Towards a Deeper Understanding of Sleep Stages through their Representation in the Latent Space of Variational Autoencoders

Luka Biedebach*
Reykjavik University
lukab@ru.is

Matias Rusanen*
University of Eastern Finland
matias.rusanen@uef.fi

Benedikt Hólm Þórðarson
Reykjavik University
benediktht@ru.is

Erna Sif Arnardóttir
Reykjavik University
ernasifa@ru.is

María Óskarsdóttir
Reykjavik University
mariaoskars@ru.is

Sami Nikkonen
University of Eastern Finland
sami.nikkonen@uef.fi

Henri Korkalainen
University of Eastern Finland
henri.korkalainen@uef.fi

Sami Myllymaa
University of Eastern Finland
sami.myllymaa@uef.fi

Juha Töyräs
University of Eastern Finland
juha.toyras@kuh.fi

Samu Kainulainen
University of Eastern Finland
samu.kainulainen@uef.fi

Timo Leppänen
University of Eastern Finland
timo.leppanen@uef.fi

Anna Sigridur Islind
Reykjavik University
islind@ru.is

Abstract

Artificial neural networks show great success in sleep stage classification, with an accuracy comparable to human scoring. While their ability to learn from labelled electroencephalography (EEG) signals is widely researched, the underlying learning processes remain unexplored. Variational autoencoders can capture the underlying meaning of data by encoding it into a low-dimensional space. Regularizing this space furthermore enables the generation of realistic representations of data from latent space samples. We aimed to show that this model is able to generate realistic sleep EEG. In addition, the generated sequences from different areas of the latent space are shown to have inherent meaning. The current results show the potential of variational autoencoders in understanding sleep EEG data from the perspective of unsupervised machine learning.

1. Introduction

During sleep, we wander through different stages, characterized by certain physiological features. These features and their temporal variation is traditionally recorded in polysomnography (PSG), which is a multi-signal sleep study based on multiple sensors. The results of the PSG outline the gold standard diagnostic method for many sleep disorders (Arnardottir, Islind, & Óskarsdóttir, 2021; Schmitz et al., 2022). One feature that varies significantly between different physiological sleep stages is the brain's electrical activity, recorded with electroencephalography (EEG). The EEG outlines a vital part of the PSG enabling scoring of sleep stages with the inclusion of eye movements and chin muscle tone (Berry et al., 2018). Currently in clinical practice, sleep technologists classify 30-second epochs of PSG recordings into five sleep stages; wakefulness (Wake), three non-rapid eye movement sleep (Stages N1, N2, and N3) and rapid eye movement (REM) sleep. The classification is done according to the rules set by the American Academy of Sleep Medicine (AASM) (Berry et al., 2018). However, the current five-stage and 30-second epochs process is a simplification that is needed to alleviate the workload of manual sleep staging, and both aspects lack a complete scientific justification (Himanen & Hasan, 2000). Therefore, the details of underlying feature variation of the complex sleep EEG recordings remains a subject of research. In this paper, we propose a method to explore the relationship between scored sleep stages and physiological sleep stages.

State-of-the-art machine learning models such as deep convolutional neural networks (CNNs) are capable of classifying sleep stages with similar reliability as sleep technologists (Perslev et al., 2021; Korkalainen et al., 2019; Phan & Mikkelsen, 2021; Fiorillo et al., 2019). This is a major achievement for sleep research in general and has the potential to reduce the manual workload in clinical practice. However, these models rely on supervised learning using labelled sleep recordings (Korkalainen et al., 2019). As a result, they express high classification accuracies but are limited to repeating the manual sleep staging which they are

URI: https://hdl.handle.net/10125/103013
978-0-9981331-6-4

trained with, in an automatic manner. In addition, the learning process and the used features are often untraceable and difficult to visualize.

Samek et al. pointed out, that due to the lack of transparency at the machine learning models, we can neither verify them nor learn from them (Samek, Wiegand, & Müller, 2017). Consequently, there is a rising demand toward explainable artificial intelligence (XAI) (Gerlings, Shollo, & Constantiou, 2020), i.e. machine learning models that not only provide an output but also enable the understanding how the output was achieved (Shaban-Nejad, Michalowski, Brownstein, & Buckeridge, 2021; Linardatos, Papastefanopoulos, & Kotsiantis, 2021; Rudin, 2019). There is a long withstanding discussion of the need for unpacking technology (Orlikowski, Iacono, et al., 2001; Kallinikos, 2002), and more specifically, on unpacking artificial intelligence (AI) and moving away from the black-box mentality (Castelvecchi, 2016).

In this paper, we aim to take a first step towards making machine learning models in sleep research more traceable, which is strongly needed especially in the healthcare sector. In sleep, just like in any other medical application of machine learning, the decisions made by a model come with a high responsibility, as they directly affect the health of a patient. For this reason, XAI helps medical professionals to gain trust and increase the actual usage of those systems (Xie, Gao, & Chen, 2019). Lately, generative machine learning models have been helpful in XAI through visualizations (Kahng, Thorat, Chau, Viégas, & Wattenberg, 2019). One of the most studied type of generative models is a Variational Autoencoder (VAE), a specifically structured generative autoencoder (Kingma & Welling, 2019). VAEs have been used for example to generate interpretable features of electrocardiography (ECG) (Kuznetsov, Moskalenko, Gribanov, & Zolotykh, 2021). Moreover, VAEs have increased the classification accuracy of EEG-based speech recognition systems (Krishna, Co, Carnahan, & Tewfik, 2020).

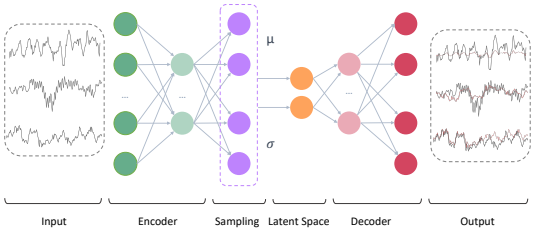

Figure 1. General Architecture of a Variational Autoencoder

Based on the previous findings, we hypothesize that VAEs have the potential to learn the underlying feature variations of sleep EEG recordings. In this proof of concept study, we aim to generate realistic sleep EEG using VAEs. In addition, we aim to show that VAEs can make an interpretable latent space using sleep EEG inputs. We furthermore discuss how this method could pave the way for a deeper understanding of sleep stages.

2. Related Work

Previously, autoencoders have been used in the context of sleep staging as a preprocessing step or as an unsupervised classifier. Najdi et al. used an autoencoder to learn a compact feature vector of PSG data in a sleep stage classification algorithm (Najdi, Gharbali, & Fonseca, 2017). Moreover, Perslev et al. utilized a typical architecture of convolutional autoencoders in another supervised sleep stage classification model (Perslev et al., 2021). Similarly, Prabhudesai et al. developed a method to automatically learn features from the raw EEG data with an autoencoder, which were then used to cluster the data to different sleep stages (Prabhudesai, Collins, & Mainsah, 2019). Autoencoders can also be used for unsupervised pre-training before supervised classification as shown by Wei et al. (Wei, Zhang, Wang, & Dang, 2018). In this study, we do not want to outperform the previous models in terms of classification accuracy but rather deepen the understanding of the sleep stages by investigating the properties learned by the autoencoder.

Variational autoencoders have shown their ability to create a meaningful latent space in other domains such as language processing (Song, Sun, Chen, Peng, & Song, 2019), image generation (Razavi, Van den Oord, & Vinyals, 2019), and cancer diagnosis (Way & Greene, 2018). In the medical field, VAEs are used to gain an understanding of ECG data (Kuznetsov et al., 2021). VAEs can also be used for emotion recognition based on EEG (Li et al., 2020) and extracting features for speech recognition on EEG data (Krishna, Tran, Carnahan, & Tewfik, 2020). To the best of our knowledge, VAEs have not been applied to sleep EEG data before as a generative model.

3. Theoretical Background

3.1. Variational Autoencoders

Autoencoders are artificial neural networks, which encode the data into a latent space and then decode it as closely as possible back into its original shape. It is a reconstruction-based form of representation learning, since the model is trained by comparing the reconstructed output with the original input (Bengio, Courville, & Vincent, 2013). The fundamental

concept of autoencoding lies within the autoencoder's architecture, consisting of an encoding function, an intermediary latent space, and a decoding function as illustrated in Figure 1 (Goodfellow, Bengio, Courville, & Bengio, 2016).

VAEs make an addition to this architecture by adding a probabilistic manipulation to the latent space variables (Kingma & Welling, 2013). In VAEs, the encoder's architecture comprises two fully connected layers connected into two latent vectors. The vectors' elements represent the mean and variance of a normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ for each latent space dimension. Furthermore, the encoder comprises a sampling layer, which maps the measures of the probability distribution into the final latent space samples. These samples also compose the input of the decoder (Spinner, Körner, Görtler, & Deussen, 2018). In contrast to normal autoencoders, VAEs also function as generative models. The generative nature of VAEs emerges from the sampling layer, which enables the sampling of the probabilistic latent space, as well as from the decoder, that can be used to generate reconstructions from the latent space samples (Kingma & Welling, 2013).

Although the latent space has a simple probabilistic nature, a reconstruction loss-based optimization alone can lead to an overly complex, non-continuous, and unorganized latent space structure. In this case, generated representations of the latent space can be hard to interpret or completely unrealistic (Kingma & Welling, 2013). Therefore, VAEs introduce a regularization term in the total loss of the model. This term is added to the reconstruction loss and controls the structure of the latent space during optimization. The total loss is therefore a combination of two parts, i.e.

$$\text{Total loss} = \text{reconstruction loss} + \text{regularization},$$

where reconstruction loss is usually the mean squared error (MSE) or mean absolute error (MAE) between the input and the output of VAE for one-dimensional signals (Kuznetsov et al., 2021; Krishna, Tran, et al., 2020).

In the case of VAEs, the regularization term is defined using Kullback-Leibler (KL) divergence, which is a statistical distance measure between two distributions (Kullback & Leibler, 1951). The distance is computed in each iteration of weight optimization between distributions of the latent space samples $X \sim \mathcal{N}(\mu, \sigma^2)$ and a unit normal distribution $I \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$. Thus, by minimizing the KL divergence, we force the latent probability distributions to follow a normal distribution, making the latent space more organized and continuous (Kingma & Welling, 2013). The total loss can be written as follows:

$$\begin{aligned}\text{Total loss} = \text{MSE}(\text{input}, \text{reconstruction}) \\ +\text{KL}(X, I).\end{aligned}$$

Because of the difference in dimensionality between the input data and the latent space, the MSE and KL loss are averaged before summation. The additional regularization enables the creation of a continuous and potentially meaningful latent space, but reduces the autoencoder's ability to accurately create reconstructions (Asperti & Trentin, 2020). Nevertheless, we can adjust the balance between good reconstructions and more continuous latent space (Alemi et al., 2018). However, better reconstructions come with the cost of possibly overlapping latent space clusters and noisier encodings. One of the methods used in balancing between these two factors is called β-VAE (Higgins et al., 2017). This method multiplies the KL loss with a constant β. It has also been shown that monotonically or cyclically updating the β value increases the performance of VAEs as well as helps with an easily vanishing KL term (Fu et al., 2019).

As explained, the latent space of the VAE is more continuous in contrast to the sparse latent space created in normal autoencoders. As the latent space is distributed around the origin and generally shares a similar value range, a valid output can be generated from decoding points in the latent space (Spinner et al., 2018). Due to these special properties of the latent space in variational autoencoders, the newly generated samples and their position in the latent space become interpretable.

3.2. Convolutional Layers

The advantage of convolutional neural networks (CNN) is that they extract visually meaningful information. Even though CNNs are most commonly used for image processing, they have also shown to be a suitable approach for transforming EEG data (Bashivan, Rish, Yeasin, & Codella, 2015). A convolutional layer of a CNN slides a kernel of a filter over the input to extract features at each position. A filter is therefore a stack of matrices, the kernels, which factors are learned during training (O'Shea & Nash, 2015). The kernel size defines the size of the sliding window which is passed over the data. Smaller kernels tend to collect more local information, while larger kernels extract the global, high-level features (Gu et al., 2018). CNNs usually comprise multiple convolutional layers with a different number of filters and different kernel sizes. In this way, the architecture of the CNN is constructed to extract information on multiple scales. Furthermore,

using convolutional layers in VAEs, the size of the input can be gradually decreased towards the latent space to reduce dimensionality while extracting information.

4. Method

4.1. Data

For this paper, we used 50 PSG recordings, which totals in 381.13 hours of EEG data. The data collection was approved by the National Bioethics Committee of Iceland (21-070). Informed written consent was obtained from all participants before measurements. We have a diverse study population with 27 male, 19 female, and 4 unspecified-gendered participants. The study population included participants with and without diagnosed sleep disorders. More information about the study population can be found in Table 1.

Table 1. Demographic information of the study population (n=50)

Variable	Mean ± SD
Age [years]	44.2 ± 13.4
Weight [kg]	84.1 ± 21.7
Height [cm]	174.9 ± 9.9
BMI [kg/m^2]	27.3 ± 5.3
AHI [1/h]	12.0 ± 13.2

SD = standard deviation, BMI = body mass index, AHI = apnea-hypopnea index

The PSG recordings were conducted at Reykjavik University as part of the Sleep Revolution project. The PSG was set up by a professional sleep technologists and the participants slept at home in their natural sleeping environment. The Type II PSG recordings were conducted using a portable PSG device (Nox A1, Nox Medical, Reykjavik, Iceland) and included EEG as recommended by the AASM (Berry et al., 2018). We used the F4-M1 channel from the EEG recordings as a single-channel input to the VAE, as it is commonly used in manual sleep staging. Only one channel was used to keep the feature variation of the input EEGs reasonable. For visualization and exploration of the latent space, we used the manual scoring of sleep stages, which was conducted by an experienced sleep professional according to the AASM scoring manual (Berry et al., 2018).

4.2. Preprocessing

The EEG signals were originally saved using 200 Hz sampling frequency in the Noxturnal (Nox Medical) software and exported to EDF format. The signals were then preprocessed with Python according to the following steps. First, we downsampled the signals to 64 Hz to reduce the computational burden and complexity of EEG signals. Second, we applied high-pass filters with a cut-off frequency of 0.3 Hz, as recommended in the AASM scoring manual (Berry et al., 2018). Finally, we scaled the signal amplitudes into a range between 0 and 1 using min-max scaling. We confirmed that the EEG signals appeared normal after each preprocessing step as illustrated in Figure 2. These preprocessing steps were conducted per subject to preserve the amplitude variation in each recording.

The recording has a length of approximately 7 hours per participant. To work with this data in a machine learning context, we split it into smaller 10-second sub-sequences. Sleep stages were manually scored in 30 second windows, but we chose the time window of 10 seconds to reduce the length of the time series processed by the VAE. The 10-second segments were randomly divided into training (90%) and testing (10%) sets, resulting in 357.9 hours or 128857 segments of EEG data for training and 39.8 hours or 14318 segments of EEG data for testing. In unsupervised machine learning the division to train and test sets is not mandatory but we chose to include it for experimental reasons.

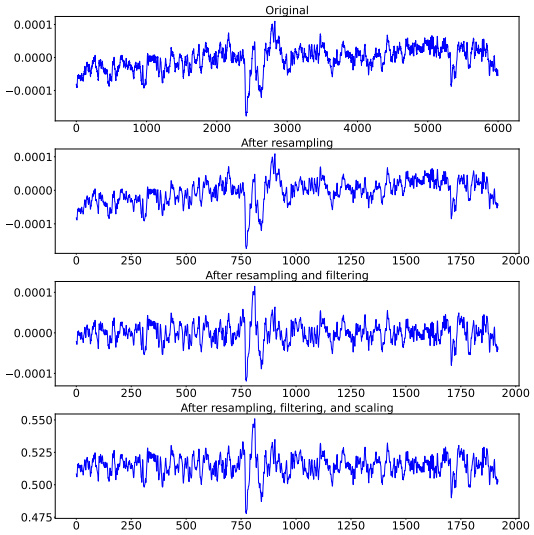

Figure 2. A 30-second sample of the input EEG after each of the preprocessing steps.

4.3. Optimization

The models were implemented using TensorFlow version 2.8.0 (Abadi et al., 2015) and Keras application programming interface. We optimized the VAEs using the Adaptive Moment Estimation (Adam) algorithm (Kingma & Ba, 2015) with default Keras configurations.

The weights were updated in batches of size 100 until the total loss converged or until 100 training epochs.

During the training of the VAE, both the reconstruction error as well as the KL divergence were taken into account. A common problem with VAEs is the KL divergence collapse problem (Asperti & Trentin, 2020; Alemi et al., 2018), which arises from unequal scales of the reconstruction error and the KL divergence. To ensure a balance between them, we used the β-VAE method.

5. Experiments

5.1. Dense VAE

As a proof-of-concept of VAEs operating with sleep EEG data, we ran experiments on the most simple VAE architecture. This architecture comprised only a single dense connection between the input and latent space parameters as well as the latent space sample and the output of the decoder. Furthermore, we used the input size of the signal (640 samples) as the latent space dimension to further increase the simplicity of our method. We increased the weight of reconstruction loss in the total loss using constant $\beta = 0.0001$ multiplying the KL term. The learning rate was decreased from 0.01 with 0.001 steps after each iteration until optimization stopped or the learning rate reached a value of 0.001.

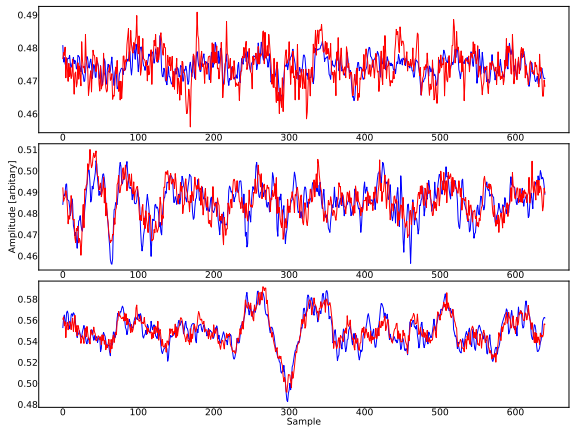

Figure 3. Three exemplary inputs of 10-second EEG segments (blue) and their reconstructions (red) using the dense model.

Our experiment to reconstruct EEG with a simple dense VAE clearly showed the ability of VAEs to work with highly complex EEG inputs. The reconstructions shown in Figure 3 were achieved after 100 epochs of training. Despite the desired reconstructions, this model was unsuitable for the intention to explore the sleep EEG data through latent space, as the latent space had no dimension reduction relative to the input data.

5.2. CNN VAE

Moving from high-dimensional latent space to reducing the dimensions into something that can be visualized, we chose to experiment with three dimensions in the latent space. Following the major change in the dimensional reduction of input data, the purpose was not to reach similar reconstructions as shown with our simple dense model. Instead, we experimented with whether the VAE can still extract features relative to input data and generate realistic EEG samples. For meaningful feature extraction, we included a CNN layer in the VAE's architecture. Keeping the experiment simple, only one convolutional or respectively deconvolutional layer was added to both the encoder and decoder. We used 256 filters with kernels of size 5 for the convolutions. Then, we reduced the length of the sequence with max-pooling. In the following layer, we flattened the sequence into one dimension, before connecting it to the latent vectors. Here, instead of forwarding the exact position in the latent space to the decoder, the mean and variance of a normal distribution were used to sample a position in the latent space. At this point the data was compressed to a vector of length three, which was then transformed back into its original shape by the following layers.

In the decoder, we used a dense layer which mirrored the transformation of the sampling and a reshape layer that mirrored the transformation of the flattening layer in the encoder. Then, an up-sampling layer was used to mirror the max-pooling. Finally, a deconvolutional layer with one filter and kernel size 5 brought the data back into their original shape. Both in the encoder and in the decoder, we used Rectified Linear Units (ReLU) as activation functions. For the optimization of this model, we gradually increased the weight of the KL divergence from $\beta = 0.01$ to $\beta = 1$ in 100 epochs. In this manner, the model should first learn the reconstructions, after which the latent space is made regular (Higgins et al., 2017). A constant learning rate of 0.001 was used for optimization of this model. In the following Results section, we refer to the results achieved with the CNN VAE.

6. Evaluation

6.1. Turing Test

The Turing test is an experimental set-up developed by Alan Turing to test the intelligence of a machine (Turing, 2009). Originally, the test was designed to

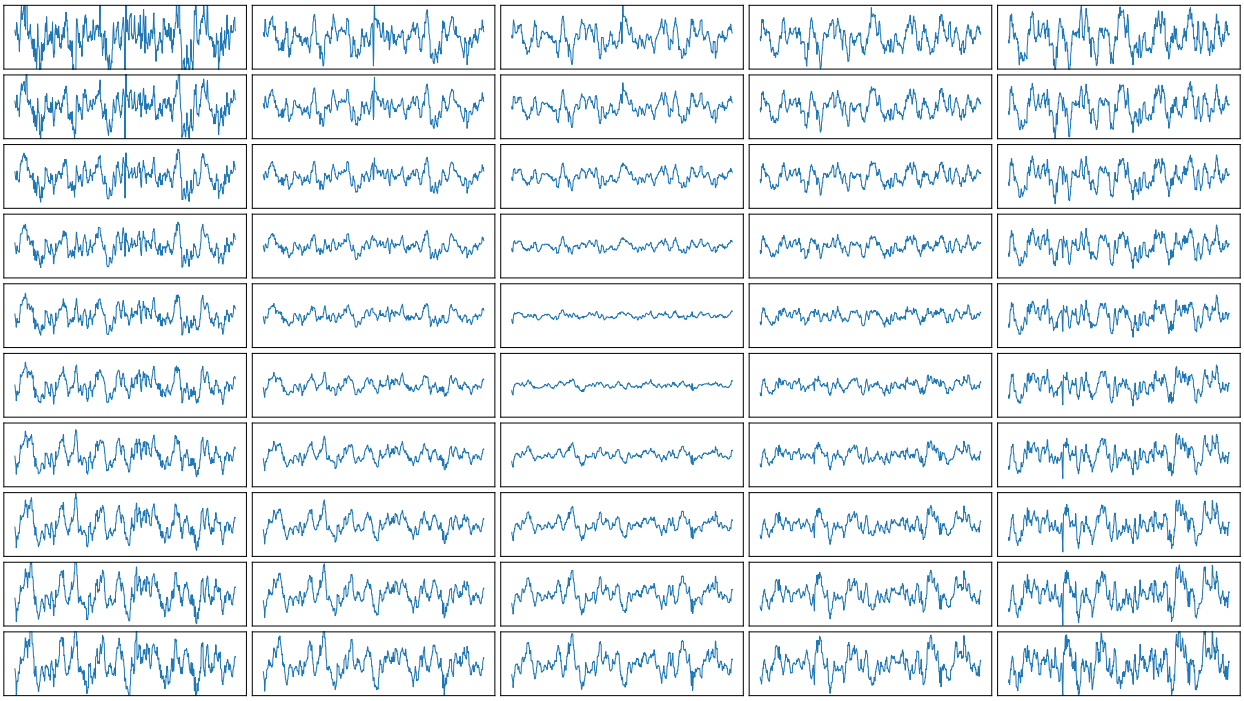

Figure 4. Generated artificial EEG segments (10 seconds) sampled with linear intervals from the first and the third axis of the latent space, keeping the second dimension coordinate constantly as zero.

determine whether an interrogator can distinguish a human from an AI in a dialogue with both of them. We used the principles of this test by confronting a sleep technologist with both real and artificially created EEG sequences. This way, we evaluated whether the sequences generated by our VAE were realistic.

We sampled 10-second segments randomly from the input EEG signals. In addition, we created artificial EEG segments by randomly sampling each of the latent space coordinates from a uniform distribution between -3 and 3 and passing the resulting point to the decoder. In the first step, the totalling 50 signal segments were then distributed randomly on a 5x10 grid including 46 real and four artificial EEG sequences. A sleep expert was asked to point the four artificial EEG sequences out. In a second step, we confronted the sleep technologist with a 3x6 grid (18 sequences) including likewise four artificially created EEG sequences.

6.2. Manual Review

In order to verify that our model could not only generate realistic EEG sequences, but also created a meaningful latent space, we manually reviewed the generated sequences with the sleep technologist. In this experiment, we showed the sleep technologist two maps of artificial EEG segments on a 5x10 grid sampled and decoded from the latent space. One map showed samples from the first and second axis of the latent space, while a second map showed the first and third axis. In both maps, the excluded dimension was kept at a constant value of zero. For visualisation purposes, we extracted the mean from each segment and fixed the y-axis limits of the subplots to be constant. To gain an understanding of different axes, sleep technologist was asked to give an estimation of the sleep stage of different EEG sequence. We furthermore asked for the presence of sleep stage specific patterns and artifacts.

7. Results

The Turing test-like experiment showed that the sleep technologist was not able to identify any of the four artificial EEG sequences from the real examples in a set-up of 50 sequences. Also, in a set-up of 18 sequences, the sleep technologist was unable to identify the four artificial ones. From this, we conclude that our VAE can generate realistically looking EEG sequences.

Using the CNN VAE, the generated artificial EEG sequences showed different features according to the latent space position they originated from. Figure 4 shows a map of points sampled from the first and third axis of the latent space, keeping the second axis coordinate constant at zero. The variation of the

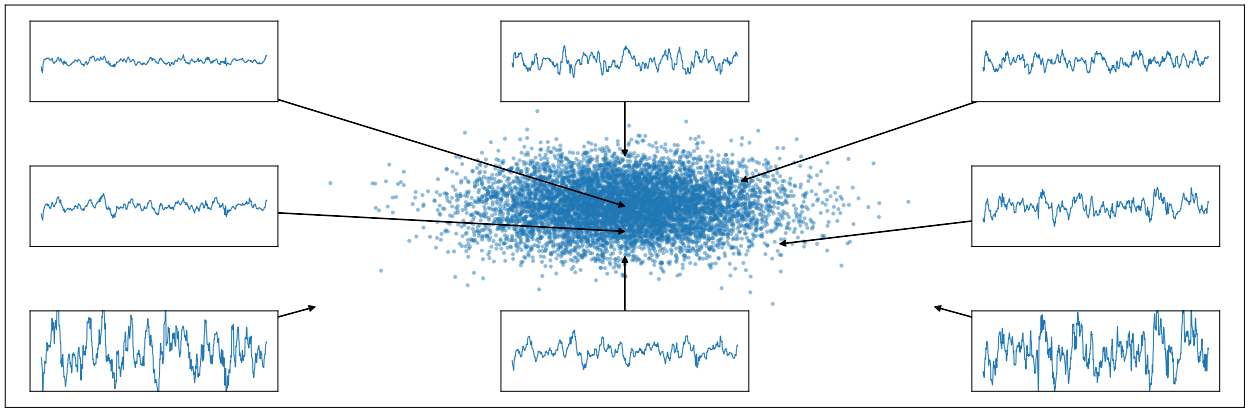

Figure 5. Signal segments (10 seconds) sampled from the latent space axes 0, and 2.

sequences sampled from different positions were clearly visible. We can observe lower amplitude signals to originate from the center of the latent space while amplitudes increased when increasing values of the first and third axis. In some parts of the latent space the VAE also generated sequences that do not look like EEG at all or resemble artifacts. These sequences might arise due to noncontinuous area of latent space or might be learned from artifacts in the training data.

The sleep technologist confirmed by manual review that samples from certain positions along axes of the latent space resemble certain sleep stages. It needs to be noted, that this was not a proper scoring according to the AASM rules, but instead a subjective estimation based solely on the shape of the signal. However, this variation showed that the samples were not randomly generated, and that the axis of the latent space contains meaning and reflects features typical for different sleep stages. Figure 5 shows samples generated from exemplary positions in the latent space. In this visualization, the second axis was held at a constant value of 0. The sequence in the top left corner was perceived as REM or N1 sleep by the sleep technologist, while the sample in the bottom left corner was perceived as deep sleep (N3).

Figure 6 shows clustering of the training data along the first axis of the latent space. In the three-dimensional visualization of the latent space, the sleep stages were slightly organized into clusters. However, depending on which axis was visualized, more clusters not related to the sleep stages become visible.

8. Discussion

In this paper, we aimed to show that VAEs can be applied to sleep EEG data. This is an important contribution to the field of information technology in sleep research as this model proposes novel methods to generate insights into sleep structure. The main novelty of this work is that it follows the principles of XAI, by making the latent space interpretable and the learning process traceable. Secondly, the proposed model is fully unsupervised, and hence does not require any manually scored data nor carry the bias introduced by manual scoring during the training.

The present results indicate that VAEs are able to generate realistic but synthetic samples of EEG with varying features that are common to sleep EEG data. In addition, we showed that the created latent space was not random, but reflected features of different sleep stages in different positions along the axes. Finally, the results illustrate that a simple convolutional VAE was capable of generating preliminary clustering of the EEG data in the latent space. Therefore, we suggest that this method could open a way to understand sleep stages in a more sophisticated manner than previously achieved through manual analysis and unpack some of the criticized mystery related to AI. It also shows preliminary potential of comparing unsupervised sleep staging to supervised sleep staging and manual sleep staging through labeling of the latent clusters.

The artificial EEG sequences generated with the VAE were not distinguishable from real EEG segments to an expert sleep technologist. Although this observation contains the bias of a subjective opinion of a single sleep technologist, it shows that the artificial sequences can be considered realistic. It furthermore indicates that even the simple neural network model can learn some features representing the input data. It should also be noted, that the method is scalable to studying neural networks of different architectures in the encoder and decoder. This method could therefore help in understanding some of the already existing sleep staging models (Perslev et al., 2021; Korkalainen et al.,

2019; Phan & Mikkelsen, 2021; Fiorillo et al., 2019).

The variation within our latent space, representing features that might be attributed to different sleep stages, shows how the model built an understanding of the EEG sequences in an interpretable way. The high complexity inherent in the combination of thousands of weights within the neural network prevents us from fully tracing what is learned during the training. However, the continuous latent space created by the VAE is a sophisticated visual approach to understand the features that are learned by the model. In order to get a more holistic evaluation of the methodology, it would require an in-depth analysis by multiple sleep technologists. That is however outside the scope of this paper. In addition, more quantitative analyses of the generated EEG signals are needed in the future.

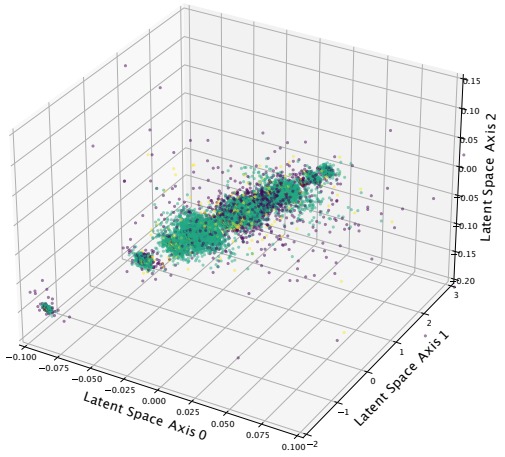

Figure 6. Embedding of EEG sequences used in training within the three-dimensional latent space, colored by sleep stage. Values on the axes represent means of the latent space samples. Colors: Yellow = REM, Green = N3, and Purple = Wake.

Compressing the data into a three-dimensional and continuous latent space comes with a certain cost. First, the low dimensionality creates a tight bottleneck in which inevitably information is lost. Second, the sampling layer introduces randomness to the model. Hence, the autoencoder faces a trade-off between accurate reconstructions of the input signals (as seen in the dense model) and a meaningful latent space suitable for generating data (as seen in the CNN model). Regarding the high complexity of our input signal, a sequence of a multi-frequency biosignal with a length of 640 samples, it is hard to achieve accurate reconstructions even with the dense model that operated without any dimensionality reduction. However, the purpose of the model is not to perfectly reconstruct compressed input signals, but to create a latent space from which realistic EEG can be generated. From this perspective, the reconstructions are not an issue, as the individual peaks and troughs are irrelevant for realistic EEGs, while the general frequency and recurring patterns matter more (Berry et al., 2018).

Moving away from manual review, we can also perceive an irregular distribution of sleep stages in the latent space when using the manual sleep stage scoring as labels. The slight clustering of sleep stages within the latent space hints towards further possibilities for unsupervised sleep staging. However, to achieve this, more sophisticated models that better capture the feature variation of input EEGs are likely needed. One possibility could be a concatenated model with separate branches of convolutional and dense layers, as proposed in (Kuznetsov et al., 2021). However, adding complexity to the model might make the optimization of the model more difficult. This in turn highlights the need for more adaptive optimization methods such as VAEs with calibrated decoder (Rybkin, Daniilidis, & Levine, 2020) or VAEs utilizing single-parameter, continuous Bernoulli distributions (Loaiza-Ganem & Cunningham, 2019). Furthermore, some clusters we observed in the latent space, which were not related to sleep stages might represent other factors, e.g. patient demographics. In order to study this assumption, other attributes such as age, gender, and sleeping disorders need be considered in the future. