SViT

a Spectral Vision Transformer for the Detection of REM Sleep Behavior Disorder

Gunter, Katarina Mary; Brink-Kjær, Andreas; Mignot, Emmanuel; Sorensen, Helge B.D.; During, Emmanuel; Jennum, Poul

Published in:
IEEE Journal of Biomedical and Health Informatics

Link to article, DOI:
10.1109/JBHI.2023.3292231

Publication date:
2023

Document Version
Peer reviewed version

SViT: a Spectral Vision Transformer for the Detection of REM Sleep Behavior Disorder

Katarina Mary Gunter, Andreas Brink-Kjær, *Member, IEEE*, Emmanuel Mignot, Helge B.D. Sørensen, *Senior Member, IEEE*, Emmanuel During, and Poul Jennum

***Abstract*— REM sleep behavior disorder (RBD) is a parasomnia with dream enactment and presence of REM sleep without atonia (RSWA). RBD diagnosed manually via polysomnography (PSG) scoring, which is time intensive. Isolated RBD (iRBD) is also associated with a high probability of conversion to Parkinson's disease. Diagnosis of iRBD is largely based on clinical evaluation and subjective PSG ratings of REM sleep without atonia. Here we show the first application of a novel spectral vision transformer (SViT) to PSG signals for detection of RBD and compare the results to the more conventional convolutional neural network architecture. The vision-based deep learning models were applied to scalograms (30 or 300 second windows) of the PSG data (EEG, EMG and EOG) and the predictions interpreted. A total of 153 RBD (96 iRBD and 57 RBD with PD) and 190 controls were included in the study and 5-fold bagged ensemble was used. Model outputs were analyzed per-patient (averaged), with regards to sleep stage, and the SViT was interpreted using integrated gradients. Models had a similar per-epoch test F1 score. However, the vision transformer had the best per-patient performance, with an F1 score 0.87. Training the SViT on channel subsets, it achieved an F1 score of 0.93 on a combination of EEG and EOG. EMG is thought to have the highest diagnostic yield, but interpretation of our model showed that high relevance was placed on EEG and EOG, indicating these channels could be included for diagnosing RBD.**

***Index Terms*— Computer vision, deep learning, Parkinson's disease, polysomnography, RBD, vision transformer.**

I. Introduction

The prevalence Parkinson's disease (PD) in the population, along with other neurodegenerative diseases, is expected to increase by 23% by 2025 [1], and yet the lack of understanding of the disease process limits treatment options. Patients who are diagnosed with Rapid eye movement (REM) sleep behavior disorder (RBD), present a unique population in which to study the spread of neurodegeneration and associated symptoms, as isolated RBD is associated with a high risk of phenoconversion to one of the alpha-synucleopathies – most commonly Parkinson's disease (PD) and Dementia with Lewy Bodies (DLB) but also Multiple System Atrophy (MSA) [2], [3]. An unmet need is to diagnose these disorders before significant detoriation is evident which has the potential for preventive or protective treatment.

Rapid eye movement (REM) sleep behavior disorder (RBD) is a sleep disorder mainly characterized by abnormal motor activity during REM sleep, most noticeably exhibited through dream enactment. Under normal conditions, except for diaphragm and extraocular muscles, all skeletal muscle motor activity signals are inhibited during REM sleep, resulting in paralysis (or REM sleep atonia). The clinical progression is also aligned with the theoretical Braak staging of PD [4]. The exact mechanism behind phenoconversion is not currently understood, and a greater understanding of this process is needed to develop neuroprotective therapeutics.

Deep learning and computer vision has now evolved to the point where multiple algorithms, including state-of-the-art vision transformers, can achieve very high performance on a whole host of problems, such as image classification, object detection, image segmentation, and video analysis.

II. Motivation

Attaining high performance when implementing commonly used computer vision architectures is largely agreed to be attributed to training on vast amounts of correctly labelled image data. This is particularly true for the state-of-the-art image transformer, which was shown to outperform other algorithms, when pre-trained with a large amount of data [5]. The data sets upon which state-of-the-art algorithms are applied to in these papers are also curated for a particular task and relatively structured. While the potential for deep learning to have a significant effect on patient care and treatment is there, the issues that arise with attaining or dealing with real-world data sets from patients - such as consent, study withdrawal and noise - result in clinical data sets which are

K. M. Gunter was with the Technical University of Denmark, and is currently with the Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, Oxford, U.K. (e-mail: katarina.gunter@ndcn.ox.ac.uk).

A. Brink-Kjær is with the Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark (e-mail: andbri@dtu.dk).

E. Mignot is at the Center for Sleep Sciences and Medicine, Stanford University, CA, USA (e-mail: mignot@stanford.edu).

H. B. D. Sørensen is with the Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark (e-mail: hbds@dtu.dk).

E. During was at the Department of Psychiatry and Behavioural Sciences, Stanford University, and is currently at Mount Sinai, New York (e-mail: emmanuel.during@mssm.edu).

P. Jennum is at the Danish Center for Sleep Medicine, Glostrup University Hospital, Glostrup, Denmark (e-mail: poul.joergen.jennum@regionh.dk).

E. During and P. Jennum contributed equally to this work.

small, biased, and carry a risk of sample annotation error. It is thus non-trivial to imply that these algorithms are even remotely comparable to field-expert assessors, or that they can achieve such high performance on real world clinical data sets. PSG data is a good example of clinical data which is typically noisy, and cohorts of RBD subjects are typically small compared to more widespread disease groups, and while the AASM definition of RBD is binary, the disorder is not.

