

Time–frequency ridge characterisation of sleep stage transitions: Towards improving electroencephalogram annotations using an advanced visualisation technique

Christopher McCausland^{a,*}, Pardis Biglarbeigi^b, Raymond Bond^c, Golnaz Yadollahikhales^d, Alan Kennedy^a, Anna Sigridur Islind^{e,f}, Erna Sif Arnardóttir^{e,f}, Dewar Finlay^a

^a Nanotechnology and Integrated BioEngineering Centre (NIBEC), Ulster University, Belfast, BT15 1ED, UK

^b School of Science and Engineering, University of Dundee, Dundee, DD1 4HN, UK

^c School of Computing, Ulster University, Belfast, BT15 1ED, UK

^d Jona Goldrich Center for Alzheimers and Memory Disorders, Neurology Department, Cedars-Sinai Medical Centre, Los Angeles, CA 90048, USA

^e Department of Computer Science, Reykjavík University, Reykjavík, Iceland

^f Reykjavík University, Sleep Institute, Reykjavík, Iceland

ARTICLE INFO

Keywords:

Automatic scoring
Electroencephalography
Sleep staging
Time–frequency analysis
Interpretable AI

ABSTRACT

Manual sleep stage scoring of polysomnography recordings is an expensive and time-consuming process, further complicated by inconsistent sleep stage agreement among sleep experts (clinicians and sleep technologists). Hence, development of automated sleep scoring algorithms are an emerging topic of interest. Automation typically mimics the clinical decision path by implementing a series of predefined rules, such as the American Academy of Sleep Medicine's (AASM) scoring manual. Recently, data driven methods have emerged using machine or deep learning. Both manual and automated methods of scoring have known limitations; primarily, unacceptable variation in agreement between different scorers and algorithms. Within the literature, electroencephalogram (EEG) frequency is an important feature considered by both sleep experts and automated approaches for classifying sleep stages.

This study presents a novel approach to sleep stage analysis, by developing a methodology to precisely determine the temporal location of sleep stage transitions. The current gold standard fails to identify such transitional changes, which leads to poor inter-scorer reliability. Therefore, development and implementation of such methodologies is a crucial, but overlooked, step in improving the consistency of scoring within sleep studies. In this work, EEG time–frequency ridge analysis was used to characterise the dominant frequency component of EEG signals in time, at the point of sleep stage transition. An in-depth analysis of N3 → N2 and N2 → N3 transitions in the 2018 PhysioNet challenge “You Snooze, You Win” and the Wisconsin Sleep Cohort (WSC) datasets ($n = 994$, $n = 742$; approximately 13,888 h of sleep data) showed consistent time–frequency patterns at the point of transition, from one sleep stage to another. This methodology allows simple and ‘interpretable’ features to be generated in future work, to precisely identify the temporal location of sleep stage transitions with the aim of improving inter-scorer reliability.

1. Introduction

Sleep is a basic function required for survival (Liew & Thidar, 2021; Pienaar et al., 2021; Shattuck & Sparks, 2022). While the specific purpose of sleep continues to be a topic of much discussion (Anafi et al., 2019; Galván, 2020; Hudson et al., 2020), there is a consensus that a lack of sleep, or fragmented sleep, due to repeated waking, is

detrimental to quality of life (D'Ambrosio et al., 1999; Medic et al., 2017). Sleep disorders, therefore, represent a significant public health issue, with a high prevalence within society (Heinzer et al., 2015; Peppard et al., 2013). Several negative health outcomes for patients with sleep disorders are noted in literature (Medic et al., 2017). As with any disease, early detection and treatment is critical for improved statistical outcomes for patients.

* Correspondence to: Nanotechnology and Integrated BioEngineering Centre (NIBEC), Ulster University, Belfast, BT15 1ED, Northern Ireland, UK.

E-mail addresses: c.mccausland@ulster.ac.uk (C. McCausland), pbiglarbeigi001@dundee.ac.uk (P. Biglarbeigi), rb.bond@ulster.ac.uk (R. Bond), golnaz.yadollahikhales@cshs.org (G. Yadollahikhales), a.kennedy@ulster.ac.uk (A. Kennedy), islind@ru.is (A.S. Islind), ernasifa@ru.is (E.S. Arnardóttir), d.finlay@ulster.ac.uk (D. Finlay).

<https://doi.org/10.1016/j.eswa.2024.125490>

Received 27 November 2023; Received in revised form 2 September 2024; Accepted 1 October 2024

Available online 22 October 2024

0957-4174/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Table 1
Overview of typical clinical features for manual sleep stage identification (Berry et al., 2020).

Sleep stage	EOG activity	EEG activity
Wakefulness (W)	Eye movement and blinking	Mixed frequency high amplitude, alpha waves (8–13 Hz)
Stage N1	Slow, rolling eye movement	Alpha waves disappear, Theta waves (4–7 Hz) appear
Stage N2	Eye movement ceases	Theta waves continue, sleep spindles (11–16 Hz) appear as rhythms in central channels, k-complexes appear as >0.5 s high sharp waves in frontal channels
Stage N3	Not typical	Delta waves appear (0.5–2 Hz) with peak-to-peak amplitude > 75 μ V, sleep spindles and k-complexes can persist
Stage REM	Rapid, irregular eye movement	Sawtooth waves (2–6 Hz) predominate in central electrodes, muscle noise may also be present

Polysomnography (PSG) is the current gold standard for clinical detection and severity rating of many sleep related disorders (Cook, 2021). A PSG consists of an overnight recording of several physiological signals including, but not limited to; electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), and electrooculography (EOG) (Arnardottir et al., 2021). The American Academy of Sleep Medicine (AASM) is widely regarded as the leading authority on sleep scoring guidelines. The last major revision of guidelines was in 2007, with annual minor revisions to date. Since publication of the revised AASM ‘Manual for the Scoring of Sleep and Associated Events’ (Berry et al., 2020), the manual has widely been adopted worldwide, thus providing an international standard for professionals to conform to. The AASM scoring process divides overnight recordings into 30-second windows known as ‘epochs’. Each epoch is visually assessed by trained sleep experts and assigned one of five sleep stages (wakefulness (W), stage N1, stage N2, stage N3 or rapid eye movement (REM)), a costly and time consuming process. Each stage includes characteristic events, a subset are noted in Table 1.

The development of automated diagnostic algorithms for sleep studies are therefore an expanding topic of interest in literature (Alsolai et al., 2022; Malekzadeh et al., 2022; Mendonça et al., 2019). Traditionally, such algorithms seek to mimic the clinician by implementing a series of decision based rules. However, modern data driven methods have emerged, with both Machine Learning and Deep Learning algorithms proposed, as reviewed in Alsolai et al. (2022), Malekzadeh et al. (2022).

A limitation of both automated and manual methods is the current qualitative nature of PSG signal analysis, for sleep staging and abnormal respiratory event (apnea) identification. As noted, the prime method for clinical investigation is visual analysis of time domain patterns within EEG signals, with no supporting quantitative material (Berry et al., 2020). Therefore, sleep experts are saddled with determining frequency information from time domain data. Considering sleep stage scoring, the limitations of quantitative methods can be calculated as the Inter-scorer reliability (ISR). ISR is the percentage likelihood of multiple sleep experts to score the same epoch with the same sleep stage label. Typically, ISR ranges from 70%–86% for manual scoring (de Chazal et al., 2022; Magalang et al., 2013; Nikkonen et al., 2023; Penzel et al., 2013; Rosenberg & Hout, 2013). Considering the study in Rosenberg and Hout (2013) an ISR of 83% was reported using an AASM Inter-scorer reliability dataset. This dataset consists of 1800 epochs and 2500 individual scorers. ISR for stage W, N2 and REM exceeded 83%, however, reduced for stages N1 and N3 to 63% and 67%, respectively. Automated scoring performance has a similar ISR range to manual scoring. However, it has been argued that this perceived ISR ceiling could be due to limitations of clinical labels, as input ‘ground truth’ clinical labels for machine and deep learning algorithms feature only ~80% ISR (de Chazal et al., 2022). Inaccurate input results in an inaccurate learned output. Therefore, improving manual ISR could also improve automated ISR through enhanced reliability of labels for automation. It is also important to consider the location of such labelling inconsistencies; the consensus for these lower ISR scores is that most disagreement occurs in determining the epoch of

transition between adjacent sleep stages (Penzel et al., 2013; Rosenberg & Hout, 2013); i.e., when moving from one sleep stage to a different sleep stage, experts may disagree by several epochs as to when the exact transition occurs. Minimising such disagreements is therefore an important step to improve the quality of scoring and further develop automated approaches, however, has yet to receive significant attention in literature.

Quantitative analysis of physiological bio-signals is a developing field, with applications emerging for quantitative EEG (qEEG) for epilepsy, anaesthesia and neurological disease and injury monitoring (Choi et al., 2020; Geraedts et al., 2018; Mesraoua et al., 2019; Noor & Ibrahim, 2020). Such methods were developed using frequency domain, time–frequency domain and non-linear methods. However, much work is still required to develop complementary methods for sleep analysis. In addition, many of these data driven features are purely mathematical concepts, which do not fall under the remit of ‘interpretable’ artificial intelligence. Creating interpretable artificial intelligence models is critical to increase clinical trust in automated approaches (Ferrario & Loi, 2022). Increasing trust, through interpretability, is required to subsequently proliferate widespread use of automated algorithms within a clinical setting (Reddy et al., 2019). Development of methodologies which can fuse qEEG in an interpretable manner represent a significant breakthrough, which is more likely to be accepted by sleep experts, when compared to purely data driven methods.

Given the low ISR across both manual and automated scoring – especially around the point of transition – coupled with advances in qEEG, robust simple features could significantly improve the ISR. Before such a feature is developed, a clear understanding of the characteristics of sleep stage transitions is required, as detailed in previous work (McCausland et al., 2022).

1.1. Comparison and limitations of existing methods

Current automated approaches for sleep staging, as independently reviewed by Gaiduk et al. (2023), Yue et al. (2024), Zhang et al. (2024), predominantly employ supervised, epoch-by-epoch classification. In these methods, epochs are classified into one of the five sleep stages based on statistical similarity to expert-labelled epochs, using either machine learning or deep learning techniques. However, as (Gaiduk et al., 2023) argued; despite the significant research efforts in this area, there has not been a substantial breakthrough in performance improvements of such algorithms.

Furthermore, Penzel (2022) highlighted the disparity in accuracy expectations between domains, noting that while errors below 5% are expected in fields such as mammography image classification, sleep stage scoring often tolerates errors of 15% or more. This issue is further exacerbated by the well-documented challenges in ISR discussed within the Introduction.

Epoch-by-epoch methods have inherent limitations, particularly given that sleep stage transitions are gradual and do not always align neatly with fixed-duration epochs. Recognising these limitations, Stephansen et al. (2018) introduced the concept of “hypnodynensity” graphs, which represent sleep staging as a probability distribution

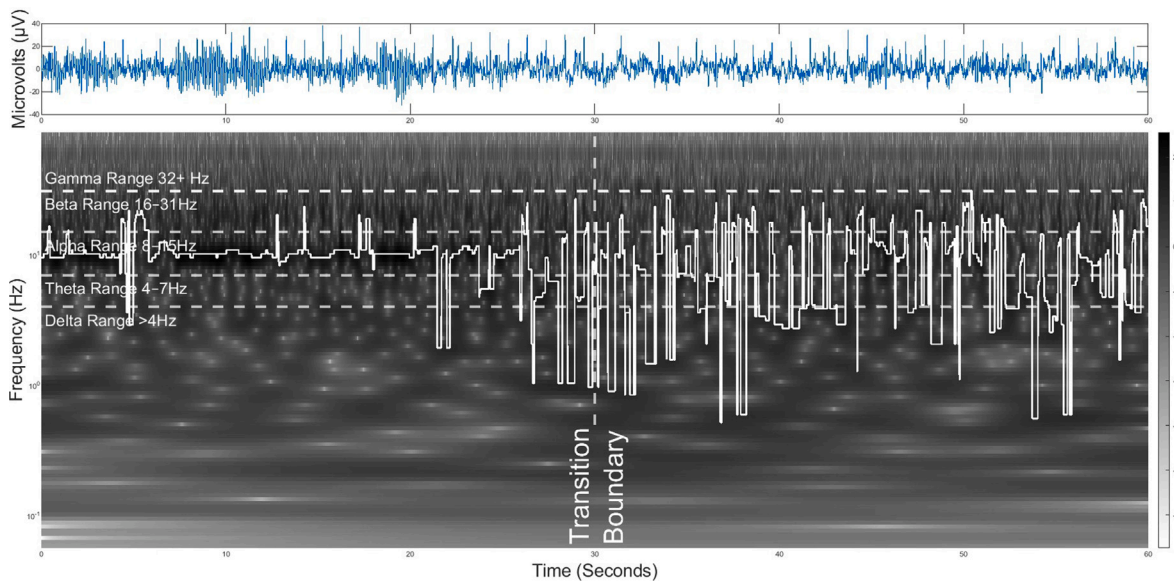


Fig. 1. A comparison of visually scoring EEG (blue line, top) vs time-frequency ridges (white solid line, bottom). The same two epoch transition from Wakefulness to sleep stage N1 is shown in both raw EEG and time-frequency ridge. This highlights the difficulties of visually scoring EEG data, as the exact location of the transition is not obvious, in contrast, even a casual observer can see clear alpha activity indicating wakefulness, which transitions to low amplitude, mixed-frequency (LAMP) wave forms commonly seen in sleep stage N1 at approximately 26 s. This transition is not readily apparent when considering the EEG trace.

rather than a categorical classification. This approach, which extends traditional hypnograms, marked a significant shift in how sleep data is interpreted.

Building on this work, Bakker et al. (2023) incorporated the hypn-density approach into automatic scoring systems, reasoning that moving away from categorical sleep stage certainty allows for the quantification of “sleep stage ambiguity”. Anderer et al. (2023) further justified this approach by noting that scorers are often required to assign sleep stages even when the underlining signals are ambiguous.

While hypn-density represents a major advancement in sleep analysis, concerns remain about the limitations of automated sleep scoring systems. As de Chazal et al. (2022) noted, the maximum performance of these systems is still constrained by the ambiguity of the input labels generated by manual scoring. Moreover, Huijben et al. (2023) highlighted that the exact relationship between PSG data, hypn-density, and manual scoring remains unclear, which hinders clinical acceptance of these methods.

It is increasingly evident that traditional epoch-by-epoch approaches are becoming outdated in the field of sleep staging, with probability approaches becoming the favoured method for automation. The reliance on manually scored training labels is a significant limitation, and improving the accuracy and reliability of these labels is essential for the development of more effective automated sleep staging systems.

1.2. Study aim and novelty

Unlike typical sleep stage algorithms, which operate on an epoch-by-epoch classification basis, the method presented in this study focuses on identifying the precise temporal location of transitions between different sleep stages, to reducing sleep stage ambiguity. Currently, EEG signals (as shown by the blue line in Fig. 1) are used to determine sleep stages, however, EEG signals are inherently chaotic, leading to significant variability in visual scoring. This variability often results in disagreements between expert scorers, spanning several epochs, regarding the exact point at which one sleep stage ends and another begins (i.e., sleep stage transitions). This lack of consensus lowers inter-scorer reliability (ISR).

Our proposed methodology leverages time-frequency ridges (TFR) (solid white line in Fig. 1) to more accurately determine these transitions. With this approach, and considering Fig. 1, even untrained observers can clearly identify changes in the TFR, such as the transition from alpha rhythm; which represents wakefulness, to low-amplitude, mixed-frequency (LAMP) waveforms; commonly seen in sleep stage N1, around 26 s into the recording. This provides an output which is also interpretable to manual scorers.

This study introduces a novel method to precisely characterise the temporal location of sleep stage transitions, addressing a significant gap in the current gold standard for sleep stage scoring and also in current epoch-by-epoch classification approaches. Presently, no standardised methodology exists for accurately pinpointing the timing of these transitions, leading to imprecise scoring and contributing to low ISR, particularly around transition points. To improve the accuracy of sleep staging, methodologies like the one proposed here should be developed and integrated into existing guidelines and automated approaches.

To meet these aims; this study proposes to (1) Characterise the profile of sleep stage transitions through the use of time-frequency ridge analysis (2) Ensure characterisation is generalised and representative of a population by using multiple independent datasets (3) Propose this characterisation for further work, such as clinical and artificial intelligence approaches for determining sleep stages and improving ISR.

2. Dataset

Clinically annotated PSG data were sourced from two independent, large scale, sleep databases; The 2018 PhysioNet challenge “You Snooze, You Win” and the Wisconsin Sleep Cohort (WSC) (Ghassemi et al., 2018; Goldberger et al., 2000; Young et al., 2009). The use of multiple, independent databases ensured population based validity, considering the statistical principle of large numbers. Both databases contained whole night recordings, approximately eight hours each. Annotations were generated by sleep experts, using the AASM’s guidelines, with labels for sleep stage and arousal events scored in 30 s epochs and instantaneously, respectively.

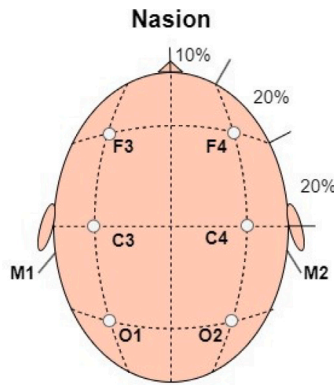


Fig. 2. Electrode placement as defined by the International 10–20 system, note ‘M’ electrodes are placed on the respective mastoid.

Table 2

Summary Table of the PhysioNet challenge “You Snooze, You Win” and WSC datasets.

	PhysioNet n (% of population)	WSC
Total Subjects	994 (100)	2570 (100)
Male subjects	666 (67)	1385 (54)
Female subjects	328 (33)	1185 (46)
	Mean	
Age (years)	55	60
BMI	33	28
AHI	19	12
Recording time (h)	7.7	8.1
Time in bed (h)	7.5	–
Sleep time (h)	6.2	6.1

2.1. 2018 PhysioNet challenge “You Snooze, You Win”

The 2018 PhysioNet challenge “You Snooze, You Win” database consisted of 994 clinically annotated overnight polysomnographic recordings. Each recording consisted of 13 channels of physiological measurements; including six electroencephalography (EEG) channels based on electrode pairs, F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, and O2-M1; one electrooculography (EOG) channel based on electrode pair E1-M2; three electromyography (EMG) channels of abdominal, chin and chest movements; electrocardiology (ECG) based on lead II; respiratory airflow and oxygen saturation (SaO2) given as a percentage. Signals were sampled at 200 Hz, with the exception of SaO2, which was recorded at a lower frequency and resampled to 200 Hz as described in Ghassemi et al. (2018). Each sample point in the recording was matched by two clinical labels, the first label described the sleep stage — six categories were possible; wakefulness, stage N1, stage N2, stage N3, REM and Undefined scored in 30 s epochs. It should be noted that the undefined stage is typically used to indicate clinically irrelevant data, such as set up time. Undefined stages were included in this analysis to ensure this study was exhaustive of the data. The second label indicated sleep arousals, scored at the point of arousal. Demographically, the PhysioNet database contained a 666–328 (67%–33%), male–female ratio, with a median age of 56 and a range between 18 and 90 years old. The mean apnea–hypopnea index was 18.9 events per hour. The dataset is summarised in Table 2.

2.2. Wisconsin sleep cohort

The WSC was hosted in an online data repository courtesy of The National Sleep Research Resource (Zhang et al., 2018). WSC contained 2570 PSG clinically annotated overnight recordings conducted between 2000 and 2015. As the AASM scoring guidelines were significantly

Table 3

Exhaustive sleep stage transition categories.

	N1	N2	N3	R	U	W
N1	N1 → N1	N1 → N2	N1 → N3	N1 → R	N1 → U	N1 → W
N2	N2 → N1	N2 → N2	N2 → N3	N2 → R	N2 → U	N2 → W
N3	N3 → N1	N3 → N2	N3 → N3	N3 → R	N3 → U	N3 → W
R	R → N1	R → N2	R → N3	R → R	R → U	R → W
U	U → N1	U → N2	U → N3	U → R	U → U	U → W
W	W → N1	W → N2	W → N3	W → R	W → U	W → W

revised in 2007, recordings made under pre-2007 guidelines were excluded from this analysis. This reduced the number of WSC recordings to 742. 16 Channels were recorded for the majority of recordings. Eight recordings contained only 14 recorded channels. The eight 14 channel recordings were included for analysis, as EEG channel O1-M2 was present. The 16 channels of physiological measurements included; three EEG channels F3-M2, C3-M2 and O1-M2; two EOG channels E1 and E2; two electromyography (EMG) channels, left chin and right leg movement; (ECG) based on lead II; one audio channel for snoring; respiratory airflow; nasal pressure; three plethysmography channels, thorax, abdomen and summation; a positional channel and SaO2, given as a percentage. All signals were sampled at 200 Hz. Demographically, the WSC database contained a 390–333 (54%–46%), male–female ratio, with a median age of 64.6 and a range between 44 and 85 years old. The median apnea–hypopnea index was 6.3 with a range between 0 and 87.8.

With the individual databases considered, they were standardised into the PhysioNet format before processing. 1736 PSG studies were included with a total of 174,786 sixty second sleep stage transitions (PhysioNet - 82,554 & WSC - 92,232). However, it should be noted that the distribution of sleep stage transitions were not uniform detailed later in Section 4.

3. Methodology

A graphical overview of the processing pipeline is shown in Fig. 3.

Briefly; clinical annotations were initially assessed, with any difference between successive sleep stage annotations regarded as a transition. Sixty seconds of data centred on each transition point was then extracted. This process was repeated for all transitions in each patient recording. The raw EEG transition signals were processed via the continuous wavelet transform (CWT), producing a two-dimensional time–frequency matrix (scalogram). Time–frequency ridges were computed from the scalogram, with each transition undergoing this time–frequency ridge extraction process. The ridges were then segregated, based on sleep stage transition categories, with 23 out of 36 possible categories populated, due to the non-uniform probability distribution of sleep stage transitions. To ensure signal integrity, excessive bandwidth was applied during CWT calculation, allowing some artefacts into the analysis. An outlier detection method was subsequently employed to exclude transition data that contained extreme frequency outliers. The remaining data underwent final analysis, excluding artefacts, allowing characterisation of the time–frequency ridges per sleep stage transition. An in-depth explanation of each step is as follows:

3.1. Pre-processing & lead selection

Pre-processing is a common initial step within Digital Signal Processing work, to ensure signals are free from artefact before analysis. However, as the time–frequency ridge response is the primary analysis method for this study, no pre-processing was implemented to prevent data skewing or temporal shifting (de Cheveigné & Nelken, 2019). Further, the CWT methodology implemented, ensured a significantly large bandwidth to avoid influencing the output signal via fidelity loss (Luo & Johnston, 2010). This ensured that detected characteristics

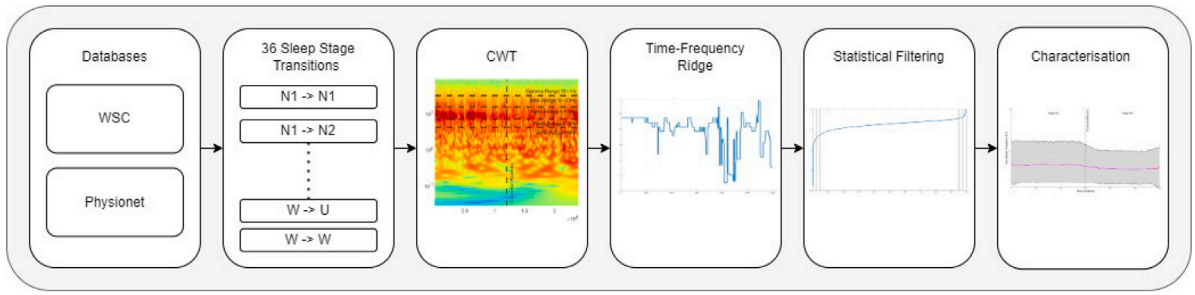


Fig. 3. An overview of the processing pipeline: two independent databases were sought; 60 s of EEG data were extracted for each sleep stage transition, 30 s before and after the point of transition; a CWT was computed; time–frequency ridge extracted from CWT; outlier detection applied to exclude data with artefacts; final characterisation performed.

were not the result of pre-processing filters. Signals were therefore analysed directly from the raw recordings. Statistical methods discussed in Section 3.5 were later used to exclude data with obvious artefacts, such as muscle artefacts and DC wandering. For analysis, EEG channel O1-M2 was selected, as this lead is one of the three recommended EEG leads in the AASM gold standard (Berry et al., 2020). Electrode placement is shown in Fig. 2. Limitations based on posterior dominant rhythm are noted in Section 4.5, however this channel was present in both databases and as this study is novel and exploratory, this limitation was acceptable.

3.2. Sleep stage transitions

Sleep stages were scored in 30 s windows (epochs), as recommended (Berry et al., 2020). Each epoch falls into one of six clinical labels; Wakefulness (W), N1, N2, N3, REM (R) and Undefined (U). With the transition from one sleep stage to another being of interest for this article, 36 unique transitions categories were possible, detailed in Table 3. However, six of these transitions did not contain a change of sleep stage, i.e. N1 → N1; these six categories were therefore excluded. To begin analysis, sleep stage transitions were identified in the clinical annotations, i.e. where there was a different label for successive epochs. If this condition was true, the corresponding EEG data were extracted. First and second epoch represent 30 s before and after the label change. This results in a total of 60 s of data for each extraction, with all sleep stage transitions extracted in the same manner.

3.3. Continuous wavelet transform

EEG signals are non-stationary, therefore, must be decomposed with respect to this property. In time–frequency decomposition, short duration windows are utilised to compute the power spectrum of a signal. The windowing effect results in an approximately stationary signal. The CWT was selected for this work due to the variable time–frequency resolution, which offered improved granularity when compared to traditional approaches, such as the short-time Fourier transform.

The CWT is defined as the convolution between a signal of interest, $x(t)$ and a wavelet, ψ as a measure of similarity. The transform produces a series of coefficients, computed by continuous scaling and shifting of the wavelet, with the signal to be analysed. A Morse wavelet was selected, as Morse wavelets are commonly used to analyse signals with varying amplitude and frequency, due to their complementary properties (Cherian & Kanaga, 2022; Martinez-Ríos et al., 2023a, 2023b; Wachowiak et al., 2018). The Fourier transform of the generalised Morse wavelet (ψ) is defined as (Olhede & Walden, 2002);

$$\psi_{P,\gamma}(\omega) = U(\omega) a_{P,\gamma} \omega^{\frac{P^2}{\gamma}} e^{-\omega^\gamma} \quad (1)$$

Where ω is the angular frequency, $U(\omega)$ is the unit step, $a_{P,\gamma}$ is a normalising constant, P^2 is the time-bandwidth product, and γ characterises the symmetry of the Morse wavelet. The Morse wavelet used had the following parameters selected; a symmetry parameter

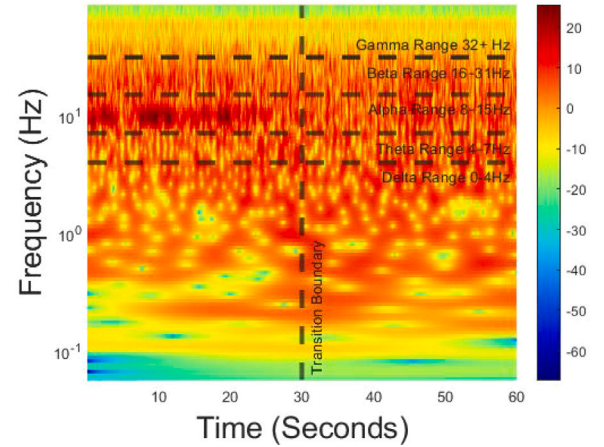


Fig. 4. An example of CWT performed on electrode O1-M2 for sleep stages W→N1, note the clear alpha rhythm representative of wakefulness visible as high energy (dark red) within the alpha band (approximately 10 Hz) before the transition boundary.

(γ) equal to 3 and time bandwidth product (P^2) equal to 60. These parameters were chosen to optimise the balance between time and frequency resolution, ensuring that the CWT captures the key features of sleep-stage transitions outlined in Table 1, further ℓ_1 normalisation was applied as defined in Eq. (2) (Boyd, 1974).

Where v is a vector to be normalised, n is the number of elements in the vector and i is the index of the vector element being normalised.

$$\ell_1 \text{ norm } ||v||_1 = \sum_{i=1}^n |v_i| \quad (2)$$

With the wavelet selected, the CWT can be considered. Mathematically, the CWT is defined by the convolution integral (Daubechies, 1992; Grossmann & Morlet, 1984);

$$T(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi\left(\frac{t-b}{a}\right) dt \quad (3)$$

Where $x(t)$ is a one dimensional signal of interest; ψ is the selected mother wavelet; a is the scale factor; b is the shifted position in time t . 107 frequency bins were used with logarithmic spacing between approximately 0–90 Hz, this frequency range was in excess of the expected nominal EEG frequency bandwidth to ensure that data was not lost. A scalogram is the absolute value of the CWT signal. The scalogram representation of a time–frequency matrix can be computed as Salles and Ribeiro (2023);

$$E_{\psi, CWT}(x) = \sum_i |\hat{\phi}_i(x)|^2 \quad (4)$$

Where $E_{\psi, CWT}$ represents the sum of energies of wavelets, point-wise, x is the signal of interest, $\hat{\phi}$ is the dual function of ψ and i represents the discrete frequency bins. An example of the CWT scalogram output is shown in Figs. 4 and 5.

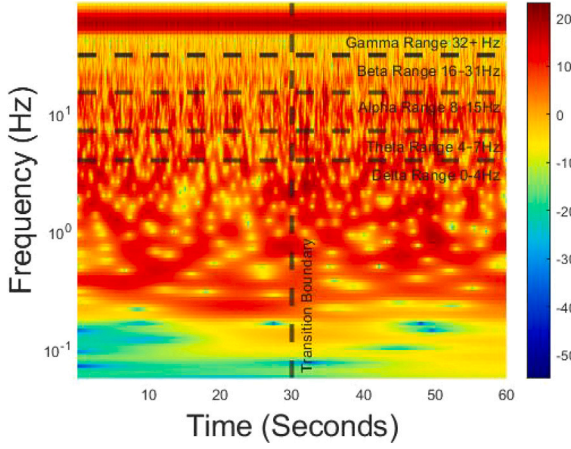


Fig. 5. An example of CWT performed on electrode O1-M2 for sleep stages N1→N2, note the high energy (dark red) EMG noise in the gamma frequency range. This obscures the underlining EEG information from the time–frequency ridge algorithm.

3.4. Time–frequency ridges (instantaneous frequency)

Time–frequency ridges can be described as a line traced across the CWT transform, which seeks to plot the maximum energy within the CWT with respect to time (Carmona et al., 1997). A secondary time–frequency ridge can also be defined by generating a plot of the second most energetic signal with respect to time and so forth.

Time–frequency ridges were extracted from the CWT scalograms. The extraction was based on a forward–backward greedy path optimisation algorithm, with penalised frequency jumps. The simplest implementation of the algorithm seeks to find an initial random point on the time–frequency ridge, and, sample by sample, work both forwards and backwards to find the next point in the time–frequency ridge. For an initial point, $\omega_p(t_{n0})$; mathematically this can be defined as (Iatsenko et al., 2016);

for $n = n_0 + 1, \dots, N$:

$$m_c(t_n) = \arg \max_m (F[t_n, Q_m(t_n), v_m(t_n), \omega_p(t_n - 1), \omega_p(t_n - 2), \dots, \omega_p(t_{n0})]) \quad (5)$$

$$\omega_p(t_n) = v_{m_c(t_n)}(t_n)$$

Where n is the discrete sample, N is the final discrete sample, $m_c(t_n)$ is the sequence of selected peak indices, at each discrete time step t_n , F is the chosen path function of discrete time t_n , the peak positions $v_m(t_n)$ and amplitudes $Q_m(t_n)$ at t_n . By selecting the maximum amplitude point in the time–frequency matrix, as the initial starting condition, (as it was likely that the dominate ridge included this point regardless) it is possible to simplify the algorithm to Iatsenko et al. (2016);

$$\begin{aligned} \omega_p(t_{n_0}) &= v_{m_0}(t_{n_0}), \\ \{m_0, n_0\} &= \arg \max_{\{m,n\}} (F_0[t_n, Q_m(t_n), v_m(t_n)]) \end{aligned} \quad (6)$$

Where F_0 denotes the maximum ridge point, and acts as a new zero in the forward–backward sequence.

The function is optimised as a continuous profile, rather than sample by sample to ensure that erroneous curves are not created. This is completed by selecting the ridge curve that maximises the integral of F with respect to time, given by the expanded form;

$$\begin{aligned} \{\omega_p(t_1), \dots, \omega_p(t_N)\} &= \{v_{m_c(t_1)}(t_1), \dots, v_{m_c(t_N)}(t_N)\}, \\ \{m_c(t_1), \dots, m_c(t_N)\} &= \arg \max_{\{m_1, m_2, \dots, m_N\}} \sum_{n=1}^N F[t_n, Q_{m_n}(t_n), v_{m_n}(t_n), \{\omega_{m_1}(t_1), \dots, \omega_{m_N}(t_N)\}] \end{aligned} \quad (7)$$

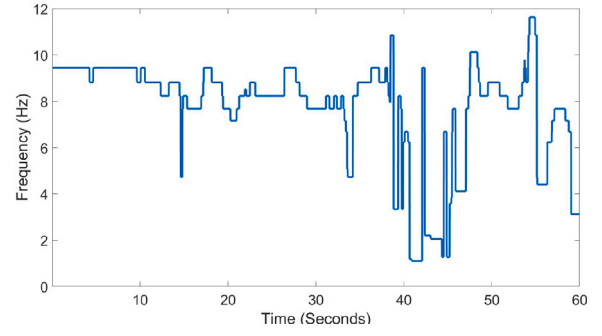


Fig. 6. An example of an extracted time–frequency ridge of sleep stage transition W→N1, note the decreased frequency beyond 30 s, this reduction is indicative of the transition to N1. A second increase at 48 s is also seen as this lasts approximately ten seconds before decreasing, which is less than 50% of the 30 s epoch it is correct to label this transition as W→N1.

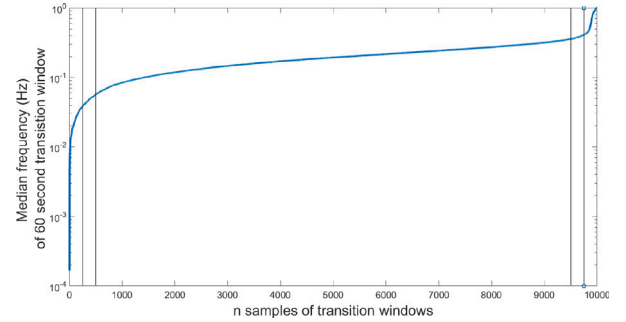


Fig. 7. The median value of each of the 10,000 N2→N3 transitions, high and low frequency artefacts are clearly present. Vertical lines have been added at $n = 250, 500, 9500$ and 9750 .

Finally, the penalisation function should be considered; this function uses a hidden Markov model to suppress frequency jumps between secondary ridges, by constraining the time–frequency ridges to local maxima, unless the penalisation function is overcome, allowing global maxima points to be reached within the CWT matrix.

$$F[\dots] = \log Q_m(t_n) + \omega(v_m(t_n) - \omega_p(t_n - 1), \alpha) \quad (8)$$

Where $F[\dots]$ denotes the time–frequency ridge, $\omega(v_m(t_n) - \omega_p(t_n - 1), \alpha)$ is a weight function, α is the adjustable parameters used by the weight function. Further detailed information on the process can be found in Carmona et al. (1997), Iatsenko et al. (2016). Implementations in both Python and MATLAB are readily available. Initially, five time–frequency ridges were explored, one ridge for each known wave frequency range; Delta, Theta, Alpha, Beta, Gamma and Mu shown in Fig. 4. However, this was later reduced to one dominant time–frequency ridge, across all brain wave patterns, to determine the dominant ridge of the signal. An example of an extracted time–frequency ridge is shown in Fig. 6.

3.5. Outlier detection

As raw data were used for analysis, it was important to remove outlying data, without introducing distortion from typical filtering methods. To do so successfully, data exclusion of outliers was the preferred method. Caution should be taken when using this approach, as it is incorrect to remove data to improve performance metrics. However, as performance metrics were beyond the aim of this study and instead generation of representative time–frequency characteristics were of importance, data exclusion was appropriate.

The Antonyan Vardan Transform (AVT) was implemented as a standardised approach to data exclusion (Cene & Balbinot, 2018). Several steps are required to preform AVT;

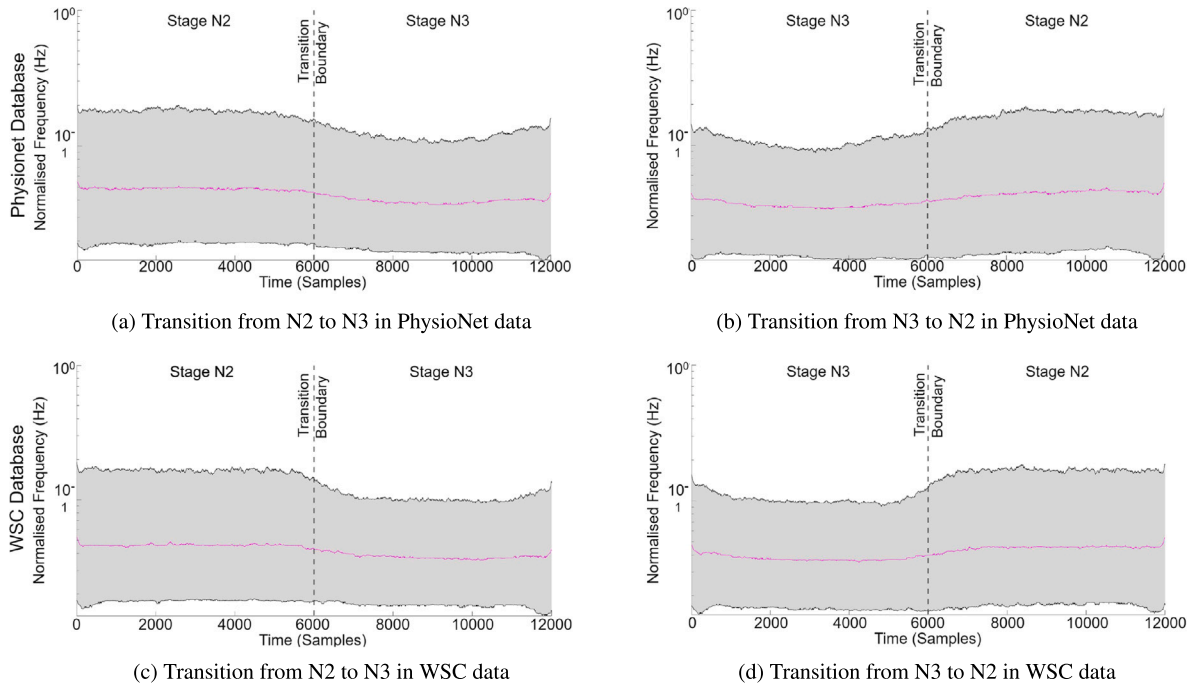


Fig. 8. A comparison of characterisation frequency curves of sleep stage transition N2→N3 and N3→N2 across both databases, median marked in magenta with 95% confidence intervals marked in grey.

1. Collection of data
2. Calculate standard deviation and median of data
3. Exclude data which falls outside of the median $\pm$ one standard deviation
4. Calculate final results from trimmed data.

Considering the sleep data, statistical outliers were identified by determining the median frequency of each transition window. Median values were less prone to skewing from a small number of erroneous data points. The median frequency values were plotted in ascending order in Fig. 7, three obvious groups emerged: approximately linear data between sample point 500 and 9750; high frequency data at sample point 9750 and greater, attributed to EMG artefact noise; and low frequency artefact data at sample points 500 and lower.

Time-frequency ridges of each individual sleep stage transition that met the criteria for exclusion were identified, and the associated time-frequency plot was removed from further analysis, as this data was deemed to include artefacts.

3.6. Evaluation

For evaluation, Pearson correlation coefficient and the Mean Squared Error were calculated in the Results section. These equations were defined as;

Pearson correlation coefficient (r);

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (9)$$

Where x_i and y_i are individual sample points at index i and $\bar{x}$ and $\bar{y}$ are the mean sample values respectively.

Mean Squared Error (MSE);

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (10)$$

where n is the number of data points, Y is the first dataset and transition input and $\hat{Y}$ is the second dataset and transition input. Thus determining the MSE between the time-frequency ridges. Further details are found in the Results section.

Table 4

Mean squared error (MSE) and Pearson correlation coefficient (r) for the selected transitions in Fig. 8 and Fig. 9. Pearson correlation coefficient p -value (p) was omitted as the large sample size and low frequency information lead to artificial decrease in p .

Dataset 1	Dataset 2	Transition 1	Transition 2	MSE	r
PhysioNet	PhysioNet	N2→N3	N3→N2	2.42E−05	−0.74
WSC	WSC	N2→N3	N3→N2	3.53E−06	−0.91
WSC	PhysioNet	N2→N3	N2→N3	1.77E−05	0.93
WSC	PhysioNet	N3→N2	N3→N2	3.09E−06	0.92

4. Results and discussion

4.1. Characterising sleep stage transition frequency

From the remaining distilled data post-exclusion processing, the median and 95% confidence interval were calculated for all transitions. Due to the statistical principle of large numbers (Yao & Gao, 2016); a large enough sample size should allow for a population based transition characterisation. A sample size of 10,000 samples was chosen empirically. To ensure this was an appropriate sample size, a second independent database was analysed to determine that the characteristic patterns identified were consistent across independent data sources.

In-depth, graphical analysis and discussion was performed on two sleep stage transitions to improve understanding. Presentation of all sleep stage transitions was outside of the scope of this article and instead is summarised in Tables 5 and 7. Discussion of results centres around transition N2→N3 and N3→N2 as these transitions;

1. exist in both datasets
2. were prevalent
3. feature only moderate ISR (Penzel et al., 2013; Rosenberg & Hout, 2013).

It should be noted that all available transitions were included. However, transitions which feature lower prevalence i.e., $n \leq 10,000$, as shown in Tables 5 and 7, are unlikely to be representative of a generalised population. As all data were considered up to this point,

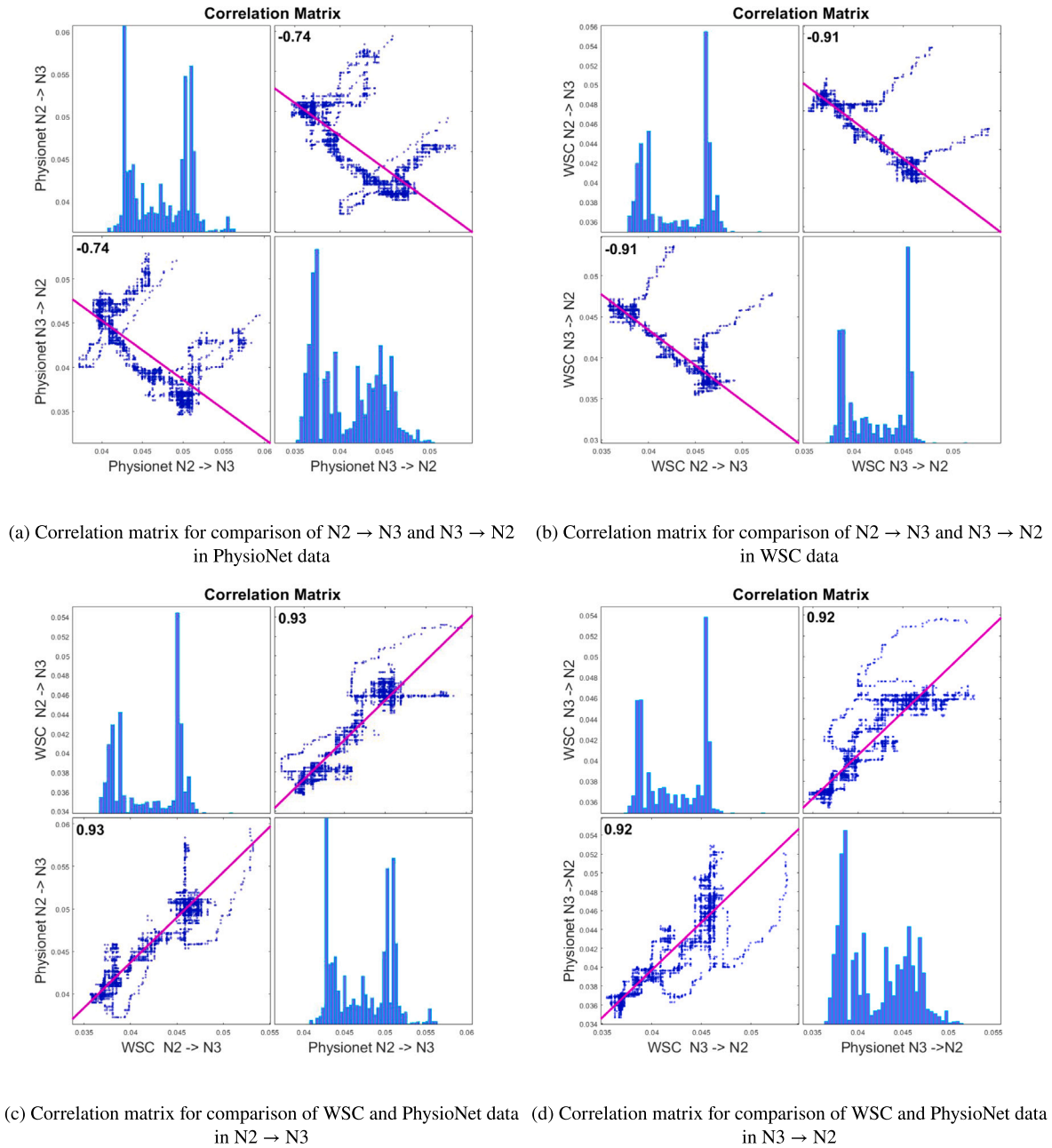


Fig. 9. Correlation matrices of the median time-frequency ridges.

it was not possible to increase ‘n’ for all transitions, without additional data collection, which fell outside of the scope of this article. To fully consider the visual data within the four graphs of Fig. 8, additional metrics were calculated, included in Fig. 9 and Table 4, which are included in discussions. Contrast and comparison was drawn between the two databases and the two sleep stage transitions. This results in a total of four assessments:

Comparison of Fig. 8, allows for visual analysis of the time-frequency patterns within the data. Several patterns are clearly visible in the data.

Assessment One — Reciprocal Transitions: Comparing Fig. 8(a) with Fig. 8(c) and Fig. 8(b) with Fig. 8(d) shows that both databases produced sleep stage transitions within the same approximate confidence interval range and median. This suggests that data saturation for these populations was reached, as no new themes emerged from a second large, independent source. Considering the correlation matrix

within Fig. 9(c) and 9(d), a correlation of +0.93 and +0.92 was calculated; this showed a high correlation between the median of the two independent databases lending further support to the fact that the sleep stage transitions in both databases were statistically similar. A mean squared error value of $1.77\text{E}-05$ and $3.09\text{E}-06$ was calculated in Table 4 for the data, showing little difference between medians. One interesting noted difference between the two databases, was the gradient at the point of transitions. This gradient was steeper in the WSC data than the PhysioNet data. A steeper gradient is indicative of a more sudden sleep stage transition within the underlining EEG data. Several physiological reasons could be responsible for this, such as age. However, a steeper gradient could also indicate higher ISR in the WSC database, as a 100% sleep stage ISR in the time domain would tend towards a ‘step’ response. Additional studies are required to fully determine the reasons behind this.

Assessment Two — Generalising to Alternative Data Sets: By comparing Fig. 8(a) with Fig. 8(b) and Fig. 8(c) with Fig. 8(d), it

is possible to see generalised trends in the transitions N2→N3 and N3→N2 across datasets. Consistency in both pairs of transitions, in terms of general frequency ridge shape and confidence interval range, was observed. It was also evident that both reciprocal transitions start and end at approximately the same frequency, within the considered 60 s window. This is another indicator that a generalised approach has yielded saturation. Considering the correlation matrices in Fig. 9, correlation values of -0.74 and -0.91 were computed. The rotated 'U' shape within the correlation matrix is generated due to the transition within the data, which was expected, considering the input data, it was argued that Simpson's paradox applied (Blyth, 1972). As the effect was consistent, it was expected to cancel out. A detailed analysis of Simpson's paradox with respect to this data fell outside of the scope of this article. It should be noted that the transition gradient is consistent among likewise databases, but dissimilar between databases.

With the transitions compared, they should also be assessed independently of one another.

Assessment Three — N2→N3 Transition Analysis: Considering transition N2→N3 of the PhysioNet dataset, and cross referencing with Table 5, the median frequency decreased from 4.32 Hz to 3.28 Hz; a reduction of 1.04 Hz, or approximately 25%, when comparing a one second average of the time–frequency ridge 15 s before the point of transition, to a one second average of the time–frequency ridge 15 s after the point of transition. Similarly, a frequency reduction of 10.4% and 14.4% is observed for the minimum and maximum confidence interval, respectively. The range of the confidence interval also reduced by 8 Hz, or 14%, over the course of the transition. Similarly, considering the same transition (N2→N3) of the WSC dataset, it was observed that median frequency reduced from 4.07 Hz to 3.27 Hz, a 19.6% reduction. Frequency reductions in the minimum and maximum confidence intervals were also seen, reducing by 2.2% and 18.6% respectively. An overall confidence interval reduction of 7.0 Hz, or 18.8% was also noted over the course of the transition.

Assessment Four — N3→N2 Transition Analysis: Moving to transition N3→N2 of the PhysioNet data, the median frequency ridge increased from 3.06 Hz to 3.90 Hz, an increase of 0.85 Hz, or 27.7%. The bandwidth increased, with minimum and maximum confidence intervals changing by -9.8% and 19.7% , respectively. This increased the overall bandwidth by 9.3 Hz, or 19.9%, over the course of the transition. Finally, considering transition N3→N2 of the WSC data, the median frequency ridge also increased by 0.78 Hz from 3.26 Hz to 4.03 Hz, a 23.8% increase. The bandwidth increased, with minimum and maximum confidence intervals changing by -6.7% and 32.1% respectively, a similar pattern to the PhysioNet data. Again this increased overall bandwidth by 9.55 Hz, or 32.8%, over the course of the transition.

Considering the results above and further information in Tables 5 and 7, it is clear that frequency ridge's can be used to characterise sleep stage transitions, with respect to large diverse populations. It is expected that these features could offer several benefits for sleep experts and automated approaches, as further discussed in Future work.

4.2. Proof of concept: Automated statistical comparison of transitioning epochs vs non-transitioning epochs

With the transitions characterised, a subset of the data was used as a proof of concept, to further validate the methodology. Transitional epochs N2→N3 and N3→N2, and non-transitional epochs N2→N2 and N3→N3 were selected for consistency. A total of 10,000 samples were chosen, comprising of 2500 transitional epochs of each label. A proof of concept algorithm was developed utilising simple, lightweight statistical features. The features were: mean, median, variance, kurtosis, standard deviation, skewness, gradient, minimum and maximum. Features were calculated on the time–frequency ridges, the associated amplitude was also considered. From the 60 s windows of transition

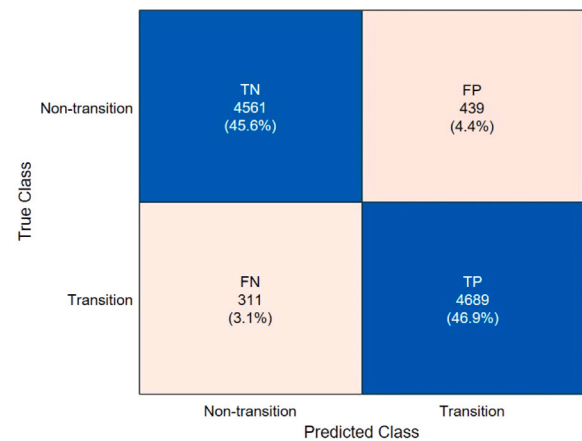


Fig. 10. A confusion matrix showing binary classification results for the proof of concept transition vs non-transition detection system.

data outlined previously, thirty seconds of data were extracted centrally, (in three ten seconds sub-epochs) from the middle time point of the data, as this was most likely to capture the transitional gradient changes shown earlier in Fig. 8. All features were normalised between the standard bounds of 0 and 1. Finally, a Support Vector Machine (SVM) algorithm was used for binary classification of transition/non-transition epochs. The SVM model was selected as SVM's generalise well to large feature sets, which have not undergone feature selection (Navot et al., 2006). The SVM model had a cubic kernel function, and a box constraint level of one. Five fold cross-validation was used, ensuring no data leakage. Model optimisation was not undertaken, this decision was made to ensure that overfitting did not occur, and as maximum performance was not an objective of this study. Results showing the average of five fold cross-validation, rounded to the nearest whole integer, are shown in Fig. 10, accuracy, sensitivity, specificity and F1 score were further calculated as 92.5%, 93.8%, 91.2% and 92.6% respectively. The training costs associated with the SVM classifier are detailed in Table 6. The model exhibits a favourable balance between complexity and efficiency, as indicated by its 3218 support vectors and a throughput of 4900 observations per second. The training time of approximately 96 s reflects the model's efficiency in processing data. Additionally, the relatively small model size of 6 MB underscores the suitability of this classifier for deployment in resource-constrained environments.

4.3. Implications

Considering the results section and the hypothesis; can sleep stages be delineated by changes in time–frequency ridges? It is clear that sleep stage delineation can be determined from PSG data via TFR's. This is significant, as identification of the point of delineation between two sleep stages causes the highest disagreement between sleep experts (de Chazal et al., 2022; Magalang et al., 2013; Nikkonen et al., 2023; Penzel et al., 2013; Rosenberg & Hout, 2013). Typically, automated algorithms make no attempt to correct for the exact point of delineation between sleep stages, nor is there clinical guidance in the current gold standards to standardise the exact epoch of sleep stage transition (Berry et al., 2020). Therefore, this novel methodology could be applied to both automated and non-automated clinical systems (as shown in Section 4.2) to improve sleep stage scoring standardisation, reducing disagreement and improving inter-scorer reliability. A second major benefit of this approach, is the interpretable of the methodology. Considering Fig. 1 and Fig. 6, even a casual observer can identify a significant change in the time–frequency ridge, in addition, sleep experts are already familiar with raw time-domain signals being presented in a similar fashion, removing the need for extensive retraining. This allows trust to be built for both sleep experts and patients.

Table 5

A table to show the median and 95% confidence interval values for all available transitions in the PhysioNet dataset. A one second average of the characteristic time–frequency ridge for three points, the point of transition, 15 s before and 15 s after.

Transition	n	One second signal mean with respect to the point of transition								
		−15 s			0 s			+15 s		
		Median	CI		Median	CI		Median	CI	
N1→N2	10000	13.71134	0.16093	53.31273	12.39668	0.200086	53.18684	10.07624	0.200086	53.05424
N1→N3	21	9.303783	0.372311	43.75476	3.645553	0.385363	30.74149	3.256727	0.222819	36.78765
N1→R	2924	14.6758	0.144293	56.94832	12.84216	0.172778	56.9555	11.18507	0.164888	56.94832
N1→W	10000	15.16731	0.138256	53.23182	12.07227	0.116422	53.10845	14.95386	0.127308	53.10845
N2→N1	10000	5.184639	0.16423	49.57976	3.680442	0.125516	52.18017	8.514339	0.120808	52.90038
N2→N3	10000	4.327099	0.256834	56.7809	3.992286	0.256834	52.81425	3.286368	0.230101	48.70022
N2→R	3523	9.982531	0.246459	60.71723	10.83572	0.260698	60.94638	11.05735	0.229512	61.08417
N2→W	6426	3.905712	0.142421	49.77516	1.229229	0.096749	46.40894	1.244773	0.088206	49.82122
N3→N1	443	1.980472	0.223394	44.99683	1.745898	0.138659	48.95282	2.564439	0.05594	47.19348
N3→N2	10000	3.058035	0.218135	47.06002	3.395749	0.181072	52.39248	3.90442	0.196664	56.34975
N3→R	100	3.236812	0.577786	43.32369	3.227174	0.117015	52.59751	7.012308	0.05594	54.06951
N3→W	587	1.926754	0.159592	39.84767	1.149226	0.082903	42.6335	1.2568	0.085981	49.01347
R→N1	3984	7.254185	0.140261	53.21116	4.717436	0.114096	53.15112	9.875143	0.139659	53.14079
R→N2	1079	12.32294	0.213966	61.11825	10.71228	0.25615	61.15651	10.16472	0.22117	61.10756
R→N3	22	6.112158	0.073997	65.52655	5.19848	0.472293	49.67289	5.247668	0.14871	44.72665
R→W	1537	3.242386	0.096166	53.31028	0.96484	0.087448	50.85115	1.120259	0.113415	49.21661
U→N1	75	11.49704	0.068026	52.21682	17.04933	0.106279	57.47147	17.13103	0.176583	59.03497
U→R	2	9.004651	1.431398	16.5779	15.51234	1.30218	29.84306	16.13479	2.636434	29.84306
U→W	912	1.449035	0.093729	51.26038	2.77756	0.093726	49.63281	11.74819	0.11829	49.78932
W→N1	10000	18.67392	0.132758	53.2832	19.69306	0.11829	56.12081	18.18172	0.162332	56.83348
W→N2	753	15.11455	0.114445	53.33699	11.92974	0.05594	53.15179	8.924729	0.20321	52.99229
W→N3	8	6.581839	0.05594	26.7063	8.731256	0.494741	52.85886	8.080206	0.177787	63.29075
W→R	158	2.07681	0.109571	48.59383	14.49951	0.064986	78.79985	15.67658	0.137505	78.79659

Table 6

A table to show the associated training costs of the SVM classifier.

	Training cost
Support Vectors	3218
Throughput	4900 (obs/s)
Training Time	96.35 (s)
Model Size	~6 MB

4.4. Considerations

One point of consideration is the PhysioNet database. This database contains individuals who are undergoing an overnight PSG to assess for sleep disorders. As such, the prevalence for sleep disorders is higher in this group. Fragmented sleep patterns could cause skewing in the data (Kaur et al., 2019; Thabtah et al., 2020), however, it is not expected to cause a significant effect. Further, sleep disordered breathing conditions are highly prevalent in the population as noted in Heinzer et al. (2015). As such, the inclusion of this data mimics a population sample required for a generalised characteristic model. To ensure that significant skewing was not introduced, the WSC data was also used. The WSC was a longitudinal study, to determine cause, consequence and medical history associated with sleep disorders. A questionnaire was provided to Employees of the State of Wisconsin to determine those at high and low risk of sleep disordered breathing. Candidates were selected in pairs, i.e. for one high risk patient, a low risk patient of the same age and sex was included. This created a more even split in terms of participants, however, may not be fully reflective of the generalised population.

4.5. Limitations

Several limitations have been identified in this work and are discussed below.

4.5.1. Antero-posterior EEG spectral power gradient

It is known that different brain rhythms are more pronounced at certain electrodes (McCallan et al., 2021). In this study, electrode O1-M2 was selected for analysis. Alpha rhythm is known to be more

pronounced than other frequencies when considering this electrode, due to the proximity of the recording electrodes to the brain stem. As the study focuses on developing and proposing a useful technique this limitation is acceptable. The author notes that if this methodology was applied clinically, observation of additional electrodes in tandem could improve results, as more information is available for a decision making process.

4.5.2. Cone of influence

The continuous wavelet transform divides signal energy into non-equal time and frequency bins. This is an improvement over equal time and frequency bin methods, such as the Fourier transform, as energy localisation improves the temporal information resolution. One downside to this method is the so called *cone of influence* (COI), this is an area within the wavelet power spectra which is distorted, due to the influence of the end points of the finite length signals. It is possible to see effect of the COI in Fig. 8 as a slight upturn at the ends of each signal plot. To negate this phenomenon edge values were not considered for analysis.

4.6. Future work

4.6.1. Application

With the time–frequency characteristic ridge generated, clinical applications were considered. It is expected that time–frequency ridges could be displayed as an alternate, or additional channel for sleep experts to offer improved ISR. Similarly, it is expected that applying this approach to automated sleep staging, via AI algorithms, could also improve results. Preliminary work to address the question of AI algorithms for sleep stage transition detection has already showed promise (McCausland et al., 2023). Additional studies to apply this methodology to these applications are a logical next step which fell outside the scope of this manuscript.

4.6.2. Analysis of expert scorer transition vs time–frequency ridge transition point

As mentioned in Section 4.6.1, automation of sleep stage transition detection is a clear path forward, however, assessment of the temporal difference in sleep stage transition points between this methodology

Table 7

A table to show the median and 95% confidence interval values for all available transitions in the WSC dataset. A one second average of the characteristic time–frequency ridge for three points, the point of transition, 15 s before and 15 s after.

Transitions	n	One second signal mean with respect to the point of transition								
		–15 s			0 s			+15 s		
		Median	CI		Median	CI		Median	CI	
N1→N2	10 000	14.08915	0.325142	49.14065	13.24937	0.3863	48.99569	10.58283	0.412273	46.18959
N1→N3	20	7.346456	0.178425	32.12625	10.32437	0.133744	53.10717	3.391312	0.0559	26.30009
N1→R	3418	10.83688	0.233823	49.03864	12.72014	0.32625	49.33325	10.34454	0.388121	49.1672
N1→U	3	22.36535	0.503137	29.04878	20.78538	11.20415	28.41691	29.90354	11.41546	36.35267
N1→W	6916	12.94627	0.265165	48.92971	9.149463	0.199536	46.07168	11.94939	0.242303	46.29708
N2→N1	10 000	8.774817	0.320934	46.17637	7.285873	0.27565	48.99569	12.94627	0.274624	49.14065
N2→N3	10 000	4.067891	0.5363	37.66854	3.792336	0.506058	35.99832	3.269465	0.524135	30.66043
N2→R	3553	9.855054	0.495954	49.3823	10.56878	0.517905	52.46669	10.83569	0.44928	49.47899
N2→U	10	10.50185	2.159218	30.13238	4.637407	1.108805	25.90475	12.3786	0.084097	49.32243
N2→W	10 000	5.448755	0.264424	45.95848	1.83338	0.200045	40.32643	1.637389	0.181205	43.1907
N3→N1	56	1.969265	0.262346	38.51389	1.584422	0.158234	40.03838	10.2359	0.272459	42.9598
N3→N2	10 000	3.256688	0.503019	29.6332	3.505708	0.477746	36.24364	4.033035	0.468823	39.14485
N3→R	25	2.964044	0.200045	36.45397	3.540626	0.0559	86.8249	17.93579	0.653806	86.8249
N3→U	3	1.392978	1.088135	7.622238	1.141383	0.916061	8.075159	5.351074	1.615773	13.06552
N3→W	574	2.9492	0.494138	30.61904	1.962188	0.271924	30.58177	1.654293	0.163581	37.14167
R→N1	3322	7.175645	0.278006	45.95848	3.660898	0.206993	43.12828	9.039579	0.221021	46.11349
R→N2	1229	12.64434	0.547779	52.90037	12.24372	0.528077	55.96966	10.20661	0.466049	49.26257
R→N3	1	15.56478	15.56478	15.56478	4.370163	4.370163	4.370163	16.98756	16.98756	16.98756
R→U	3	0.743143	0.511339	1.24495	0.822564	0.312278	1.24495	15.65276	1.754	26.47195
R→W	4019	5.11231	0.190292	46.16767	1.391318	0.159818	40.23091	1.533623	0.196845	42.22713
U→N1	14	1.811668	0.512245	25.72618	10.63916	0.322988	32.71577	6.101099	0.326278	32.02858
U→N3	4	4.028383	2.599663	24.17901	2.904641	1.067049	10.49237	6.780268	4.038864	52.03389
U→R	2	5.154481	4.13664	6.172322	8.510017	1.601648	9.257171	3.071215	1.404921	4.897479
U→W	738	0.712518	0.134552	37.57689	0.770364	0.127572	37.50964	0.937199	0.1463	40.35816
W→N1	10 000	15.09433	0.221755	48.89423	15.94178	0.188765	49.21659	14.0173	0.329122	46.29708
W→N2	6620	7.525396	0.191766	48.92971	10.55982	0.172737	49.43293	8.634106	0.368881	47.71391
W→N3	17	2.964044	0.109659	34.52961	2.37484	0.633369	42.00222	4.751807	0.401842	58.73498
W→R	1669	7.003584	0.18854	48.55735	12.69605	0.183928	49.3452	12.58885	0.388499	49.73185
W→U	16	20.675	0.221637	74.34264	8.065376	0.073703	36.4989	3.993686	0.064132	39.11275

and a sleep expert could offer the possibility of further sleep scoring recommendations to improve ISR. To ensure the proposed work captures generalised issues with the current rules, and not a individual scorer, a multi-scorer database will be necessary. The output of such work would help to inform best practice for both manual scoring rules and automated principals.

4.6.3. Non-transitioning epochs

This study focused solely on characterising the profile of sleep stage transitions, due to the impact this could have on ISR agreement, therefore, non-transitioning epochs were excluded. Future work should also assess the performance of non-transitioning epochs for use as a sleep staging visualisation tool.

4.6.4. Time–frequency characteristics for all transitions

Time–frequency characteristics were generated for 23 and 29 sleep stage transitions for the PhysioNet and WSC data, respectively. Ten thousand time–frequency ridge samples were determined to provide suitable variability, as no new themes emerged beyond this number of samples. However, only six transitions met this condition. Future work should gather more data to allow all 30 sleep stage transition time–frequency characteristic ridges to be generated.

4.6.5. Proof of concept model

A proof of concept model was developed, to validate the methodology. The SVM model preformed well on the binary classification task, while an SVM model was used, several other models showed similar performance. Future work will seek to optimise this model in regard to model type, initial parameters and feature selection, to maximise algorithm performance.

4.6.6. Time–frequency ridge calculations

Time–frequency ridge (TFR) computations are an emerging area of interest. As alternative ridge optimisation strategies continue to be developed and published, future work should explore enhancing the results by optimising TFR generation, particularly with respect to the frequency bands highlighted in Table 1. Additionally, incorporating alternative two-dimensional input transforms could further improve the methodology. Unfortunately, this level of optimisation falls outside the scope of the current article, but it presents a promising direction for future research. We anticipate that this work will inspire the development of similar methodologies. These should be directly compared to establish a new baseline for sleep stage transition delineation, moving beyond traditional epoch-by-epoch classification approaches.

5. Conclusions

In conclusion, characteristic time–frequency ridges were computed for sleep stage transitions. Using two independent and large scale databases ($n = 994$, $n = 742$; approximately 13,888 h of sleep data), consistent patterns were identified at the point of transition from one sleep stage to another. It is expected that these patterns are representative of a wider population given the statistical principles of the law of large numbers. Given the current unacceptable interscorer reliability for both manual and automated approaches, at determining the exact epoch of sleep stage transition, implementing an additional time–frequency ridge channel, or an automated approach, (such as the proof of concept in Section 4.2), alongside typical PSG channels, could offer a more deterministic approach for sleep experts.

CRedit authorship contribution statement

Christopher McCausland: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft.
Pardis Biglarbeigi: Conceptualization, Methodology, Data analysis,

Supervision. **Raymond Bond:** Methodology, Supervision. **Golnaz Yadollahikhales:** Conceptualization. **Alan Kennedy:** Conceptualization, Writing – review & editing. **Anna Sigridur Islind:** Writing – review & editing. **Erna Sif Arnardóttir:** Writing – review & editing. **Dewar Finlay:** Conceptualization, Methodology development, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Erna Sif Arnardóttir reports financial support was provided European Union's Horizon 2020 research and innovation programme under grant agreement no. 965417. Erna Sif Arnardóttir reports a relationship with Nox Medical that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with Jazz Pharmaceuticals Inc that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with Linde Healthcare that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with Wink Sleep that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with Apnimed Inc that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with Vistor hf that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with ResMed that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with Philips North America LLC that includes: consulting or advisory.

Data availability

The Wisconsin Sleep Cohort dataset (Young et al., 2009) is available from the National Sleep Research Resource (Zhang et al., 2018), by request. The 2018 PhysioNet challenge “You Snooze, You Win” dataset (Ghassemi et al., 2018) is open source, and available from PhysioNet (Goldberger et al., 2000).

Acknowledgements

The authors would like to thank the Northern Ireland Department for Economy for funding this work via a Ph.D. studentship. The authors would also like to thank the Reykjavík University Sleep Institute for their constructive feedback of this work.

The Wisconsin Sleep Cohort Study was supported by the U.S. National Institutes of Health, National Heart, Lung, and Blood Institute (R01HL62252), National Institute on Aging, United States (R01AG036838, R01AG058680), and the National Center for Research Resources, United States (1UL1RR025011). The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24 HL114473, 75N92019R002).

References

- Alsolai, H., Qureshi, S., Iqbal, S. M. Z., Vanichayobon, S., Henesey, L. E., Lindley, C., & Karrila, S. (2022). A systematic review of literature on automated sleep scoring. *IEEE Access*, 10, 79419–79443. <http://dx.doi.org/10.1109/ACCESS.2022.3194145>.
- Anafi, R. C., Kayser, M. S., & Raizen, D. M. (2019). Exploring phylogeny to find the function of sleep. *Nature Reviews Neuroscience*, 20, 109–116. <http://dx.doi.org/10.1038/s41583-018-0098-9>.
- Anderer, P., Ross, M., Cerny, A., Vasko, R., Shaw, E., & Fonseca, P. (2023). Overview of the hypnodelity approach to scoring sleep for polysomnography and home sleep testing. *Frontiers in Sleep*, 2, <http://dx.doi.org/10.3389/frsle.2023.1163477>.
- Arnardóttir, E. S., Islind, A. S., & Óskarsdóttir, M. (2021). The future of sleep measurements: a review and perspective. *Sleep Medicine Clinics*, 16(3), 447–464.
- Bakker, J. P., Ross, M., Cerny, A., Vasko, R., Shaw, E., Kuna, S., Magalang, U. J., Punjabi, N. M., & Anderer, P. (2023). Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: hypnodelity based on multiple expert scorers and auto-scoring. *Sleep*, 46, <http://dx.doi.org/10.1093/sleep/zsac154>.
- Berry, R., Quan, S., AR, A., Bibbs, M., DelRosso, L., Harding, S., Mao, M., Plante, D., Pressman, M., Troester, M., & Vaughn, B. (2020). The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications.
- Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338), 364–366. <http://dx.doi.org/10.1080/01621459.1972.10482387>.
- Boyd, D. W. (1974). The power method for lp norms. *Linear Algebra and its Applications*, 9, 95–101. [http://dx.doi.org/10.1016/0024-3795\(74\)90029-9](http://dx.doi.org/10.1016/0024-3795(74)90029-9), URL <https://www.sciencedirect.com/science/article/pii/0024379574900299>.
- Carmona, R., Hwang, W., & Torresani, B. (1997). Characterization of signals by the ridges of their wavelet transforms. *IEEE Transactions on Signal Processing*, 45(10), 2586–2590. <http://dx.doi.org/10.1109/78.640725>.
- Cene, V. H., & Balbinot, A. (2018). Using the sEMG signal representativity improvement towards upper-limb movement classification reliability. *Biomedical Signal Processing and Control*, 46, 182–191. <http://dx.doi.org/10.1016/j.bspc.2018.07.014>.
- Cherian, R., & Kanaga, E. G. (2022). Theoretical and methodological analysis of EEG based seizure detection and prediction: An exhaustive review. *Journal of Neuroscience Methods*, 369, Article 109483. <http://dx.doi.org/10.1016/j.jneumeth.2022.109483>, URL <https://www.sciencedirect.com/science/article/pii/S0165027022000103>.
- Choi, H.-I., Noh, G.-J., & Shin, H.-C. (2020). Measuring the depth of anesthesia using ordinal power spectral density of electroencephalogram. *IEEE Access*, 8, 50431–50438. <http://dx.doi.org/10.1109/ACCESS.2020.2980370>.
- Cook, J. (2021). Oximetry data affect the quality of gold standard polysomnography. *Journal of Clinical Sleep Medicine*, 17, 1329. <http://dx.doi.org/10.5664/JCSM.9216>, URL <https://jcs.m.aasm.org/doi/10.5664/jcs.m.9216>, <https://jcs.m.aasm.org/>.
- D'Ambrosio, C., Bowman, T., & Mohsenin, V. (1999). Quality of life in patients with obstructive sleep apnea: Effect of nasal continuous positive airway pressure - A prospective study. *Chest*, 115, 123–129. <http://dx.doi.org/10.1378/chest.115.1.123>.
- Daubechies, I. (1992). *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, <http://dx.doi.org/10.1137/1.9781611970104>, URL <https://epubs.siam.org/doi/abs/10.1137/1.9781611970104>, arXiv:<https://epubs.siam.org/doi/pdf/10.1137/1.9781611970104>.
- de Chazal, P., Mazzotti, D. R., & Cistulli, P. A. (2022). Automated sleep staging algorithms: have we reached the performance limit due to manual scoring? *Sleep*, 45, <http://dx.doi.org/10.1093/SLEEP/ZSAC159>, URL <https://academic.oup.com/sleep/article/45/9/zsac159/6648461>.
- de Cheveigné, A., & Nelken, I. (2019). Filters: When, why, and how (not) to use them. *Neuron*, 102, 280–293. <http://dx.doi.org/10.1016/j.neuron.2019.02.039>.
- Ferrario, A., & Loi, M. (2022). How explainability contributes to trust in AI. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 1457–1466). New York, NY, USA: Association for Computing Machinery, <http://dx.doi.org/10.1145/3531146.3533202>.
- Gaiduk, M., Alarcón, Á. S., Seepold, R., & Madrid, N. M. (2023). Current status and prospects of automatic sleep stages scoring: Review. *Biomedical Engineering Letters*, 13, 247–272. <http://dx.doi.org/10.1007/s13534-023-00299-3>.
- Galván, A. (2020). The Need for Sleep in the Adolescent Brain. *Trends in Cognitive Sciences*, 24(1), 79–89. <http://dx.doi.org/10.1016/j.tics.2019.11.002>.
- Geraedts, V. J., Boon, L. I., Marinus, J., Gouw, A. A., van Hilten, J. J., Stam, C. J., Tannemaat, M. R., & Contarino, M. F. (2018). Clinical correlates of quantitative EEG in parkinson disease. *Neurology*, 91, 871–883. <http://dx.doi.org/10.1212/WNL.0000000000006473>.
- Ghassemi, M. M., Moody, B. E., Lehman, L. W. H., Song, C., Li, Q., Sun, H., Mark, R. G., Westover, M. B., & Clifford, G. D. (2018). You snooze, you win: the PhysioNet/Computing in cardiology challenge 2018. In *2018 computing in cardiology conference (cinC)*: vol. 45, (pp. 1–4). <http://dx.doi.org/10.22489/CinC.2018.049>.
- Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101, <http://dx.doi.org/10.1161/01.CIR.101.23.e215>.
- Grossmann, A., & Morlet, J. (1984). Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM Journal on Mathematical Analysis*, 15(4), 723–736. <http://dx.doi.org/10.1137/0515056>, arXiv:[10.1137/0515056](https://arxiv.org/abs/10.1137/0515056).
- Heinzer, R., Vat, S., Marques-Vidal, P., Marti-Soler, H., Andries, D., Tobback, N., Mooser, V., Preisig, M., Malhotra, A., Waeber, G., Vollenweider, P., Tafti, M., & Haba-Rubio, J. (2015). Prevalence of sleep-disordered breathing in the general population: the HypnoLaus study. *The Lancet Respiratory Medicine*, 3, 310–318. [http://dx.doi.org/10.1016/S2213-2600\(15\)00043-0](http://dx.doi.org/10.1016/S2213-2600(15)00043-0).
- Hudson, A. N., Van Dongen, H. P., & Honn, K. A. (2020). Sleep deprivation, vigilant attention, and brain function: a review. *Neuropsychopharmacology*, 45(1), 21–30. <http://dx.doi.org/10.1038/s41386-019-0432-6>.
- Huijben, I. A. M., Hermans, L. W. A., Rossi, A. C., Overeem, S., van Gilst, M. M., & Sloun, R. J. G. (2023). Interpretation and further development of the hypnodelity representation of sleep structure. *Physiological Measurement*, 44, Article 015002. <http://dx.doi.org/10.1088/1361-6579/acaf61>.
- Iatsenko, D., McClintock, P., & Stefanovska, A. (2016). Extraction of instantaneous frequencies from ridges in time–frequency representations of signals. *Signal Processing*, 125, 290–303. <http://dx.doi.org/10.1016/j.sigpro.2016.01.024>.
- Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys*, 52(4), <http://dx.doi.org/10.1145/3343440>.

- Liew, S. C., & Thidar, A. (2021). Sleep deprivation and its association with diseases- a review. *Sleep Medicine*, 77, 192–204. <http://dx.doi.org/10.1016/J.SLEEP.2020.07.048>.
- Luo, S., & Johnston, P. (2010). A review of electrocardiogram filtering. *Journal of Electrocardiology*, 43(6), 486–496. <http://dx.doi.org/10.1016/j.jelectrocard.2010.07.007>, URL <https://www.sciencedirect.com/science/article/pii/S0022073610002852>.
- Magalang, U. J., Chen, N.-H., Cistulli, P. A., Fedson, A. C., Gislason, T., Hillman, D., Penzel, T., Tamisier, R., Tufik, S., Phillips, G., & Pack, A. I. (2013). Agreement in the scoring of respiratory events and sleep among international sleep centers. *Sleep*, 36, 591–596. <http://dx.doi.org/10.5666/sleep.2552>.
- Malekzadeh, M., Hajibabae, P., Heidari, M., & Berlin, B. (2022). Review of deep learning methods for automated sleep staging. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference* (pp. 80–86). Institute of Electrical and Electronics Engineers Inc., <http://dx.doi.org/10.1109/CCWC54503.2022.9720875>.
- Martinez-Ríos, E. A., Bustamante-Bello, R., & Navarro-Tuch, S. A. (2023). Generalized morse wavelets parameter selection and transfer learning for pavement transverse cracking detection. *Engineering Applications of Artificial Intelligence*, 123, Article 106355. <http://dx.doi.org/10.1016/j.engappai.2023.106355>, URL <https://www.sciencedirect.com/science/article/pii/S0952197623005390>.
- Martinez-Ríos, E. A., Bustamante-Bello, R., Navarro-Tuch, S., & Perez-Meana, H. (2023). Applications of the generalized morse wavelets: A review. *IEEE Access*, 11, 667–688. <http://dx.doi.org/10.1109/ACCESS.2022.3232729>.
- McCallan, N., Davidson, S., Ng, K. Y., Biglarbeigi, P., Finlay, D., Lan, B. L., & McLaughlin, J. (2021). Seizure classification of EEG based on wavelet signal denoising using a novel channel selection algorithm. In *2021 Asia-Pacific signal and information processing association annual summit and conference* (pp. 1269–1276). <http://dx.doi.org/10.48550/arXiv.2109.00852>.
- McCausland, C., Biglarbeigi, P., Bond, R., Yadollahikholes, G., & Finlay, D. (2022). Time-frequency ridge analysis of sleep stage transitions. In *2022 IEEE signal processing in medicine and biology symposium* (pp. 1–5). <http://dx.doi.org/10.1109/SPMB55497.2022.10014897>.
- McCausland, C., Biglarbeigi, P., Bond, R., Yadollahikholes, G., & Finlay, D. (2023). Towards improving inter-scorer reliability of sleep staging via time-frequency locating of sleep stage transitions with bi-LSTM networks. URL https://www.researchgate.net/publication/378300014_Towards_Improving_Inter-Scorer_Reliability_of_Sleep_Staging_via_Time-Frequency_Locating_of_Sleep_Stage_Transitions_with_Bi-LSTM_Networks, BioMedEng23.
- Medic, G., Wille, M., & Hemels, M. E. (2017). Short- and long-term health consequences of sleep disruption. *Nature and Science of Sleep*, 9, 151–161. <http://dx.doi.org/10.2147/NSS.S134864>.
- Mendonça, F., Mostafa, S. S., Ravelo-García, A. G., Morgado-Dias, F., & Penzel, T. (2019). A review of obstructive sleep apnea detection approaches. *IEEE Journal of Biomedical and Health Informatics*, 23, 825–837. <http://dx.doi.org/10.1109/JBHI.2018.2823265>.
- Mesraoua, B., Deleu, D., Hail, H. A., Melikyan, G., Boon, P., Haider, H. A., & Asadi-Pooya, A. A. (2019). Electroencephalography in epilepsy: look for what could be beyond the visual inspection. *Neurological Sciences*, 40, 2287–2291. <http://dx.doi.org/10.1007/s10072-019-04026-8>.
- Navot, A., Gilad-Bachrach, R., Navot, Y., & Tishby, N. (2006). Is feature selection still necessary? In C. Saunders, M. Grobelnik, S. Gunn, & J. Shawe-Taylor (Eds.), *Subspace, latent structure and feature selection* (pp. 127–138). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Nikkinen, S., Somaskandhan, P., Korkalainen, H., Kainulainen, S., Terrill, P. I., Gretarsdottir, H., Sigurdardottir, S., Olafsdottir, K. A., Islind, A. S., Óskarsdóttir, M., Arnardottir, E. S., & Leppanen, T. (2023). Multicentre sleep-stage scoring agreement in the sleep revolution project. *Journal of Sleep Research*, Article e13956.
- Noor, N. S. E. M., & Ibrahim, H. (2020). Machine learning algorithms and quantitative electroencephalography predictors for outcome prediction in traumatic brain injury: A systematic review. *IEEE Access*, 8, 102075–102092. <http://dx.doi.org/10.1109/ACCESS.2020.2998934>.
- Olhede, S., & Walden, A. (2002). Generalized morse wavelets. *IEEE Transactions on Signal Processing*, 50(11), 2661–2670. <http://dx.doi.org/10.1109/TSP.2002.804066>.
- Penzel, T. (2022). Sleep scoring moving from visual scoring towards automated scoring. *Sleep*, 45, <http://dx.doi.org/10.1093/sleep/zsac190>.
- Penzel, T., Zhang, X., & Fietze, I. (2013). Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. *Journal of Clinical Sleep Medicine*, 09, 89–91. <http://dx.doi.org/10.5664/jcs.m.2352>, URL <http://jcs.m.aasm.org/doi/10.5664/jcs.m.2352>.
- Peppard, P. E., Young, T., Barnett, J. H., Palta, M., Hagen, E. W., & Hla, K. M. (2013). Increased prevalence of sleep-disordered breathing in adults. *American Journal of Epidemiology*, 177, 1006–1014. <http://dx.doi.org/10.1093/AJE/KWS342>, URL <https://academic.oup.com/aje/article/177/9/1006/145450>.
- Pienaar, P. R., Kolbe-Alexander, T. L., van Mechelen, W., Boot, C. R., Roden, L. C., Lambert, E. V., & Rae, D. E. (2021). Associations between self-reported sleep duration and mortality in employed individuals: Systematic review and meta-analysis. *American Journal of Health Promotion*, 35, 853–865. <http://dx.doi.org/10.1177/0890117121992288>, URL <https://journals.sagepub.com/doi/abs/10.1177/0890117121992288>.
- Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2019). A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, 27(3), 491–497. <http://dx.doi.org/10.1093/jamia/ocz192>, URL <https://www.sciencedirect.com/science/article/pii/S0378779622008872>.
- Rosenberg, R., & Hout, S. V. (2013). The American academy of sleep medicine inter-scorer reliability program: Sleep stage scoring. *Journal of Clinical Sleep Medicine*, 9, 81–87. <http://dx.doi.org/10.5664/jcs.m.2350>.
- Salles, R. S., & Ribeiro, P. F. (2023). The use of deep learning and 2-D wavelet scalograms for power quality disturbances classification. *Electric Power Systems Research*, 214, Article 108834. <http://dx.doi.org/10.1016/j.epsr.2022.108834>, URL <https://www.sciencedirect.com/science/article/pii/S0378779622008872>.
- Shattuck, E. C., & Sparks, C. S. (2022). Sleep duration is related to increased mortality risk through white blood cell counts in a large national sample. *American Journal of Human Biology*, 34, Article e23574. <http://dx.doi.org/10.1002/AJHB.23574>, URL <https://onlinelibrary.wiley.com/doi/full/10.1002/ajhb.23574>, <https://onlinelibrary.wiley.com/doi/abs/10.1002/ajhb.23574>, <https://onlinelibrary.wiley.com/doi/10.1002/ajhb.23574>.
- Stephansen, J. B., Olesen, A. N., Olsen, M., Ambati, A., Leary, E. B., Moore, H. E., Carrillo, O., Lin, L., Han, F., Yan, H., Sun, Y. L., Dauvilliers, Y., Scholz, S., Barateau, L., Hogl, B., Stefani, A., Hong, S. C., Kim, T. W., Pizze, F., Mignot, E. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications*, 9, 5229. <http://dx.doi.org/10.1038/s41467-018-07229-3>.
- Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513, 429–441. <http://dx.doi.org/10.1016/j.ins.2019.11.004>, URL <https://www.sciencedirect.com/science/article/pii/S0020025519310497>.
- Wachowiak, M. P., Wachowiak-Smolíková, R., Johnson, M. J., Hay, D. C., Power, K. E., & Williams-Bell, F. M. (2018). Quantitative feature analysis of continuous analytic wavelet transforms of electrocardiography and electromyography. *Philosophical Transactions of the Royal Society of London A (Mathematical and Physical Sciences)*, 376(2126), Article 20170250. <http://dx.doi.org/10.1098/rsta.2017.0250>, URL <https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2017.0250>, <https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2017.0250>, [arXiv:https://arxiv.org/abs/10.1098/rsta.2017.0250](https://arxiv.org/abs/10.1098/rsta.2017.0250).
- Yao, K., & Gao, J. (2016). Law of large numbers for uncertain random variables. *IEEE Transactions on Fuzzy Systems*, 24, 615–621. <http://dx.doi.org/10.1109/TFUZZ.2015.2466080>.
- Young, T., Palta, M., Dempsey, J., Peppard, P. E., Nieto, F. J., & Hla, K. M. (2009). Burden of sleep apnea: Rationale, design, and major findings of the wisconsin sleep cohort study. *Wisconsin Medical Journal*, 108, 246–249.
- Yue, H., Chen, Z., Guo, W., Sun, L., Dai, Y., Wang, Y., Ma, W., Fan, X., Wen, W., & Lei, W. (2024). Research and application of deep learning-based sleep staging: Data, modeling, validation, and clinical practice. *Sleep Medicine Reviews*, 74, Article 101897. <http://dx.doi.org/10.1016/j.smrv.2024.101897>.
- Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., & Redline, S. (2018). The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25, 1351–1358. <http://dx.doi.org/10.1093/jamia/ocy064>.
- Zhang, X., Zhang, X., Huang, Q., Lv, Y., & Chen, F. (2024). A review of automated sleep stage based on EEG signals. *Biocybernetics and Biomedical Engineering*, 44, 651–673. <http://dx.doi.org/10.1016/j.bbe.2024.06.004>.