July 2020

Validation of a Single Channel EEG for the Athlete: A Machine Learning Protocol to Accurately Detect Sleep Stages

Kayla Thompson
Nova Southeastern University, kt986@mynsu.nova.edu

Kamil Celoch
KC- Performance, info@kc-performance.com

Frankie Pizzo
Nova Southeastern University, fp0@mynsu.nova.edu

Ana I. Fins
Nova Southeastern University, anaifins@nova.edu

Jaime Tartar
Nova Southeastern University, tartar@nova.edu

Validation of a Single Channel EEG for the Athlete: A Machine Learning Protocol to Accurately Detect Sleep Stages

Abstract

There is a large and growing movement towards the use of wearable technologies for sleep assessment. This trend is largely due to the desire for comfortable, burden free, and inexpensive technology. In tandem, given the competitive nature of professional athletes enduring high training load, sleep is often jeopardized which can result in adverse outcomes. Wearable devices hold the promise of increasing the ease of monitoring sleep in athletes which can inform health and recovery status, as well as aid performance optimization. However, wearable devices typically lack sufficient validity to assess sleep – and especially sleep stages. To address this concern, the present study aimed to validate an algorithm to detect wakefulness, light sleep, deep sleep, and REM sleep against the gold standard polysomnography (PSG), using a wearable single channel electroencephalogram (EEG). Through the single channel EEG, machine learning models were built to infer sleep staging. The model was created from training and validating EEG output and labels assigned from the PSG software. Additionally, to determine the accuracy of agreement between the devices both Random Forest and a deep learning Convolutional Neural network model were implemented. The sleep staging output was consistent with our sleep staging algorithm for the single channel EEG and more notably, the sleep versus wake agreement was strong- above 80%. Our findings show that machine learning algorithms can be used with wearable devices to accurately detect, not only the sleep versus wake cycles, but the 4 sleep stages as well. Accordingly, this technology can be applied in an athlete population for accurate assessment of full sleep architecture.

Keywords

electroencephalogram, machine learning, sports performance, polysomnography, sleep, software

Introduction

Given the competitive nature of professional sports, it is well understood that athletes are continuously in need of innovative technologies and modalities to gain an edge to optimize their performance and health (Casey et al., 2012; Park et al., 2015; Thompson et al., 2008). There has been a large and growing shift in the athletic community towards the use of wearable devices as a means to monitor training progress and recovery (Seshadri et al. (2019). This is evidenced by an ever-growing sports performance technology market which offers smart watches, bands, garments, and patches with inbuilt sensors (Peake et al., 2018). Despite this, there is limited peer reviewed validation studies for wearables in spite of their increased incorporation in sports as a mean of monitoring athletes' workload (Seshadri et al., 2019). With the advent of miniaturized sensors, integrated computing, and artificial intelligence (Peake et al., 2018), it is expected that the emerging data-driven health and performance technologies will be of increased relevance in the field of sports performance (Perez-Pozuelo et al., 2020).

Given the well-established link between sleep and athletic performance, as well as sleep and traumatic brain injury (TBI), many sports practitioners turn to brain imaging and neurophysiological measures in the hopes of improving the recovery capacity and sports performance of their athletes (Jaffee et al., 2015; Knufinke et al., 2018; Murdaugh et al., 2018). Under normal physiological conditions, exercise is thought to have a positive impact on sleep (Walsh et al., 2021). However, high training load or injury may jeopardize sleep, and consequently impair recovery. It has been noted that heavy competition schedules, stress, brain injury, commute, academic demands, circadian misalignment, and overtraining have been all identified as potential obstacles to obtaining proper sleep (O'Donnell et al., 2018). Furthermore,

previous studies have demonstrated that athletes are particularly susceptible to sleep loss around competition time, further highlighting the need for a reliable way of monitoring sleep in this demographic (O'Donnell et al., 2018; Walsh et al., 2021; Watson, 2017).

Characterized by habitual short length (<7 hours/night) and poor sleep quality (e.g., fragmentation), sleep inadequacies have been shown to negatively affect a number of variables that underpin athletic performance, including the rate of perceived exertion, motor skill acquisition, injury rate, as well as a range of cognitive skills, such as accuracy, and reaction time (O'Donnell et al., 2018; Walsh et al., 2021; Watson, 2017). Moreover, it has been demonstrated that athletes show poor self-assessment of their sleep duration, and quality (Watson, 2017) which demonstrates a need for an objective measure. In light of this, it has been postulated that athletes may require more careful monitoring to identify individuals at risk of developing allostatic overload due to insufficient capacity to recover (O'Donnell et al., 2018). Hence, validation of sleep monitoring wearables will allow for sleep architecture and characteristics to be adequately attained and used to inform recovery and performance metrics.

Polysomnography (PSG) is the gold standard tool for clinical diagnosis of sleep disorders and for accurate determination of sleep-wake stages. The parameters for scoring normal adult sleep are provided by the American Academy of Sleep Medicine (AASM) Manual for Scoring Sleep Stages and Associated Events (Iber et al., 2007). To accurately assess clinical sleep disorders, PSG takes advantage of multiple recorded parameters. Electroencephalogram (EEG) serves as the main parameter during sleep and is coincided with respiratory function, heart rate, and blood pressure. These parameters, while critical for accurate clinical assessment and diagnoses, are not necessary for the accurate determination of sleep-wake states and sleep stages since EEG is the primary parameter (Vaughn & Giallanza, 2008). EEG measures of sleep and

wakefulness reliably show predictable and patterned cycles through sleep stages (Koley & Dey, 2012). The macrostructure of sleep consists of non-rapid eye movement (NREM) and rapid eye movement (REM) sleep. Wakefulness largely consists of beta rhythm, low-voltage fast EEG activity. The EEG displays clear patterns of decreased neural activity from the transition from wakefulness to NREM sleep. Initially, in stage N1 sleep this activity reflects a decrease in alpha activity followed by a transition into stage N2 sleep, which consists of EEG sleep spindles and K-complexes. The N2 stage is classified as a period of light sleep (LS), which accounts for 50% of an entire night's sleep. Following N2 sleep, there is an observable increase in EEG amplitude and predominance of delta activity in stage N3, which is referred to as a period of deep sleep (DS). Rapid eye movement REM (R), also known as paradoxical sleep, is characterized by a loss of muscle tone accompanied by low amplitude fast EEG in the theta range.

While the PSG provides an accurate and clinically useful measure of sleep, it includes multiple factors that limit sleep assessment for research purposes (Perez-Pozuelo et al., 2020). For example, the need for a research subject to spend the night in the sleep lab results in an unnatural night of sleep. In addition, the cost and manpower needed for a full night PSG study limits the number of participants in a study. For this reason, alternative in-home polysomnography was developed (Kundel & Shah, 2017). Although cost effective, polysomnography for home use has been limited by the difficulty in setting up the device and discomfort associated with multiple wires. Consequently, it is common for researchers to utilize wearable technology for sleep assessment as an alternative to the PSG. A common alternative known for its simplicity is wrist-worn devices which measure multiple bioelectric signals such as heart rate, skin conductance, temperature, and movement/activity to provide an assessment of sleep behavior (De Zambotti et al., 2019). Multiple studies have investigated the ability of

wearable wrist-worn devices to accurately assess sleep, which approximate self-reported sleep time. Furthermore, these devices tend to correlate well with each other on total sleep time measures (De Zambotti et al., 2019; Haghayegh et al., 2019; Lee et al., 2019; Meltzer et al., 2015). Notably, however, is the tendency of wearable technology to overestimate sleep time (De Zambotti et al., 2019; Kushida et al., 2001) and sleep efficiency (Bhat et al., 2015; Haghayegh et al., 2019) in healthy adults.

The determination of sleep from wakefulness and accurate sleep staging are measures from electrophysiological signals from the scalp. Previous reports have shown that a single channel "wearable" EEG headband can also accurately detect wake and sleep stages compared to PSG and actigraphy (Kosmadopoulos et al., 2014; Shambroom et al., 2012). This creates a desire to expand on existing technologies using brain-based recording to accurately classify sleep time as well as EEG measures of sleep stages. The present study builds on this previous finding by showing that a machine learning algorithm can accurately provide sleep staging analyses, despite not having multiple electrode sites in the recording. We show that a wireless device can automate sleep staging in real-time using 30 second epochs with a single channel fabric headband. To overcome in-home difficulty, the device uses Bluetooth and features a user-friendly mobile in-app software using a smartphone. Previous studies have attempted other physiological measures to predict sleep stages, such as pulse, blood oxygen and motion sensors (Zhang et al., 2012); however, they were unable to detect a differentiation between N1 and N2. A main issue addressed in this paper, is the lack of validation for alternative sleep staging devices resulting in them being pulled off the athlete-consumer and medical market (Behar et al., 2013; De Zambotti et al., 2019). Lack of reliability and validity testing has been identified as a threat to the use of data-driven applications in sleep medicine and research (Perez-Pozuelo et al., 2020).

Accordingly, the overall goal of the current study was to investigate whether the output from the single channel EEG device is sufficient for accurate sleep stage inferences, using the proposed machine learning algorithm. Notably, the algorithm was validated through inferences of wakefulness, light sleep (LS), deep sleep (DS) and REM sleep against the gold standard polysomnography (PSG).

Methods and Materials

Subjects

Fifteen subjects were recruited from Nova Southeastern University (n = 15; 5 females, 10 males, mean age=25.2, SD =9.13), of which 8 completed an overnight sleep study and 7 completed a napping study. This study was carried out according to a protocol approved by the Nova Southeastern University Institutional Review Board. (IRB NSU-2018-646). All participants received a verbal explanation of the study procedures and signed an NSU IRB-approved written Informed Consent Form. Exclusionary criteria included a prior history of drug or alcohol abuse, neurological, psychiatric or sleep disorders. PSQI scores (M=5.73, SD= 1.94) were obtained from all subjects.

Procedure

Participants were tested in the NSU sleep laboratory, (Fort Lauderdale, FL) on one occasion. Testing arrival time was 12-5 p.m. for the daytime nap or between 10-11:30 p.m., for the overnight sleep study, according to their typical bed times. Nap participants were provided a 4-hour sleep opportunity in an individual room, while those engaging in overnight sleep were provided 9 hours of time in bed. The sleep lab was equipped with automated blackout shades

which were closed when the researcher left the room and all lighting in the room was turned off. Participants were also asked to turn off all electronic devices they had with them. Participants were connected to the polysomnography with electrodes attached to the face and scalp, along with a wireless ambulatory sleep-monitoring device on the forehead which connected behind the mastoid bone, for concurrent monitoring. Participants remained under continuous EEG monitoring by a researcher that was stationed in the monitoring room next door. To obtain corresponding epochs, both sleep monitoring devices, the PSG and wireless device, were programmed to store data in 30-s epochs. In addition, clock times were aligned by concurrent recording start times, along with synchronizing the time to the same computer clock prior to each recording. Agreement between the PSG and wireless device were then evaluated.

Materials

Pittsburgh Sleep Quality Index

The Pittsburgh Sleep Quality Index (PSQI) is a reliable self-report measure used to assess sleep quality and patterns in adults. The index consists of 19 items indicative of 7 component scores which convey sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, the use of sleep medications, and daytime dysfunction. Each component is self-rated by the participant. These components yield a score ranging from 0 to 21, in which a score above five distinguishes between those with poor sleep versus those with good sleep (Buysse et al., 1989).

Polysomnography

The PSG was conducted using the Alice 5 and G3 Sleepware (Respironics, Murraysville, PA) and Grass gold-cup electrode leads (Astro-Med, Inc., West Warwick, RI). Four channels of electroencephalography were used to measure brain activity at the central (C3-A2, C4-A1) and frontal (F3-A2, F4-A1) lobes; eye movements were monitored with right and left electro-oculo-grams, and two channels of submental electromyography placed bi-lateral to measure muscle tone. In addition, reference electrodes were placed on each earlobe. Prior to recording, a routine calibration and impedance check below 5 K Ω were performed to confirm the signal.

Wireless Sleep Monitoring Device

The wireless sleep monitoring device used in this study was the Enchanted Wave headband (Enchanted Wave, LLC, Miami, FL). The Enchanted Wave EEG device is an ambulatory, wireless sleep staging tool that includes a headband containing two dry electrodes which record signals from the forehead at the Fp1 region based on the 10-20 system of electrode placement. Alongside these sensors are two metallic fabric electrodes by the mastoid bone which require skin contact. At the end of each recording, the device's automated analysis scores these signals into sleep and wake stages; to ensure optimal performance of the algorithms analysis, consistent signal integrity is to be maintained. Classifications of wake and REM sleep were reported according to standard definitions. Time spent in each sleep stage was accounted for, along with categorization of light sleep (LS), deep sleep (DS), total sleep time (TST), sleep efficiency (SE), spindles, alpha waves, beta waves, theta waves, and delta waves.

Sleep Staging Algorithm

Random Forest (RF) is an ensemble machine learning method (Breiman, 2001). A comparison of feature and classifier algorithms for online sleep staging based on a single EEG Signal found that the random forest model outperformed the support vector machine ensemble. The Random Forest works by modeling decision rules and as such, resembles the AASM methodology of sleep staging (Radha et al., 2014). The Random Forest (RF) model learned a set of estimators from training data. Each estimator is a decision tree that can make classification decisions hierarchically based on selected feature values. Each estimator in a RF model is learned from a subset of entire data through random sampling. The inference is determined by the average of inference results from all the estimators. The model used the Sci-kit Learn Python package (version 0.23.1) to implement the RF method used in this project. Each RF model includes 100 estimators. Based on this number, the decision trees are built. Each split decision is determined based on a Gini Coefficient with up to 16 selected features (the square root value of total features from original data sets).

The deep neural network approach utilized network architecture inspired by LeNet-5 (LeCun et al., 2015). The implementation uses Tensorflow Version 1.10.0 and Keras 2.2.4. The training process was limited to 100 epochs. Additionally, the batch size was set at 32, and the learning rate was 0.005. The implementation uses the Adam algorithm, which is a stochastic gradient descent method based on adaptive estimation of first-order and second-order moments (Kingma et al., 2014). In both cases of the RF and deep neural network approach, the training set was a random sampling from the original data sets, while the remaining dataset was used for accuracy testing and validation.

Results

A total of 15 subjects (10 males, 5 females; M=25.2, SD=9.13) were included in the present analysis to evaluate the accuracy of the single channel EEG recording in comparison to the standard PSG consensus. For further descriptive statistics on sleep of the participants, refer to Table 1. Data were excluded based on incomplete PSG or EEG data, due to technical problems or technician/subject error. A Random Forest model was implemented for analysis using 100 estimators. The goal of the present analysis was to classify performance with all outputs from the EEG device of sleep architecture. Given the slight variability in PSG models classification of sleep architecture, the present study validated the single channel device output against 2 PSG software outputs (i.e. Alice 5 and G3) for generalizability. One subgroup (labeled Group A, n=10) was classified as those who had undergone the Alice 5 software for the PSG, while the other subgroup (labeled Group B, n=5) had undergone the G3 software. Group A included those who participated in the napping and overnight studies (Alice 5 software), whereas Group B only included overnight participants (G3 software). Both group A and B data sets were time-synchronized, and data was analyzed using 30-sec epochs where average values were used as feature vectors. Each night of complete data was normalized using a min-max scaler.

Given the large variance among number of measurements for various sleep staging, a balanced data set was created for each group through subsampling, to ensure scope in the validation process. In the balanced dataset, there are the same number of measurements for each sleep staging. On the other hand, full data sets included all data in the group. Analyses for all groups were then conducted using five runs of five-fold cross validation with both the balanced and full data sets. A summary, reporting accuracy of the Single channel EEG is provided in Table 2-3 and Figures 1-2. The average cross validation for Group A's four class analysis, yielded a high average of 0.71 for the balanced test. The balanced data set offered a small range

of variability and higher scores. The two-class analysis of sleep versus wakefulness agreement with the PSG for Group A resulted in an average accuracy of 0.82. In agreement with the findings yielded in Group A, the balanced data set in Group B's four class analysis, yielded robust results with an average of 0.68. In the two-class analysis of sleep versus wakefulness for Group B, the accuracy had an average 0.80. A follow up test was conducted using a curated dataset that was randomly sampled and manually scored for sleep classification by an expert against the algorithms performance. The group was labeled Group C. Following the trend of the previous results in group A and B, the balanced datasets in Group C yielded robust results, however notably, they were better results with less variation at an average of 0.75. Here, the best model performance from the balanced data set was 0.77. At the classification of only two classes: wakefulness vs sleep, the accuracy is remarkably increased, yielding results of 0.91.

For further validation and scope of the analysis, a deep learning convolutional neural network (CNN) model was implemented, whereby the full data of the current epoch and the calculated features of several previous epochs were used and assembled as a two-dimensional vector then used as input. This allows the convolutional layers to create a network that learns relevant features and/or local patterns. Unlike random forest, the neural networks consist of the data not only over a 30 second window but as well the history, as it can learn without any a priori feature selection. Accordingly, Group C could not be manually classified for the CNN model. Four convolution layers were used for the network and five runs were conducted across full data sets with groups A and B only. At convergence, the average validation accuracy yielded was 0.74 for Group A and 0.69 for Group B. In reference to the two classes, Group A yielded higher accuracy of 0.88, while Group B was 0.86.

Discussion

The performance of sleep staging was evaluated in a single channel EEG recording. Light Sleep, Deep Sleep, REM Sleep and Wakefulness were defined using a 30 second epoch comparison against the automated PSG sleep staging software (Alice 5 and G3 software). In both PSG software programs, the sleep staging output was consistent with our sleep staging algorithm for the single channel EEG. Random forest analyses resulted in complete stage agreement of 0.71 in group A (Alice 5 software) and 0.68 in group B (G3 software) balanced datasets. Notably, for the wake vs. sleep staging, there was strong agreement with the PSG- above 80%. The deep learning analyses were consistently higher than the random forest analyses, most likely due to their nature of, not only considering the current moment, but also the priori moment in tandem. Accordingly, the deep learning model had increased agreement for the complete staging with 0.74 in Group A and 0.69 in Group B. Similarly, the two-class staging agreement was higher than the 4-stage analysis, at 88% and 86% respectively. Critically, our results are within range of those previously reported in studies using actigraphy and similar single channel EEG with sleep staging algorithms (Shambroom et al., 2012; Viera & Garrett, 2005; Wang et al., 2015). Accepted ranges are reported at 0.63 and above 80% suggests strong agreement (Mikkelsen & De Vos, 2018).

Of note, we found that there was a discrepancy between the agreement of the two PSG software algorithms. Our results were more closely aligned with the Alice5 software relative to the G3 software, despite G3 being the more updated software package. Upon review of the data, more staging errors were found in G3 than in Alice5, which may relate to the discrepancy. To further investigate this possibility, we manually scored a curated data set using human experts, labelled Group C, where the agreement evaluation showed superior results compared to the PSG automated output. We speculate that the improved results might relate to the fact that manually scored sleep staging is less error prone than automated scoring by the PSG, which is especially

applied in cases of sleep disturbances or disorders (Aşık et al., 2014). Hence it is recommended, that combining automated and manual scoring offers good diagnostic agreement (BaHammam et al., 2011). Indeed, the common practice observed in sleep medicine is to have multiple PSG technologists verify the results and manually score them for inter-rater reliability. This suggests a human's judgement is still considered the "gold standard", while the PSG provides an expedited assembly of the sleep scoring process.

Although the results of the present study clearly demonstrate the ability of the machine learning algorithm to robustly classify sleep-wake stages, a larger sample size would be beneficial to further define and describe the algorithm. Furthermore, the study aimed to identify a broad range of sleep types to account for variation in sleep patterns for a reliable conclusion, thus, participants were not screened for sleep disorders. Future research should aim to replicate these findings with healthy participants, screened for sleep disorders. In addition, the population should extend to older adults to gather a more diverse training set and understand the limitations and strengths of the algorithm. This may allow for enhanced performance, given that larger and more diverse training sets increase the performance of classifiers (Mikkelsen & De Vos, 2018). Likewise, our comparison study was conducted on a single night, and therefore were unable to assess test-retest reliability. Despite these limitations, the findings suggest that the sleep staging algorithm is robust in distinguishing sleep stages with a single channel EEG. This holds the promise for sleep monitoring to be less obtrusive and more comfortable in data acquisition, with the future possibility of implementing less-intrusive and well validated monitoring in clinical and research settings. Furthermore, this study lends credibility to the use of a wearable single channel EEG device for further use amongst athletes as a valid alternative to PSG. Such validation pro-

motes the investigation of relationships which are suggested to be of primary relevance in athletes such as sleep and brain injury or recovery, as well as sleep and performance. In a similar vein, the ability to obtain PSG-concordant consecutive nights of data with a single channel EEG headband in one's home environment and on the field could open numerous possibilities for research designs that have not previously been possible.

Conflict of Interest Statement. Dr. Jaime Tartar serves as a scientific advisor for Enchanted Wave, LLC. To date she has not received any payment or resources in this role.

Tables

Table 1

Descriptive Statistics for Variables of Interest

Variable	M	SD	n
Age	25.2	9.13	15
PSQI	5.73	1.94	15
		Overnight Sleep	
TST	349	39	8
NREM	275	61	8
REM	72	58	8
		Napping	
TST	63	26	7
NREM	62	26	7
REM	2	2	7

[1] The Pittsburgh Sleep Quality Index score (PSQI) and age is reported for all participants. The total number of participants (n) and average (M) duration for total sleep time (TST), Non-rapid eye movement sleep (NREM), and rapid eye movement sleep (REM) for those in the napping and overnight sleep study are indicated.

Table 2

Complete Sleep Four-Staging

	Accuracy		
Algorithm	Group A	Group B	Group C
	Alice5	G3	Curated
Random Forest (bal.)	0.71	0.68	0.75
Convolutional Networks	0.74	0.69	N/A

[2] Cross validation for the four-class analysis using both the balanced (bal.) and unbalanced (unbal.) dataset using Random Forest (RF). Additionally, the accuracy using the deep learning Convolutional Network (CNN) was provided.

Table 3

Wake versus Sleep

	Accuracy		
Algorithm	**Group A**	**Group B**	**Group C**
	Alice5	**G3**	**Curated**
Random Forest	0.82	0.80	0.91
Convolutional Networks	0.88	0.86	N/A

[3] Cross validation for the two-class analysis using the balanced dataset for Random Forest (RF). Additionally, accuracy using the deep learning Convolutional Network (CNN) was provided.

Figures

Figure 1.

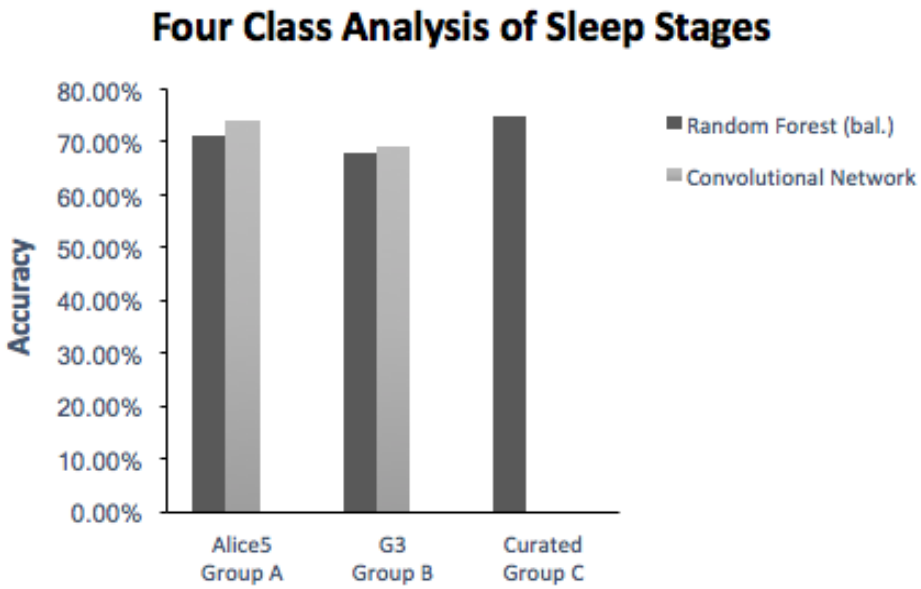
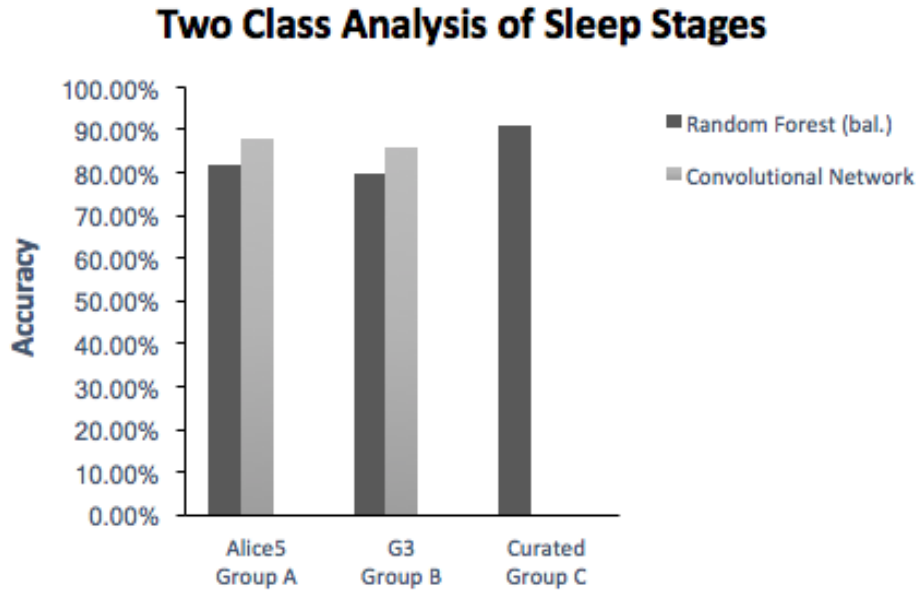


Figure 2.



References

Aşık, M., Bostancı, A., & Turhan, M. (2014). Comparison of Manual and Automated Scoring Techniques in Polysomnography. *Turkish Archives of Otolaryngology, 52,* 17-21.

BaHammam, A., Sharif, M., Gacuan, D. E., & George, S. (2011). Evaluation of the accuracy of manual and automatic scoring of a single airflow channel in patients with a high probability of obstructive sleep apnea. *Medical science monitor: international medical journal of experimental and clinical research, 17*(2), MT13.

Behar, J., Roebuck, A., Domingos, J. S., Gederi, E., & Clifford, G. D. (2013). A review of current sleep screening applications for smartphones. *Physiological measurement, 34*(7), R29.

Bhat, S., Ferraris, A., Gupta, D., Mozafarian, M., DeBari, V. A., Gushway-Henry, N., Gowda, S. P., Polos, P. G., Rubinstein, M., & Seidu, H. (2015). Is there a clinical role for smartphone sleep apps? Comparison of sleep cycle detection by a smartphone application to polysomnography. *Journal of Clinical Sleep Medicine, 11*(7), 709-715.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Buysse, D. J., Reynolds III, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry research, 28*(2), 193-213.

Casey, M., Yau, A., Barfoot, K., & Callaway, A. (2012). Data mining of portable EEG brain wave signals for sports performance analysis: An archery case study. *International Convention on Science, Education and Medicine in Sport* (ICSEMIS), 19--24 July 2012, Glasgow, UK.

De Zambotti, M., Cellini, N., Goldstone, A., Colrain, I. M., & Baker, F. C. (2019). Wearable sleep technology in clinical and research settings. *Medicine and science in sports and exercise, 51*(7), 1538.

Haghayegh, S., Khoshnevis, S., Smolensky, M. H., Diller, K. R., & Castriotta, R. J. (2019). Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis. *Journal of medical Internet research, 21*(11), e16273.

Iber, C., Ancoli-Israel, S., Chesson, A. L., & Quan, S. F. (2007). *The American Academy of Sleep Medicine manual for the scoring of sleep and associated events: rules, terminology*

and technical specifications. Westchester, Illinois: American Academy of Sleep Medicine, 2007.

Jaffee, M. S., Winter, W. C., Jones, C. C., & Ling, G. (2015). Sleep disturbances in athletic concussion. *Brain injury, 29*(2), 221-227.

Kingma, D. P., Mohamed, S., Rezende, D. J., & Welling, M. (2014). *Semi-supervised learning with deep generative models*. Advances in neural information processing systems.

Knufinke, M., Nieuwenhuys, A., Geurts, S. A., Møst, E. I., Maase, K., Moen, M. H., Coenen, A. M., & Kompier, M. A. (2018). Train hard, sleep well? Perceived training load, sleep quantity and sleep stage distribution in elite level athletes. *Journal of science and medicine in sport, 21*(4), 427-432.

Koley, B., & Dey, D. (2012). An ensemble system for automatic sleep stage classification using single channel EEG signal. *Computers in biology and medicine, 42*(12), 1186-1195.

Kosmadopoulos, A., Sargent, C., Darwent, D., Zhou, X., & Roach, G. D. (2014). Alternatives to polysomnography (PSG): a validation of wrist actigraphy and a partial-PSG system. *Behavior research methods, 46*(4), 1032-1041.

Kundel, V., & Shah, N. (2017). Impact of portable sleep testing. *Sleep medicine clinics, 12*(1), 137-147.

Kushida, C. A., Chang, A., Gadkary, C., Guilleminault, C., Carrillo, O., & Dement, W. C. (2001). Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. *Sleep medicine, 2*(5), 389-396.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Lee, X. K., Chee, N. I., Ong, J. L., Teo, T. B., van Rijn, E., Lo, J. C., & Chee, M. W. (2019). Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. *Journal of Clinical Sleep Medicine, 15*(9), 1337-1346.

Meltzer, L. J., Hiruma, L. S., Avis, K., Montgomery-Downs, H., & Valentin, J. (2015). Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep, 38*(8), 1323-1330.

Mikkelsen, K., & De Vos, M. (2018). Personalizing deep learning models for automatic sleep staging. *arXiv:1801.02645*.

Murdaugh, D. L., Ono, K. E., Reisner, A., & Burns, T. G. (2018). Assessment of sleep quantity and sleep disturbances during recovery from sports-related concussion in youth athletes. *Archives of physical medicine and rehabilitation, 99*(5), 960-966.

O'Donnell, S., Beaven, C. M., & Driller, M. W. (2018). From pillow to podium: a review on understanding sleep for elite athletes. *Nature and science of sleep, 10*, 243.

Park, J. L., Fairweather, M. M., & Donaldson, D. I. (2015). Making the case for mobile cognition: EEG and sports performance. *Neuroscience & Biobehavioral Reviews, 52*, 117-130.

Peake, J. M., Kerr, G., & Sullivan, J. P. (2018). A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. *Frontiers in physiology, 9*, 743.

Perez-Pozuelo, I., Zhai, B., Palotti, J., Mall, R., Aupetit, M., Garcia-Gomez, J. M., Taheri, S., Guan, Y., & Fernandez-Luque, L. (2020). The future of sleep health: a data-driven revolution in sleep science and medicine. *NPJ digital medicine, 3*(1), 1-15.

Radha, M., Garcia-Molina, G., Poel, M., & Tononi, G. (2014). *Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal*. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Seshadri, D. R., Li, R. T., Voos, J. E., Rowbottom, J. R., Alfes, C. M., Zorman, C. A., & Drummond, C. K. (2019). Wearable sensors for monitoring the physiological and biochemical profile of the athlete. *NPJ digital medicine, 2*(1), 1-16.

Shambroom, J. R., Fábregas, S. E., & Johnstone, J. (2012). Validation of an automated wireless system to monitor sleep in healthy adults. *Journal of sleep research, 21*(2), 221-230.

Thompson, T., Steffert, T., Ros, T., Leach, J., & Gruzelier, J. (2008). EEG applications for sport and performance. *Methods, 45*(4), 279-288.

Vaughn, B. V., & Giallanza, P. (2008). Technical review of polysomnography. *Chest, 134*(6), 1310-1319.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med, 37*(5), 360-363.

Walsh, N. P., Halson, S. L., Sargent, C., Roach, G. D., Nédélec, M., Gupta, L., Leeder, J., Fullagar, H. H., Coutts, A. J., & Edwards, B. J. (2021). Sleep and the athlete: narrative review and 2021 expert consensus recommendations. *British journal of sports medicine, 55*(7), 356-368.

Wang, Y., Loparo, K. A., Kelly, M. R., & Kaplan, R. F. (2015). Evaluation of an automated single-channel sleep staging algorithm. *Nature and science of sleep, 7*, 101.

Watson, A. M. (2017). Sleep and athletic performance. *Current sports medicine reports, 16*(6), 413-418.

Zhang, J., Chen, D., Zhao, J., He, M., Wang, Y., & Zhang, Q. (2012). *Rass: A portable real-time automatic sleep scoring system.* 2012 IEEE 33rd Real-Time Systems Symposium.