Currently, RBD is diagnosed via the visual inspection of polysomnography (PSG) data – a range of physiological signals recorded during sleep over one night. Visual inspection of the PSG signals also requires an experienced technician and trained sleep specialist. This method presents a host of limitations, as manual annotations are subject to inter-scorer variation, and it is not a truly objective method. Visual inspection of the PSG signals also requires an experienced technician and trained sleep specialist. Furthermore, PSG-based diagnosis is currently based on the combination of RSWA and clinical history of dream-enactment or, alternatively, evidence of complex behaviors and/or vocalizations during REM sleep identified on PSG video recordings, per the American Academy of Sleep Medicine guidelines [6]. However, recent studies suggest that RBD may have a wider range of neurophysiological abnormalities beyond REM sleep that could be detected on PSG, such as loss of hypotonia during during non-REM (NREM) sleep [7], changes in resting state EEG [8],] including micro-sleep event abnormalities [9], abnormal EEG oscillations in NREM [10], EOG abnormalities [11], and changes in autonomic activity, e.g. heart rate variability (HRV) [12]. Thus, the discovery of novel PSG biomarkers could potentially be used for diagnosis, and for advancing our understanding of the neurodegenerative mechanisms underlying synucleinopathies.

Given that many relevant aspects of physiological signals are captured within the time-frequency domain, computer-vision models were applied to the scalograms of the PSG signals from patients with RBD and CC. While to date, multiple studies have used computer vision to analyse scalogram data for the purpose of sleep stage classification [13], [14], to our knowledge, there is limited published work on applying computer vision to spectral data for classifying RBD and none have applied a vision transformer to this type of data. Ruffini et. al. showed that CNNs and RNNs could distinguish between control subjects and RBD phenoconversion to PD, based on scalograms [15]; however, this was based on only a few minutes of resting state EEG, rather than PSG data and a full electrode montage. Consequently, the aim of this study was to implement novel deep learning architectures to classify patients with RBD versus clinical controls without RBD (CC). Here, both a simple convolutional neural network (CNN) and a dilated CNN, are compared to the state-of-the-art vision transformer. To understand the classification of a model, as well as possibly elucidate novel pathophysiological insights, an interpretation method was applied.

TABLE I
DEMOGRAPHICS

Variable	**STNF CC [n=98]**	**STNF RBD [n=71]**	**DCSM CC [n=92]**	**DCSM RBD [n=82]**
Age (mean ± SD)	63.8 ± 9.2	66.5 ± 9.1	51.5 ± 16.5	64.1 ± 12.5****
Sex: Male (%)	65 (66 %)	51 (72 %)	50 (54 %)	59 (72 %)
AHI (mean ± SD)	14.1 ± 13.0	23.3 ± 20.3*	7.6 ± 9.8	12.1 ± 16.9
PLM (%)	11 (11 %)	0 (0 %)**	41 (45 %)	5 (1 %)****
iRBD (%)	—	61 (86 %)	—	35 (43 %)
RBD+PD (%)	—	10 (14 %)	—	47 (57 %)
Sleep Stage				
Wake %	19.7	22.8****	14.5	22.3
REM %	24.4	13.1	17.8	12.9****
N1 %	10.6	8.3	8.4	12.0****
N2 %	34.7	49.5	43.1	39.1****
N3 %	10.7	5.6****	14.7	11.9****

Summary of demographics, co-morbidities and patient distributions. STNF: Stanford cohort. DCSM: Danish Center for Sleep Medicine cohort. AHI: Apnea-hypopnea index. PLM: Periodic leg movement index $>$ 15/hour. SD: Standard deviation. Sleep stage indicated as percentage of total hypnogram within defined stage in RBD (REM-Sleep Behaviour Disorder) and Controls (C). Significance between cohort (STNF/DCSM) RBD and controls as determined by the Mann-Whitney U test for continuous variables and independent t-test for binary variables. $p < 0.0001$ (****), $p < 0.01$ (**), $p < 0.05$ (*). 2 patients did not have hypnograms.

III. METHODS

A. Data

Large, good quality data sets within the medical sector are sparse, particularly when it comes to in-clinic data. The limitations behind this are multi-faceted: data sets are limited by collection of data from patients and consent, and data that is collected is affected by annotation error, inter-scorer variability and noise related to the nature of data from human subjects, e.g. artifarcts due to movement and electrode displacements. This is particularly relevant for PSG data. Thus, here two different data sets have been collated, to evaluate whether a state-of-the-art model can generalize to this type of data from multiple sources, as well as an evaluation of if any value can be extracted using advanced models in this low data, noisy regime. A total of 343 full night PSG recordings/patients were included, and comprised of 96 iRBD, 57 RBD+PD and 190 CC. Data was sourced from two different sleep clinics, the Danish Center for Sleep Medicine (DCSM), Department of Clinical Neurophysiology, Rigshospitalet, Denmark, and the Stanford Center for Sleep Sciences and Medicine (STFD), Stanford University, Redwood City, California. The demographics, patient groupings, and co-morbidities are shown in Table I. All participants provided written informed consent. This study was approved by the Institutional Review Board of Stanford (protocol #56218) and the Danish Health Authorities, as well as the Data Protection Agency.

B. Preprocessing of Polysomnographic Signals

The v-PSG signals used as input to the neural network models included electroencephalogram (EEG) (C3, C4, F3,

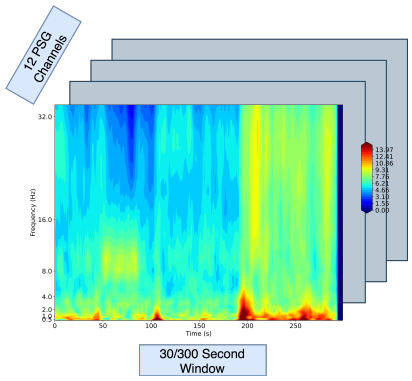

Fig. 1. Illustration of data set up. Whole night scalograms were computed for each patient for 6 EEG, 3 EMG and 3 EOG channels and stacked. These stacks were then segmented to produce stack epochs (e.g. 30 seconds) and one epoch stack is regarded as one input sample to the model.

F4, O1, O2), electromyography (EMG) (chin, left tibia, and right tibia), and electrooculography (EOG) (left and right) signals. These were raw signals, without any removal of artifacts or epochs. Scalograms were computed between 'lights off' and 'lights on' annotations. If 'light off' and 'light on' annotations were not available for a given patient, sleep staging scores were used to truncate the signals to the first and last sleep stage which was not scored as wake. As the data set included different cohorts, PSG signals were also standardized. Standardization was based on the most common settings; thus, signals were re-sampled to 250 Hz, using the polyphase method provided by the SciPy library, and a standard unit of measurement (μV) was set.

Furthermore, the unreferenced PSG signals were referenced using the contralateral mastoid (for EEG and EOG signals). As most of the relevant information given in EOG signals is directional, a third EOG signal was also included, which was created by subtracting the two given EOGs. This was done to account for the loss of phase information which occurs when calculating the scalogram of a signal. The scalograms/scalograms were computed using the Continuous Wavelet Transform (CWT), using the Morlet wavelet [16]. For each channel, 25 frequency bins were considered, ranging from 0.31-33.8 Hz for EEG, 10.1-101.5 Hz for EMG, and 0.2-31.2 Hz for EOG. Due to the computational expense of the deep learning algorithms, the time-axis was re-sampled to 5Hz using a moving average. The resulting dimensions were Nc x Nf x window, where Nc is the number of channels (12), Nf is the number of frequency bins, and window is the epoch length (150 or 1500 samples, corresponding to 30 or 300 seconds). An example of the data set up is shown in Figure 1.

C. Deep Learning Models and Training Methods

Three different deep learning model architectures were explored for this work, two of which were convolutional neural networks (CNN), which are commonly implemented for the task of image classification. All models were implemented from scratch, without pre-training. A simple CNN model was implemented, based on the architecture shown in Cesari

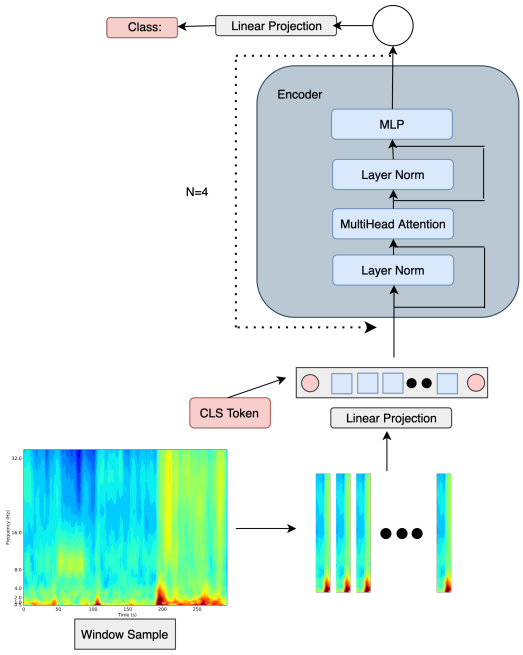

Fig. 2. Illustration of transformer model architecture. CLS token: classification token. MLP: Multi-Layer Perceptron (feed forward neural network).

et al. [14], in which it attained high performance on similar data. In Cesari et al., this network architecture was used for sleep stage classification of short epochs of EEG and EOG scalograms, in PD/PD+RBD patients. This architecture is analogous to applying a frequency domain and subsequent time-domain filter bank to the scalogram, a convolution over frequency dimension and time dimension, respectively (see Appendix II, Figure 5). In addition to the simple CNN, a more complex model, as well as larger scalogram segmentation window, were explored. To increase the receptive field without increasing the computational cost, a dilated CNN architecture was implemented. See Table VII in Appendix II for an overview of the architecture.

While the dilated CNN results in a larger receptive field, relevant non-local interactions will only be captured by the model if the dilation factor is appropriate. The state-of-the-art vision transformer model by-passes this problem. Transformers have typically been used for natural language processing (NLP) tasks, and also recently been used for the task of image classification [5]. For a transformer, the receptive field is the entire input, and the model can learn all non-local interactions. With sufficient training data, the vision transformer has been found to outperform more traditional CNN architectures. An illustration of the adapted spectral vision transformer (SViT) implemented in this work is shown in Figure 2. See Appendix II for more details and https://github.com/katarinamg/svit.

Libraries for CNNs and other image classification architectures typically have a hard-coded 3 channel input, and thus the architectures described were self-implemented to be compatible with 12 channel input PSG data. All models

were trained on segments of the full nights' scalogram, as inputting the entire scalogram as one sample would both be very computationally expensive and would significantly reduce the number of training samples. Input segments of 30 seconds were used in both the simple CNN and the vision transformer, whereas the dilated CNN allowed for exploration of how larger segments (5-minute windows) of the scalograms would influence the classification. For comparison, the vision transformer was also trained on 5-minute windows. The data set was split to a 70:10:20 ratio, for training, validation, and test, respectively. This split was on a patient basis, such that all epochs belonging to a given patient were in the same set. During training, a batch size of 64 for the CNN models and 30 second transformer, and 8 for the 5 minute transformer, was used. To increase the stability of training, a 5-model ensemble was implemented, whereby the model was trained and validated on different splits of the 80% allocated to train/validation. The models were trained and evaluated on the validation set after each training epoch, and ensembled to give a robust performance metric of each architecture on the test set. Early stopping was used with a patience of 3 epochs. The binary cross-entropy loss function was used to optimize the models and as the evaluation metric on the validation set.

All models were optimized using the adaptive moment estimation (Adam) [17]. A cosine learning rate decay scheduler was also implemented, which has been found to often outperform stepwise decay [18]. A dropout of 10% was also applied to all three models. Data augmentation has also been shown to be particularly useful for improving the robustness of image classification problems and when dealing with low data problems; here a scalogram specific augmentation is applied to the input samples. SpecAugment [19], a method developed by Google Brain, applies random time and frequency masking to the input samples and was applied during training with a maximum frequency and time mask of 15 Hz and 10 seconds, respectively.

