Explainable AI for EEG Biomarkers Identification in Obstructive Sleep Apnea Severity Scoring Task

Luca La Fisca[1], Celiane Jennebauffe[1], Marie Bruyneel[2], Laurence Ris[3], Laurent Lefebvre[4], Xavier Siebert[5], Bernard Gosselin[1]
[1]*Department of Information Signal and Artificial Intelligence, University of Mons*, $Belgium$
[2]*Department of Pneumology, CHU Saint-Pierre*, $Brussels$, $Belgium$
[3]*Department of Neuroscience, University of Mons*, $Belgium$
[4]*Department of Cognitive Psychology and Neuropsychology, University of Mons*, $Belgium$
[5]*Department of Mathematics and Operational Research, University of Mons*, $Belgium$

***Abstract*—The assessment of Obstructive Sleep Apneas and hypopneas (OSAs) severity has known an increasing interest over the last decade with the use of Apnea-Hypopnea Index (AHI) being highly criticized by the majority of sleep scientists. To go beyond the single AHI, alternative metrics such as hypoxic burden, arousal intensity, odds ratio product, and cardiopulmonary coupling have been investigated in the literature. However, no consensus has currently been found for a common efficient metric. In this paper, we propose a novel architecture of deep learning model aiming at discovering an objective metric for OSAs severity assessment. We demonstrate the efficiency of this method by identifying features of interest in the Electroencephalographic (EEG) signals while training the model based on biomarkers not or indirectly derived from the EEG, i.e. the desaturation area, the arousal events and the respiratory event duration. By inspecting what the model looks for to make the different classifications, we identified that EEG signals from posterior and medial regions in low frequency bands (0-8 Hz) are highly affected by the apnea-hypopnea severity. With this proof of concept, we pave the way towards the use of Explainable Artificial Intelligence (xAI) to make OSAs severity assessment more objective and find a consensus metric adopted across the community of sleep scientists as well as to boost EEG biomarkers discovery in multiple tasks.**

***Index Terms*—EEG, obstructive sleep apnea, explainable AI, semi-supervised learning, proof-of-concept**

I. Introduction

Obstructive Sleep Apnea-hypopnea (OSA) is a common sleep disorder associated with multiple medical conditions from excessive daytime sleepiness to cognitive or cardiovascular disorders [1]. The assessment of how OSAs affect patients' health, i.e. its severity, is currently stemmed from the number of apnea and hypopnea events occurring overnight through the Apnea-Hypopnea Index (AHI) [2]. However, the use of AHI has been largely criticized over the last decade as it fails to estimate the impact of OSAs on related medical conditions [3]. This issue has triggered many works aiming to find better metrics to characterize OSAs severity, such as hypoxic burden, arousal intensity, duration of apneic events, odds ratio product, heart rate variability and cardiopulmonary coupling [4]. Despite all the efforts made in the direction of discovering the most efficient metric for OSAs severity, no consensus has been found across the sleep research community [4]. We therefore propose a novel approach to pave the way towards a common severity metric relying on Explainable Artificial Intelligence (xAI). The purpose of this work is to demonstrate the relevance of using explainable Deep Learning (DL) models to identify the features of importance for OSAs severity assessment task in order to get rid of the subjective biases the metrics proposed by clinicians suffer from. This demonstration is carried out by identifying Electroencephalographic (EEG) biomarkers considered of high importance by our DL model when performing a severity classification task defined by Polysomnographic (PSG)-derived features not or indirectly related to EEG signals. In fact, several research works have shown that OSA events trigger specific EEG power variations that differ between patients with severe OSA syndrome and patients with moderate one [5].

To the best of the authors' knowledge, xAI principles have never been applied to OSAs severity assessment. Most of the studies involving DL algorithms on sleep data aim at either automatically detect sleep stages or apnea-hypopnea events from PSG signals, distinguish OSAs from other sleep conditions such as insomnia [6], estimate a specific underlying symptom, e.g. excessive daytime sleepiness [7] or estimate AHI from arbitrary chosen signals like the oxygen saturation signal [8]. In this work, we rather propose to reduce the subjectivity of the diagnosis by inspecting how a DL algorithm makes its decisions.

xAI applied to medical data can be divided into two subtypes: inherent and post-hoc explainability. Models with inherent explainability relies on simple decision process, but struggle to consider higher order information and post-hoc explainability, often relying on saliency maps, suffers from the human tendency of ascribing overly positive interpretation. In this work, we introduce explainability in a human-centric approach, which relies on similarity between items in their entirety, instead of reducing it to statistics which rely on averages of specific and often inhuman features.

In this paper, we propose a proof of concept for a novel xAI approach applied to PSG data for OSAs severity assessment task by identifying EEG biomarkers using our xVAEnet model,

This research was funded by the "Fonds pour la Formation à la Recherche dans l'Industrie et l'Agriculture" (FRIA), Belgium.

a new DL architecture combining Convolutional Neural Network (CNN) for feature extraction, Variational Auto-Encoder (VAE) for dimension reduction, and Multilayer Perceptron (MLP) for classification. All the open-source codes and supplementary materials are available using the following link: *https://github.com/numediart/xVAEnet*.

II. Dataset

The dataset built for this research consists of PSG data of 72 patients who had undergone in-lab PSG ($\geq$ 8 hours) in 2022. The recordings, realized in the Sleep Laboratory, Centre Hospitalier Universitaire Saint-Pierre (CHU S^t-Pierre), Brussels, Belgium, have been manually annotated by clinicians to identify sleep stages, apnea and hypopnea events and arousal events according to international guidelines. All the selected patients exhibited excessive obstructive respiratory events (apneas or hypopneas) during the night, with at least AHI$\geq$5. The sleep onset was determined when the first epoch of sleep occurs. A preliminary sleep questionnaire was performed and the protocol CE/22-03-03 was approved by the local ethical comittee of the CHU S^t-Pierre on March 14th 2022. The PSG sensors are composed of 6 EEG electrodes, 2 Electrooculograph (EOG) electrodes (EOG1 under the left eye and EOG2 above the right eye) , thoracic and abdominal belts (VTH and VAB) to monitor respiratory motions, an Electrocardiogram (ECG) sensor, a pulse oxymetry sensor recording the pulse rate (PR) and the oxygen saturation (SAO2), and a pressure probe measuring the nasal airflow (NAF2P).
The raw signals were recorded at 200 Hz and downsampled at 50Hz for storage requirements using Medatec Brainnet Winrel 5.0 system. The data were then converted to Python-friendly files using the MNE-Python package [9], which was used to preprocess the signals. As our analysis focuses on the differences between apneic events, the studied database consists of OSA trials only, each of them corresponding to a 60 seconds segment extracted from the manually labeled signals and starting 4 seconds before an OSA event. The selected EEG signals were the 3 left-hand side electrodes, a frontal (FP1), a central (C3) and an occipital (O1), with the reference electrode being placed just above the nasion and the derivations being performed with a right mastoid electrode. These signals were divided into bands with a width of 2Hz, ranging from 0Hz to 10Hz. The 3 right-hand side electrodes were not analyzed for clarity and simplicity of this proof-of-concept research. The signals have been preprocessed, as described in the supplementary materials, for artifacts removal, patients exclusion and addition of Pulse Rate Variability (PRV) signal from the PR signal and phase shift (Pshift) signal from the VAB and VTH signals. The normalization has been performed by channel independently as a z-score normalization with clamping in the [-3; 3] range. After the preprocessing phase, the final dataset is composed of 6992 OSA trials of 23 channels (15 filtered EEG channels and 8 non-EEG PSG channels) and 3001 timestamps from 60 patients divided into a training set of 4660 trials from 48 patients, namely the trainset, and a validation set of 2332 trials from the 12 remaining patients, namely the testset.

III. xVAEnet Architecture

The principle of the proposed approach is to make sense of the feature space the classifier use to perform the desired severity classification task. In fact, when going forward in an artificial neural network, the input data undergoes different transformations across the hidden layers, this phase is called *feature extraction*. The purpose of these computed features is to be maximally discriminative for the categories that have to be classified. The last layer, the classifier, is often a fully-connected layer that allocates specific weights to the extracted features to make the best final classification.
In this work, the feature space is manipulated to acquire some required properties: 1) *reconstruction* ability, 2) *generative* ability, 3) *Gaussian* distribution.
The reconstruction ability ensures a direct relationship between the feature space and all the independent input channels, allowing clinicians to evaluate the relevance of the proposed classification. The generative ability ensures the model to be usable for new patients by avoiding part of the space not to be characterized. The Gaussian distribution allows a fair comparison between OSA trials by performing a directional study within the feature space, i.e. looking for the direction responsible for severity encoding. As shown in Figure 1, the aforementioned properties are added to the feature space by sharing the encoder part of the model between three sub-networks: a VAE, a Generative Adversarial Network (GAN) and a classifier. In fact, the VAE is responsible for bringing the reconstruction ability to the encoder latent space (Z_e), while the generative and the Gaussian distribution properties are part of the decoder latent space (Z_d). The GAN allows the transfer of the latter properties from Z_d to Z_e, and the classifier forces the separation of samples from different conditions in the latent space of interest (Z_e). The architecture details are presented in the supplementary materials.

A. VAE

As described by Kingma and Welling [10], a VAE model is designed to learn a latent representation of the input data with a desired probability distribution and generative ability. In this paper, the VAE model used is inspired by the *Stagernet* model proposed by Banville *et al.* to analyze long EEG sequences of sleep recordings [11]. The encoder part of our VAE is therefore a replica of the Stagernet CNN adapted to the 23x3001 format of our input data. The decoder is the mirrored version of the encoder where the convolutions are replaced by transposed convolutions and the max pooling layers by max unpooling ones. Contrarily to classical VAE models, the mean (μ) and standard deviation (σ) are not directly derived from the last convolutional layer of the encoder, but an intermediary latent vector is added to the encoder side (Z_e). In fact, as stated by Zhang *et al.*, the latent representations in VAEs are stochastically sampled from the prior distribution instead of being directly rendered from the input data [12]. This property compromises the further classification of the input data from their latent representation. The purpose of adding Z_e is therefore to make available a deterministic latent

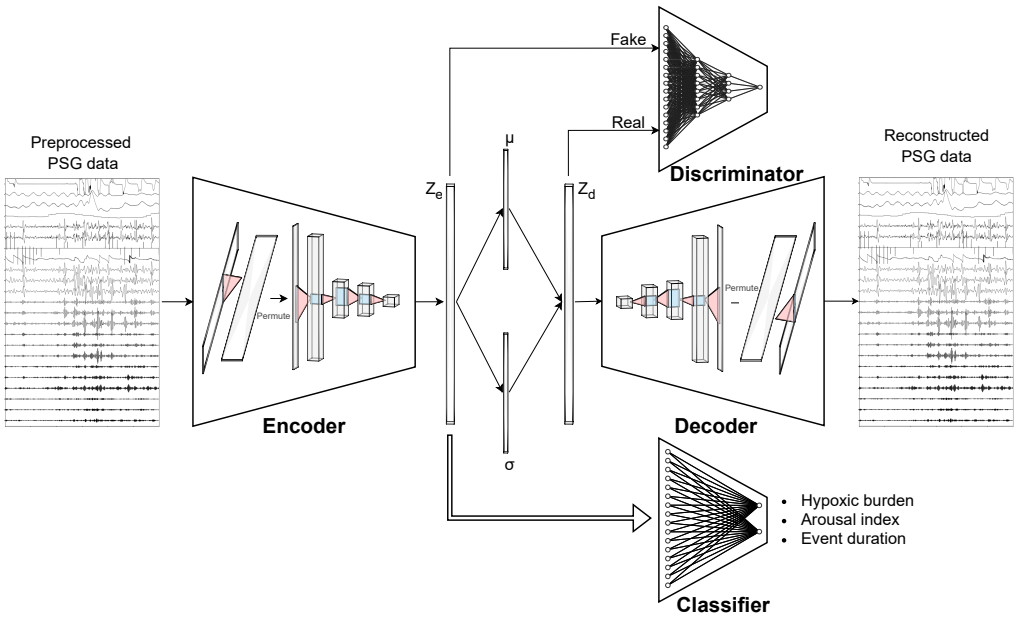

Fig. 1. xVAEnet Architecture. The model is composed of 3 parts: a VAE, a GAN and a classifier, all of them making use of the convolutional encoder that encodes the input data into an embedding, the encoder latent space (Z_e). The **VAE** (center) first estimates the mean (μ) and the standard deviation (σ) of the dataset distribution from Z_e using dense layers to obtain the decoder latent space (Z_d), then Z_d is decoded to derive a reconstructed version of the input data using a deconvolutional decoder. The **GAN** (top) exploits the encoder as a generator and discriminates Z_d (real distribution) from Z_e (fake distribution) using an MLP discriminator. The **Classifier** (bottom) uses the features extracted by the encoder in Z_e to classify the desaturation area, the arousal events and the respiratory event duration with a unique single-layer perceptron.

representation of the input data for the classification task. The decoder latent space (Z_d) is obtained by the reparameterization trick classically used in VAEs, i.e. $Z_d = \mu + \sigma\epsilon$ where ϵ is a random variable with Gaussian distribution. The loss function used to train the VAE part of our model is a combination of reconstruction loss and Kullback-Leibler divergence, as described in Equation 1:

$$\mathcal{L}_{VAE} = 0.5 \cdot MSE(output, input) + \\ 0.5 \cdot \frac{1}{bs} \sum_{i=1}^{bs} -0.5 \cdot \sum_{latent\_dim} (1 + log(\sigma) - \mu^2 - \sigma) \quad (1)$$

with *bs* the batch size and *MSE* the mean-squared error. As only Z_d acquires the generative ability and follows a Gaussian distribution during the VAE training phase, another process should transfer these properties to Z_e that is the purpose of the GAN module.

B. GAN

By considering Z_d as the real latent representation and Z_e as the fake one, the training of the GAN module will force the encoder to directly generate a latent vector Z_e that mimics the properties of Z_d, as inspired by Zhang *et al.* [12]. Adversarial networks requires a generator that generates fake data as close as possible to real data and a discriminator that differentiates between real and fake data. In the proposed xVAEnet architecture, the generator is the encoder shared with the VAE part and the discriminator consists of a 3-layer MLP each of them using leaky ReLU activation function with a negative slope of 0.2 and batch normalization. The output activation function is a sigmoid function. The loss function of the generator is a weighted sum of the VAE loss and the mean of correct predictions by the discriminator within a batch: $\mathcal{L}_{gen} = 0.2 \cdot \mathcal{L}_{VAE} + 0.8 \cdot \frac{1}{bs} \sum_{i=1}^{bs} (1 - fake_i)$ with *fake* the output of the discriminator when the fake latent representation (Z_e) is given as input, which should be equal to 1 for an ideal generator. The loss function of the discriminator is the difference between the mean fake predictions and the mean real predictions: $\mathcal{L}_{discrim} = \frac{1}{bs} \sum_{i=1}^{bs} fake_i - \frac{1}{bs} \sum_{i=1}^{bs} real_i$ with *real* the output of the discriminator when the real latent representation (Z_d) is given as input. For an ideal discriminator, *real* = 1 and *fake* = 0. The combination of both loss functions is described in Section IV.

C. Classifier

The classifier module forces the encoder to output a latent vector where trials presenting a low level of severity are maximally distant from trials exhibiting a high level of severity. In this work, the level of severity is characterized by the Desaturation Area (DA) (i.e., the area over the curve of the SAO_2 signal [13]), the arousal events (i.e., the presence or not of an arousal occurring just after the OSA) and the duration of the respiratory event. For clarity and simplicity of this proof-of-concept research, we simply defined two severity levels (low or high). Except for the arousal events that are binary by definition, we computed the median values of each severity features across trials and considering trials below the median value as low severity level trials and trials above the

median values as high severity level trials. Finally, we obtain four levels of severity: 1) *Very Low* (low severity level on all 3 features), 2) *Low* (high level on 1 feature), 3) *High* (high level on 2 features), 4) *Very High* (high level on all 3 features). The classifier block consists of a single-layer perceptron that performs a linear combination of the 128 latent vector values, outputting a probability for the sample to belong to each class via a softmax activation function, allowing the encoder to perform the majority of the classification task. As described in Section IV, the training process is done separately for each severity feature with the loss function being a binary cross-entropy loss: $\mathcal{L}_{classif} = BCE(predicted, target)$.

IV. Experiments

A. Training

The training process of the proposed model consists in a semi-supervised curriculum learning framework. In fact, every block of the architecture described in Section III is trained separately, and the initialisation of the following block's training process is done using the updated weights obtained at the end of the previous stage. The VAE and the GAN blocks are trained with non-supervised learning, while the classifier is trained in a supervised manner, making the whole model training semi-supervised. The training parameters are detailed in the supplementary materials. The VAE module has been trained using a random initialization until convergence. Then, the GAN module has been trained by initializing the generator with the best weights of the encoder obtained during the VAE training phase and the discriminator has been randomly initialized. At each batch, the discriminator is first trained by freezing the generator and using the loss function of the discriminator described in Section III, then the generator is trained by freezing the discriminator and using the corresponding loss function. Every 15 epochs, the updated network is used in inference to compute a new Z_d vector given as real input for the 15 following epochs in order to avoid the deterioration of the "real" space to be responsible for the increase of the GAN performance. Finally, the classifier module is initialized with the weights of the best generator previously obtained and the single-layer perceptron is randomly initialized. In the philosophy of curriculum learning, the classifier is trained on each severity feature sequentially, starting with the low vs. high severity classification on the DA, then on the arousal events and finally on the event duration. The learning rates were 10^{-3}, $5 \cdot 10^{-4}$, and $2 \cdot 10^{-4}$ for the first, second, and third stages, respectively. For each classification stage, a global loss, calculated every 5 epochs, combines the VAE, GAN, and classifier losses: $\mathcal{L}_{global} = \frac{1}{3}\mathcal{L}_{VAE} + \frac{1}{3}\frac{1}{bs}\sum_{i=1}^{bs}(1 - fake_i) + \frac{1}{3}\mathcal{L}_{classif}$. On the first stage, DA was classified without considering the other severity features. On the second stage, the classification has been performed on both the DA and the arousal events every 2 epochs: $\mathcal{L}_{classif_2} = \frac{1}{2} \cdot \mathcal{L}_{DA} + \frac{1}{2} \cdot \mathcal{L}_{arousal}$. On the third stage, the classification has been performed on all the severity features twice every 3 epochs: $\mathcal{L}_{classif_3} = \frac{1}{3} \cdot \mathcal{L}_{DA} + \frac{1}{3} \cdot \mathcal{L}_{arousal} + \frac{1}{3} \cdot \mathcal{L}_{duration}$.

B. Explainability

To explain how our xVAEnet model makes its decisions, we rely on a human-centric approach consisting in highlighting the similarities and differences between samples from different parts of the encoded latent space. In fact, the latent vector encoded by the encoder (Z_e) can be considered as a projection of the input data into a 128-dimensional space. This space being non-sparse owing to the Gaussian and generative properties, we can navigate through it in any direction to explore the specificity of each region. As the feature of interest of this work is the severity of the OSAs a patient undergoes, we have looked for the principal direction where the severity is encoded by performing a Linear Discriminant Analysis (LDA) on Z_e that maximizes the discrimination between the 4 classes of severity. The result of this process is a vector giving the direction of the severity encoding, namely the severity direction. By comparing input samples along the severity direction, we can highlight the channels that vary the most as well as the time windows that are the most affected by the OSAs severity. By analyzing non-EEG channels, we can validate that the model actually looks at the important features for severity scoring. By analyzing EEG channels, we can identify the best biomarkers of OSA in the EEG signals.

V. Results

The results of this research, essentially qualitative, can be divided in two parts: 1) the severity scoring efficiency and 2) the EEG biomarkers identification. All the quantitative results are detailed in the supplementary materials. In Figure 2, we can observe the evolution of the latent space distribution across the different training phases, allowing the qualitative evaluation of how Z_e acquires the required properties. For illustration purpose, the 128-dimensional latent space has been projected to a 2D space using the t-distributed stochastic neighbor embedding (t-SNE) transform. As shown in Figure 2A, the training process of the VAE module leads to a sparse encoder latent space (Z_e). The training process of the GAN module leads to a non-sparse Z_e getting closer to a Gaussian distribution, but with the samples of different severity scores randomly distributed (Figure 2B). The training process of the classifier module leads to a non-sparse, quasi-Gaussian Z_e where the samples of similar severity level tend to be gathered together and separated from samples of different severity levels, as illustrated in Figure 2C. From this well-designed latent space, we have performed an LDA aiming at classifying the 4 severity levels. With the classifier module trained, this LDA reaches a mean accuracy of 54.0% (trainset) and 48.8% (testset), and a mean F1-score of 56.5% (trainset) and 48.4% (testset). The direction of highest severity score variance, namely the *severity direction*, is represented with an arrow in Figure 2C and is responsible for 78.14% of the explained variance. This ability to estimate the severity score from a trial representation in Z_e is the first proof of the relevance of the proposed framework in severity scoring task. By navigating along the severity direction, we can sort the OSA trials by severity score to generate a severity scale

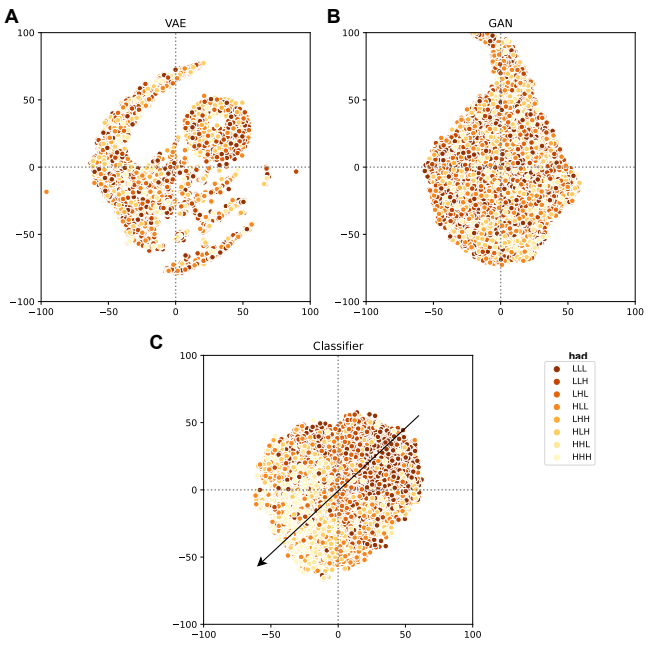

Fig. 2. 2D representations of the encoder latent space (Z_e) using t-SNE. Each sample represents one of the 6992 OSA trials. (**A**) Z_e with the training phase of the VAE module completed. (**B**) Z_e with the training phase of the GAN module completed. (**C**) Z_e with the training phase of the classifier module completed on every severity feature. The arrow represents the severity direction obtained using LDA. In the legend, each letter of "**had**" represents a severity feature: "**h**ypoxic burden", "**a**rousal event", and "**d**uration of the respiratory event". The "L" means "Low-level severity", the "H" means "High-level severity"

and compare the trials depending on their position on this scale. Figure 3A provides a summary of the influence of the severity score on each PSG channels based on their power signal. This comparison is performed by computing the mean power difference of each channel separately as described in Equation 2:

$$Pdiff_{c[i]\ dist[j]} = \frac{1}{N-j}\sum_{k=0}^{N-j} P_{c[i]\ t[k+j]} - P_{c[i]\ t[k]}$$
$$Pdiff_{c[i]} = \frac{1}{N}\sum_{j=0}^{N} Pdiff_{c[i]\ dist[j]} \quad (2)$$

with trials being sorted based on their severity score, *N* the number of trials, *t* the trial number, *c* the channel and *dist* being the distance on the severity scale. The second operation allowing the evaluation of the severity scoring efficiency consists in identifying PSG channels and time windows that vary the most with the severity score. The non-EEG PSG are used to evaluate the consistency between clinical studies and the proposed framework, while the EEG channels allow the biomarkers discovery. In Figures 3A and B, the high positive power difference on the SAO2 signal suggests deeper and/or longer desaturations of severe OSA trials, as stated by Kulkas *et al.* [13]. The high negative power difference on the EOG signal is consistent with the works of Eiseman *et al.* who showed the dependence of apnea severity on REM vs. non-REM sleep stage (that highly affects the eye movement) [14]. Furthermore, Figure 3C shows that the SAO2 effect mainly appears during the respiratory events (beginning of the trial) with a spurious peak effect around 50 seconds after the start of the event (note that OSA event starts after 4s as described in Section II). The aforementioned results provide the desired second proof that the proposed framework actually extracts severity information.

The EEG biomarkers identification task is based on the information provided by Figures 3D and E where we can observe that the central electrode (C3) is the most affected by the severity of the respiratory event in the 2-8Hz frequency range, this effect being maximal in the 5-25s trial time window (corresponding to the mean respiratory event duration). The occipital electrode (O1) also varies with the severity score in the 2-8Hz frequency range, but the frontal one (FP1) seems not to be influenced by the OSA severity. The findings indicate a decrease in EEG power in parieto-occipital regions as the severity score increases. Further investigation, utilizing high-density EEG studies, may support the interpretation of this decrease as a reduction in brain activity during severe OSA events.

VI. Discussion

This research is a proof-of-concept work aiming at demonstrating the ability of xAI to identify EEG biomarkers related to a specific task. The studied experimental task is the severity scoring of Obstructive Sleep Apneas-Hypopneas (OSAs) from PSG signals. The proposed framework relies on a human-centric explainability approach based on the comparison between samples to maximize the interpretability of the results. Our xVAEnet model is composed of 3 modules (VAE, GAN and classifier) trained sequentially using a semi-supervised curriculum learning process with the objective of encoding the input data into a latent feature space maximizing the discriminability between samples of different OSA severity levels. This framework provides a latent space that: 1) contains the majority of the input signals information, 2) is usable with new patients, 3) allows a fair comparison between OSA trials through directional study. The identified "severity direction" in this encoded space is responsible for 78% of the severity score explained variance, demonstrating the ability of the model to generate a well-suited distribution for the studied task. The EEG features that have been identified as OSA severity biomarkers are the central and occipital electrodes in the 2-8Hz frequency range, which is consistent with Jones *et al.* who have shown a decreased EEG power in the parietal region, especially in slow-wave activity (1-4.5Hz) and θ band (4.5-8Hz) [15]. In this proof-of-concept study, we mainly focused on qualitative results rather than quantitative ones as the purpose of this research is to demonstrate the relevance of the proposed xAI approach for biomarkers discovery, rather than obtaining the best performing model for the studied task. In the future works, we will seek for more detailed EEG biomarkers by considering: 1) time-frequency EEG representations (Fourier

Fig. 3. Biomarkers identification performed by comparing the power signal, by channel, of the OSA trials sorted by severity score ($\in [0,1]$) along the severity direction obtained using LDA. (**A**) Mean power difference across OSA trials obtained by subtracting, for each channel separately, the power signal of each trial from the power signal of trials of higher severity scores. (**B**) Channel-by-channel mean power difference of PSG channels excluding EEG channels. The x axis represents the distance, along the severity direction, between the trials being compared. A distance of 0 means a trial is compared to itself, a distance of 1 means the comparison between the trial of lowest severity score and the trial of highest severity score. (**C**) Time window-by-time window mean power difference of the SAO2 channel (channel of highest absolute mean power difference). (**D**) Channel-by-channel mean power difference of EEG channels. (**E**) Time window-by-time window mean power difference of the C3 channel on the 4-6Hz frequency band (channel of highest absolute mean power difference).

and Wavelet transforms), 2) all the available electrodes, 3) more precise severity definition by using regression and adding other severity features such as AHI. We will also improve the scoring efficiency by exploring other encoder architectures and hyperparameters (latent space dimension, batch size, dropout, weight decay, etc.).

References

[1] A. S. Jordan, D. G. McSharry, and A. Malhotra, "Adult obstructive sleep apnoea," *The Lancet*, vol. 383, no. 9918, pp. 736–747, Feb. 2014.

[2] R. B. Berry, R. Budhiraja, D. J. Gottlieb, D. Gozal, C. Iber, V. K. Kapur, C. L. Marcus, R. Mehra, S. Parthasarathy, S. F. Quan, S. Redline, K. P. Strohl, S. L. D. Ward, and M. M. Tangredi, "Rules for Scoring Respiratory Events in Sleep: Update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events," *Journal of Clinical Sleep Medicine*, vol. 08, no. 05, pp. 597–619, 2012.

[3] D. A. Pevernagie, B. Gnidovec-Strazisar, L. Grote, R. Heinzer, W. T. McNicholas, T. Penzel, W. Randerath, S. Schiza, J. Verbraecken, and E. S. Arnardottir, "On the rise and fall of the apneahypopnea index: A historical review and critical appraisal," *Journal of Sleep Research*, vol. 29, no. 4, 2020.

[4] A. Malhotra, I. Ayappa, N. Ayas, N. Collop, D. Kirsch, N. Mcardle, R. Mehra, A. I. Pack, N. Punjabi, D. P. White, and D. J. Gottlieb, "Metrics of sleep apnea severity: beyond the apnea-hypopnea index," *Sleep*, vol. 44, no. 7, p. zsab030, Jul. 2021.

[5] S. Puskás, N. Kozák, D. Sulina, L. Csiba, and M. T. Magyar, "Quantitative EEG in obstructive sleep apnea syndrome: a review of the literature," *Reviews in the Neurosciences*, vol. 28, no. 3, pp. 265–270, Apr. 2017.

[6] M. Younes, A. Azarbarzin, M. Reid, D. R. Mazzotti, and S. Redline, "Characteristics and reproducibility of novel sleep EEG biomarkers and their variation with sleep apnea and insomnia in a large community-based cohort," *Sleep*, vol. 44, no. 10, Oct. 2021.

[7] S. Nikkonen, H. Korkalainen, S. Kainulainen, S. Myllymaa, A. Leino, L. Kalevo, A. Oksenberg, T. Leppänen, and J. Töyräs, "Estimating daytime sleepiness with previous night electroencephalography, electrooculography, and electromyography spectrograms in patients with suspected sleep apnea using a convolutional neural network," *Sleep*, vol. 43, no. 12, Dec. 2020.

[8] G. C. Gutiérrez-Tobal, D. Álvarez, A. Crespo, F. del Campo, and R. Hornero, "Evaluation of Machine-Learning Approaches to Estimate Sleep Apnea Severity From At-Home Oximetry Recordings," *IEEE Journal of Biomedical and Health Informatics*, Mar. 2019.

[9] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers in Neuroscience*, vol. 7, 2013.

[10] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," May 2014, arXiv:1312.6114.

[11] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort, "Uncovering the structure of clinical EEG signals with self-supervised learning," *Journal of Neural Engineering*, Mar. 2021.

[12] X. Zhang, L. Yao, and F. Yuan, "Adversarial Variational Embedding for Robust Semi-supervised Learning," in *Proceedings of the 25th ACM International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA, Jul. 2019, pp. 139–147.

[13] A. Kulkas, P. Tiihonen, P. Julkunen, E. Mervaala, and J. Töyräs, "Novel parameters indicate significant differences in severity of obstructive sleep apnea with patients having similar apnea–hypopnea index," *Medical & Biological Engineering & Computing*, vol. 51, no. 6, pp. 697–708, 2013.

[14] N. A. Eiseman, M. B. Westover, J. M. Ellenbogen, and M. T. Bianchi, "The Impact of Body Posture and Sleep Stages on Sleep Apnea Severity in Adults," *Journal of Clinical Sleep Medicine*, 2012.

[15] S. G. Jones, B. A. Riedner, R. F. Smith, F. Ferrarelli, G. Tononi, R. J. Davidson, and R. M. Benca, "Regional Reductions in Sleep Electroencephalography Power in Obstructive Sleep Apnea: A High-Density EEG Study," *Sleep*, vol. 37, no. 2, pp. 399–407, Feb. 2014.