D. Model Evaluation

Both the CNNs and transformers were evaluated on the test set by ensembling the 5 models trained on different splits of data. The geometric mean was used combine the predictions on the test set from each ensemble model. The evaluation was based on the F1 score metric, rather than the accuracy, as this encompasses both precision and recall. Thus, the raw logit outputs from the models were converted to binary labels by applying the sigmoid function and thresholding the resulting probabilities at 0.5 (prediction 0.5 = 1). The F1 score was calculated for each class, as well as overall (weighted average).

The performance of each model was also evaluated on a per-patient basis, by averaging the predicted probabilities over an entire night. The patient probability output was then thresholded, using a threshold which was grid-optimized on the validation set to maximise the F1 score (average threshold found from each of the 5 train/validation split models).

TABLE II
MODEL PERFORMANCE

Per Epoch	**RBD**	**CC**	**Overall**
Simple CNN	0.78	0.74	0.77
Dilated CNN	0.73	0.77	0.75
SViT[a]	0.77	0.80	0.79
Large SViT[b]	0.81	0.83	0.82
Per Patient			
Simple CNN	0.83	0.82	0.83
Dilated CNN	0.78	0.78	0.78
SViT	0.82	0.81	0.81
Large SViT	0.87	0.87	0.87

Performance of models per epoch and per patient, given as F1 score. Results given as F1 score on each class as well as a weighted average. [a]Spectral Vision transformer. [b] 5-Minute Spectral Vision transformer.

E. Interpretation of SViT

Deep learning models are often described as 'black box' models, and few studies interpret what a given model is basing its predictions on. However, there are ways in which deep learning models can be interpreted [20]. In addition to the evaluation of the models, the predictions of the large (5-minute) vision transformer (SViT) were interpreted using integrated gradients [21], which generates a relevance score for each input pixel of the scalogram inputs. The relevance scores were averaged across time to produce relevancy-frequency plots for each PSG channel. Briefly, integrated gradients are calculated by integrating the gradient with respect to the input at all points along the path from a baseline (x') to the input (x). The baseline given was a zero scalar which corresponds to each input. The final importance output is multiplied by the difference between x and x' (input-baseline).

In addition to the interpretation of the large SViT, the predictions of this model were analyzed in relation to sleep stages. For each class, the performance as given by the F1 score of the given class for wake, REM, N1, N2 and N3 sleep was calculated. As the output was given in 5 minute windows, the prediction was repeated for each 30 second sleep stage within the window.

IV. RESULTS

The performance of each type of ensembled model on epochs of the data are shown in Table II. Interestingly, the small SViT had only a very minor increase in F1 score (0.79) compared to the simple CNN (0.77). The dilated CNN had the lowest performance, despite the increase in window size. The state-of-the-art large SViT obtained the highest F1 score (0.82), an improvement over the other models. The performance of the models when optimizing the threshold for a per patient prediction can be found in Table II. We again see the same trend in performance: and the large vision transformer again outperforms the other models (F1 score of 0.87).

As we are exploring the effect of applying these models to data from multiple sources, the F1 scores of all three models with regards to the data cohort (DCSM and STFD) are given in Table III. All models achieved a significantly

TABLE III
COHORT PERFORMANCE

Model	STFD[a]	DCSM[b]	STFD PP[c]	DCSM PP[d]
Simple CNN	0.88	0.57	0.86	0.78
Dilated CNN	0.84	0.59	0.86	0.67
SViT	0.91	0.58	0.93	0.63
Large SViT	0.92	0.67	0.95	0.74

Performance of models per epoch and per patient on each cohort, given as weighted average F1 score. [a]Stanford cohort per epoch. [b]Danish Center for Sleep Medicine cohort per epoch. [c]Stanford cohort per patient. [d]Danish Center for Sleep Medicine cohort per patient.

TABLE IV
LARGE SVIT CHANNEL PERFORMANCE

Channel Subset	RBD	CC	Overall
EEG	0.87	0.89	0.88
EMG	0.77	0.80	0.78
EOG	0.87	0.87	0.87
EEG+EMG	0.86	0.85	0.86
EEG+EOG	0.93	0.92	0.93
EMG+EOG	0.80	0.82	0.81

Performance of Large Spectral Vision Transformer (SViT) channel subset models per patient given as weighted average F1 score.

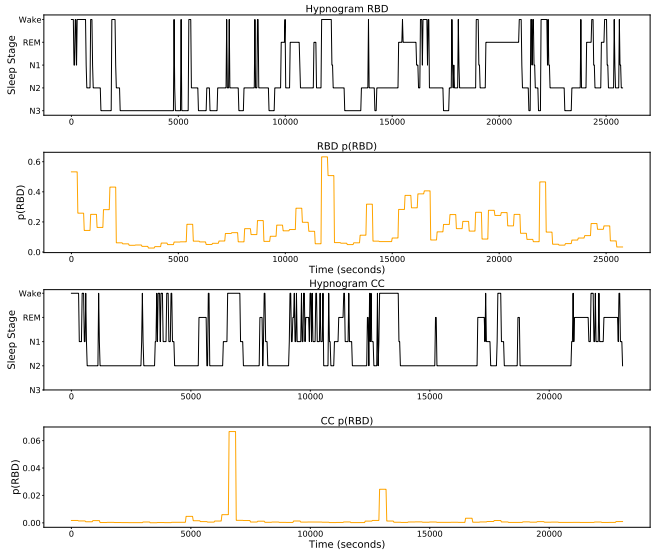

Fig. 3. Example of probability of RBD in one RBD and one control (CC) subject with corresponding hypnogram.

higher performance in the STFD cohort, with the large SViT achieving an F1 score of 0.95 on a per patient basis.

The large SViT architecture was also explored in relation to channel inputs. Channel subsets and per patient performance is shown in Table IV. Interestingly, the combination of EEG and EOG resulted in the highest weighted F1 score (0.93), and EMG alone had the lowest performance (Weighted F1: 0.78).

The outputs of the large SViT were also investigated with regards to manual sleep stage scoring. The F1 score for the test set, within each sleep stage, are outlined in Table V. We see the highest F1 score in REM and N2, while wake and N1 are not significantly lower. An example of the probability output from the large SViT in one RBD and one control patient with the corresponding hypnogram is shown in Figure 3.

To further investigate the large SViT, the average relevancy score for each channel type, over the frequency range, are plotted in Figure 4. An attribution above 0 indicates relevancy for predicting RBD, whereas an attribution below 0 indicates relevancy for classifying data as a CC. The raw and absolute relevance scores for given EEG frequency ranges are summarized in Table VI.

TABLE V
SLEEP STAGE RBD ACCURACY

Large SViT	Wake	REM[a]	N1[b]	N2[c]	N3[d]
RBD	0.84	0.79	0.82	0.85	0.59
CC	0.81	0.89	0.83	0.82	0.77
Weighted	0.83	0.85	0.82	0.84	0.70

F1 score with respects to sleep stage using the Large spectral vision transformer. [a]Rapid Eye Movement, [b]Non-REM 1, [c]Non-REM 2, [d]Non-REM 3.

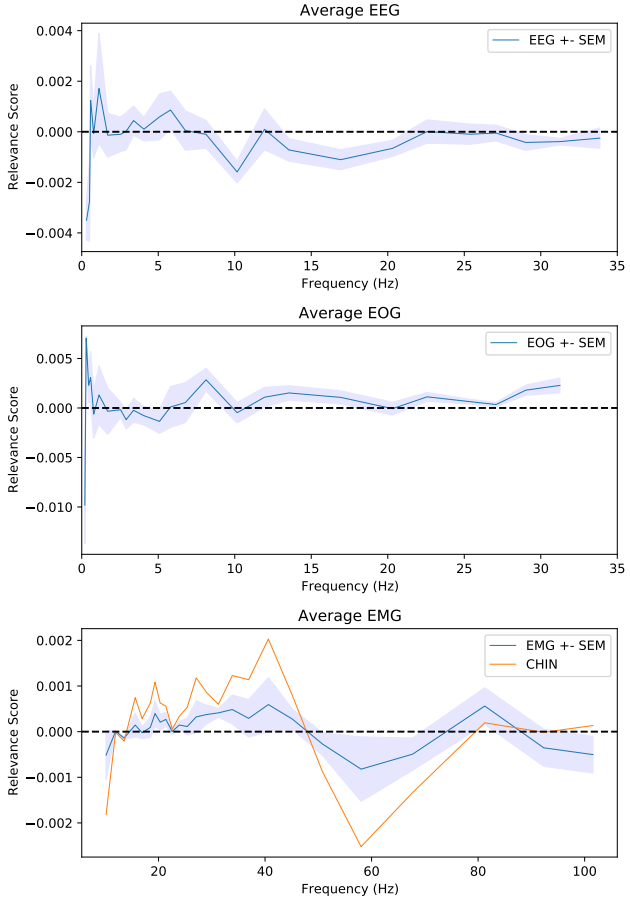

Fig. 4. Plot of relevancy scores for EEG, EMG and EOG, averaged over channels, as well as Chin alone, with regards to sleep stage and frequency. SEM: Standard error of the mean.

TABLE VI
FREQUENCY BANDS RELEVANCY

Frequency Band	Absolute Relevancy	Raw Relevancy
Delta	3.8×10^{-4}	-3.5×10^{-4}
Theta	4.6×10^{-4}	-3.9×10^{-4}
Alpha	5.2×10^{-4}	5.2×10^{-4}
Beta	1.1×10^{-3}	-3.5×10^{-6}

Average absolute and raw relevancy scores for the Large spectral vision transformer in EEG channels at corresponding frequency ranges.

V. DISCUSSION

A. RBD Detection Performance

We explored whether a state-of-the-art model could achieve high performance on PSG data, which is known to suffer from the limitations often associated with medical data. Furthermore, current methods for diagnosing of RBD by PSG rely exclusively on the motor expression of the disorder, namely RSWA and behaviors observed during video recording while in REM sleep. As former studies have proven that there are several other abnormalities in EEG and EOG signals, we evaluated the use of EEG and other correlated PSG channels to test if these contain relevant information for RBD detection and found that these measures are useful for diagnostic purposes.

Interestingly, the three deep learning model architectures implemented in this work, the large vision transformer obtained the highest per patient classification F1 score of 0.87, when distinguishing between RBD and CC patients. This shows that state-of-the-art models which have been created based on large amounts of good quality data, should still be considered when dealing with the small, noisy data regime.

We can see from Table III that there was a significant discrepancy between the per-epoch performance in the STFD and DCSM dataset, with overall per patient F1 scores of 0.95 and 0.74, respectively. This may be due to differences in the evaluation of RSWA between centers, and/or signal quality. Individuals in the STFD cohort were also older, which may be associated with a more advanced stage of neurodegeneration as compared to DCSM where cases with RSWA were included. The small CNN model and small SViT had a similar performance, with only a small difference in the per-epoch and per-patient performance, while the dilated CNN had the lowest F1 score - despite the dilated CNN using a larger epoch window (5 minutes). The dilated CNN and simple CNN obtained per patient F1 scores of 0.78 and 0.83, respectively. This may be due to using a dilated CNN with a dilation factor which is not able to capture the relevant non-local interactions within the scalogram segment, as the superior performance of the large SViT would suggest that a larger window is informative for classification. When varying the channel subset as input to the large SViT, the EEG and EOG attains the highest performance, and comfortably beats the full channel set, likely indicating that including all channels may lead to over-fitting.

As there were significant differences in AHI and age between the DCSM RBD and control groups, we examined whether this altered the ability of the model to classify subjects. The association between these variables and the binary output were tested using logistic regression models, and the coefficients and associated p-values reported (see Appendix I). We found that neither AHI or age had a significant effect.

B. Physiological Interpretation

While the large SViT cannot be directly compared to individual 30-second epochs of the manually staged hypnograms, the 5-minute predictions were analysed in relation to each sleep stage present within the window. Table V shows that the F1 score in each sleep stage closely follows what would be expected based on the literature - with the highest F1 score within REM sleep, followed by N2 and wake. As these stages could all be included in a prediction window, concrete conclusions cannot be drawn. The example in Figure 3 of the probability output of the model shows that probability of RBD is high in sections extended periods of wake and REM sleep. PD has previously been shown to cause alterations in the ratio of all sleep stages compared to healthy controls [22], [23] – thus one may speculate as to whether the models are picking up on features related to PD from the RBD with PD group. However, given the analysis shown in Table III and the ratio of iRBD to RBD with PD, it is unlikely that including these patients would have such a major effect on the relevancy scores. In the DCSM and STNF cohort we saw significant differences in the percentage of REM/N1/N2 and wake, respectively, between RBD and control subjects, which may be contributing to the high accuracy within these sleep stages. However, we also see significant difference in the amount of N3 sleep, which had the lowest F1 score, strengthening the indication that fewer physiological markers of the disease are present within this stage.

In relation to the channel and frequency relevancy, for EEG channels, a high positive relevancy is attributed to low frequencies, specifically within the delta and theta range (0.5-8 Hz). This is also reflected in the raw relevancy scores, with a high relevance for prediction of RBD found within the theta band. Both results correspond well with EEG slowing, a known marker of RBD [24]. We see a similar trend with the attribution scores for the EOG channels, with a very high relevance score for low frequencies (0.2 – 4 Hz), and in fact obtain the highest relevancy scores overall for the EOG channels within this range. This may reflect previous findings of changes in eye movements in patients with RBD and PD, with PD showing more eye movement than controls during N2 and REM sleep, and RBD showing less eye movement than controls during wake [8]. The relevancy scores for EMG channels are less informative. Given that PLM may be a confounder, we also plot the relevancy for the chin channel alone - however, this is not significantly different from the average trend.

C. Comparison with Previous Works

While other studies [25] have obtained similar performance using simpler model architectures using various PSG channels, it should be noted that it is difficult to compare automatic methods for the detection of RBD without applying all methods to the same dataset, as PSG datasets in RBD tend to be small and heterogenous. However, other measures for RBD detection have been wide-spread, including, but not limited to, frequency component analysis [26], topic modelling [27], sleep spindle density [9], and arousal characteristics [28]. These studies focus on certain elements related to PSG data and rely on feature extraction. The method presented here combines EEG, EMG and EOG and requires no feature extraction or engineering and is also relatively interpretable, and methods such as this are more attractive for biomarker discovery. The interpretation of the large SViT implies that EEG and EOG have high relevance for classification, suggesting that these modalities may be used rather than EMG for diagnosis.

A potential use for these findings is that diagnosing RBD can further be improved including EEG and EOG channels, and potentially solely using these channels for identification; this has implication for simpler diagnostic devices identifying RBD and earlier diagnosis of PD and related disorders.

D. Limitations and Future Work

One aspect of the picture that this study lacks, is incorporating sleep architecture into the model. This could be implemented by way of including both a CNN and recurrent neural network (RNN) in the method, as can be seen for PSG data in Brink-Kjær et al. [29], or by way of using a transformer to classify the entire nights scalogram using "chunked" data. However, this is likely not ideal when dealing with a small data set – as one would be optimizing the model based on only a few hundred samples.

There was a clear difference in the performance on the two cohorts. The reason for this is unclear, but may be due to data quality and recording methods - which often differ between countries or sites. Furthermore, we cannot draw clear conclusions from the sleep stage analysis due to the nature of the predictions on larger windows. We hypothesize that including sleep stage information within the model would increase the performance, and in terms of biomarkers, sleep stage specific models may be informative. Here we also use only a subset of the electrophysiological signals collected during the polysomnography, and the inclusion of other signals and information (e.g. heart rate variability) could result in a higher F1 score. However, this is outside the scope of this work. More analysis on which epochs were correctly or incorrectly classified could also provide more information on specific wave forms which contribute to the prediction. By adapting existing packages, such as GradCAM to this novel transformer would allow us to visualise the attributions in given epochs.

E. Conclusion

In this paper we implemented an adapted spectral vision transformer (SViT), which can be directly applied to N-Channel PSG data, which we show to comfortably beat a baseline CNN and a dilated CNN, when applied to 5-minute scalogram windows, a on a relatively small and noisy PSG data set. We also investigated whether a novel image classification model interpretation methods can be applied to PSG data to discover new biomarkers which could be of interest to clinicians and medical researchers. Whereas diagnostic criteria for RBD exclusively relies on abnormal RSWA and behaviors demonstrated during REM sleep, our model found relevant biomarkers are also expressed during NREM sleep, and this should be further investigated.

APPENDIX I
DEMOGRAPHICS

To test whether the significant difference in age and the apnea-hypopnea index played a significant role in the output of the model, both the raw logit output from the model and the demographic variable were fed into a logistic regression model. The model was fit to predict the binary output. This method resulted in three models with different input combinations: logit alone, logit and age, logit and AHI. To examine how this affected the prediction, the weight applied in the logit only model was compared to the weights applied in the logit+variable models. There was no reduction in the weight applied to the logit, indicating the the variables are not correlated. Furthermore, we also examined the 95% CI, as well as p-values, of the coefficients. Table VII shows that the difference in these variables did not have a significant effect the models ability to classify subjects.

TABLE VII
LOGISTIC REGRESSION ANALYSIS OF ASSOCIATION BETWEEN DEMOGRAPHIC VARIABLES AND MODEL OUTPUT

Model	**Coefficients**	**95% CI**	**p-value**
Logit	0.57	[0.26-0.90]	$\mathbf{4.3 \times 10^{-4}}$
Logit + Age	0.64, 0.05	[0.31-0.96, -0.01-0.13]	$\mathbf{1.2 \times 10^{-4}}$, 0.11
Logit + AHI	0.57, -0.01	[0.26-0.92, -0.04-0.03]	$\mathbf{4.6 \times 10^{-4}}$, 0.71

Association between transformer output and variables using logistic regression.

APPENDIX II
MODEL ARCHITECTURES AND PERFORMANCE

In Figure 5. the architecture of the simple CNN is illustrated. A drop out of 10% was applied prior to the last fully connected layer. Table VIII shows an overview of the dilated CNN. Drop out of 10% was applied prior to the second fully connected layer and prior to the last fully

TABLE VIII
DILATED CNN

Layer	1	2	3	4	5	6	7	8	9
Conv[a]	96x1	1x3	1x3	1x3	1x3	1x3	1x3	1x3	1x1
Dilate[b]	1x1	1x1	1x1	1x2	1x4	1x8	1x16	1x1	1x1
Stride[c]	1x1	1x2	1x1	1x2	1x1	1x1	1x1	1x1	1x1
Output[d]	1x300	1x149	1x147	1x72	1x64	1x48	1x16	1x14	1x14
Field[e]	96x1	96x3	96x5	96x13	96x21	96x38	96x71	96x73	96x73

Overview of dilated convolutional neural network (CNN) architecture. [a]Size of convolutional kernel. [b]Dilation factor. [c]Stride (or step size) of the kernel over the input. [d]Size of the output from convolution. [e]Expansion of the receptive field of the input at each layer.

TABLE IX
FULL SUMMARY OF PERFORMANCE

Large SViT	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Train Loss	0.43	0.21	0.34	0.41	0.30
Validation Loss	0.35	0.45	0.49	0.43	0.37
Train F1	0.79	0.91	0.81	0.80	0.87
Validation F1	0.85	0.79	0.76	0.80	0.83
Test Loss	0.44	0.61	0.42	0.52	0.49

Full summary of the performance of each of the 5 fold models using the Large spectral vision transformer architecture.

connected layer.

As in the original ViT [5], here we consider only the encoder. 1 second slices of the input sample were split into (N x W x H) patches, where N is the number of channels, W is the width (1 second) and H is the number of frequency bins. The patches were flattened and mapped to the constant embedding vector length (768) using a linear projection. A classification token is added at this stage. The encoder included 4 layers, each consisting of alternating layers of LayerNorm, MultiHead Attention, an MLP layer. The final output from the encoder was then send through a final linear projection to give the output class. Each MultiHead Attention layer consisted of 12 heads.

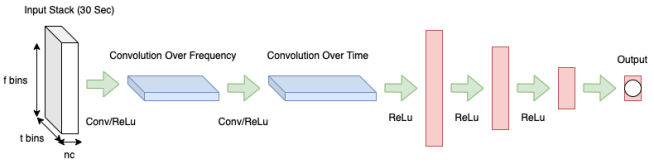

Fig. 5. : Illustration of simple convolution neural network model architecture. Input stack - stack of scalograms with dimensions frequency bins (f bins: 96), time bins (t bins: 30), and number of channels, or scalograms in the stack (nc). Convolutional layers, conv, followed by rectified unit activation functions, RELU, and three fully connected layers (red).

In Table IX we show the loss and per epoch F1 score on the training and validation for each of the 5 SViT models.

REFERENCES

[1] NICE, "Parkinson's disease: How common is it?" 2022, accessed = 2022-08-16. [Online]. Available: https://cks.nice.org.uk/topics/parkinsons-disease/background-information/prevalence

[2] C. H. Schenck, S. R. Bundlie, and M. W. Mahowald, "Delayed emergence of a parkinsonian disorder in 38% of 29 older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder," *Neurology*, vol. 46, no. 2, p. 388–393, 1996.

[3] C. H. Schenck, B. F. Boeve, and M. W. Mahowald, "Delayed emergence of a parkinsonian disorder or dementia in 81% of older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder: A 16-year update on a previously reported series," *Sleep Medicine*, vol. 14, no. 8, p. 744–748, Aug 2013.

[4] H. Braak, K. D. Tredici, U. Rüb, R. A. de Vos, E. N. Jansen Steur, and E. Braak, "Staging of brain pathology related to sporadic parkinson's disease," *Neurobiology of Aging*, vol. 24, no. 2, p. 197–211, 2003.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[6] M. J. Sateia, "International classification of sleep disorders-third edition," *Chest*, vol. 146, no. 5, p. 1387–1394, 2014.

[7] D. Levendowski, B. Boeve, D. Shprecher, J. lee Iannotti, D. Salat, J. Hamilton, T. Neylan, C. Walsh, D. Tsuang, P. Westbrook, C. Berka, G. Mazeika, E. Angel, C. Guevarra, P. Timm, and E. St. Louis, "Non-rem sleep with hypertonia: A potential prodromal biomarker for -synuclein-related neurodegenerative disease (1943)," *Neurology*, vol. 96, no. 15 Supplement, 2021. [Online]. Available: https://n.neurology.org/content/96/15_Supplement/1943

[8] M. T. Pascarelli, C. Del Percio, M. F. De Pandis, R. Ferri, R. Lizio, G. Noce, S. Lopez, M. Rizzo, A. Soricelli, F. Nobili, and et al., "Abnormalities of resting-state eeg in patients with prodromal and overt dementia with lewy bodies: Relation to clinical symptoms," *Clinical Neurophysiology*, vol. 131, no. 11, p. 2716–2731, Sep 2020.

[9] J. A. Christensen, J. Kempfner, M. Zoetmulder, H. L. Leonthin, L. Arvastson, S. R. Christensen, H. B. Sorensen, and P. Jennum, "Decreased sleep spindle density in patients with idiopathic rem sleep behavior disorder and patients with parkinson's disease," *Clinical Neurophysiology*, vol. 125, no. 3, p. 512–519, Mar 2014.

[10] J.-S. Sunwoo, K. S. Cha, J.-I. Byun, J.-S. Jun, T.-J. Kim, j.-w. Shin, S.-T. Lee, K.-H. Jung, K.-I. Park, K. Chu, M. Kim, S. Lee, H.-J. Kim, C. Schenck, and K.-Y. Jung, "Nrem sleep eeg oscillations in idiopathic rem sleep behavior disorder: A study of sleep spindles and slow oscillations," *Sleep*, vol. 44, 08 2020.

[11] J. A. Christensen, M. Cesari, F. Pizza, E. Antelmi, R. A. Frandsen, G. Plazzi, and P. Jennum, "Nocturnal eye movements in patients with idiopathic rapid eye movement sleep behaviour disorder and patients with parkinson's disease," *Journal of Sleep Research*, vol. 30, no. 3, Aug 2020.

[12] P. Bugalho, M. Mendonça, T. Lampreia, R. Miguel, R. Barbosa, and M. Salavisa, "Heart rate variability in parkinson disease and idiopathic rem sleep behavior disorder," *Clinical Autonomic Research*, vol. 28, no. 6, p. 557–564, Aug 2018.

[13] A. Malafeev, D. Laptev, S. Bauer, X. Omlin, A. Wierzbicka, A. Wichniak, W. Jernajczyk, R. Riener, J. Buhmann, and P. Achermann, "Automatic human sleep stage scoring using deep neural networks," *Frontiers in Neuroscience*, vol. 12, p. 781, 11 2018.

[14] M. Cesari, J. A. Christensen, F. Sixel-Doring, M.-L. Muntean, B. Mollenhauer, C. Trenkwalder, P. Jennum, and H. B. Sorensen, "A clinically applicable interactive micro and macro-sleep staging algorithm for elderly and patients with neurodegeneration," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019.

[15] G. Ruffini, I.-S. David, M. Castellano, L. Dubreuil Vall, A. Soria-Frisch, R. Postuma, J.-F. Gagnon, and J. Montplaisir, "Deep learning with eeg spectrograms in rapid eye movement behavior disorder," *Frontiers in Neurology*, vol. 10, p. 806, 07 2019.

[16] A. Narin, "Detection of focal and non-focal epileptic seizure using continuous wavelet transform-based scalogram images and pre-trained deep neural networks," *IRBM*, vol. 43, no. 1, p. 22–31, 2022.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[18] A. Lewkowycz, "How to decay your learning rate," 2021. [Online]. Available: https://arxiv.org/abs/2103.12682

[19] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*. ISCA, sep 2019. [Online]. Available: https://doi.org/10.21437\%2Finterspeech.2019-2680

[20] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, p. 1–15, Feb 2018.

[21] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and

O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," 2020. [Online]. Available: https://arxiv.org/abs/2009.07896

[22] A. Papp, A. Horváth, M. Virág, Z. Tóth, C. Borbély, F. Gombos, A. Szűcs, and A. Kamondi, "Sleep alterations are related to cognitive symptoms in parkinson's disease: A 24-hour ambulatory polygraphic eeg study," *International Journal of Psychophysiology*, vol. 173, pp. 93–103, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167876022000186

[23] Y. Zhang, R. Ren, L. D. Sanford, L. Yang, J. Zhou, L. Tan, T. Li, J. Zhang, Y.-K. Wing, J. Shi, and et al., "Sleep in parkinson's disease: A systematic review and meta-analysis of polysomnographic findings," *Sleep Medicine Reviews*, vol. 51, p. 101281, Feb 2020.

[24] M. Livia Fantini, J.-F. Gagnon, D. Petit, S. Rompré, A. Décary, J. Carrier, and J. Montplaisir, "Slowing of electroencephalogram in rapid eye movement sleep behavior disorder," *Annals of Neurology*, vol. 53, no. 6, p. 774–780, Jun 2003.

[25] N. Cooray, F. Andreotti, C. Lo, M. Symmonds, M. T. Hu, and M. De Vos, "Detection of rem sleep behaviour disorder by automated polysomnography analysis," *Clinical Neurophysiology*, vol. 130, no. 4, p. 505–514, Apr 2019.

[26] S.-Y. Gong, Y. Shen, H.-Y. Gu, S. Zhuang, X. Fu, Q.-J. Wang, C.-J. Mao, H. Hu, Y.-P. Dai, C.-F. Liu, and et al., "Generalized eeg slowing across phasic rem sleep, not subjective rbd severity, predicts neurodegeneration in idiopathic rbd," *Nature and Science of Sleep*, vol. Volume 14, p. 407–418, Mar 2022.

[27] J. Christensen, H. Koch, R. Frandsen, M. Zoetmulder, L. Arvastson, S. Christensen, H. Sorensen, and P. Jennum, "Sleep stability and transitions in patients with idiopathic rem sleep behavior disorder and patients with parkinson's disease," *Sleep Medicine*, vol. 16, Jan 2016.

[28] A. Brink-Kjær, M. Cesari, F. Sixel-Döring, B. Mollenhauer, C. Trenkwalder, E. Mignot, H. B. Sorensen, and P. Jennum, "Arousal characteristics in patients with parkinson's disease and isolated rapid eye movement sleep behavior disorder," *Sleep*, vol. 44, no. 12, Dec 2021.

[29] A. Brink-Kjaer, E. Mignot, H. B. Sorensen, and P. Jennum, "Predicting age with deep neural networks from polysomnograms," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 146–149.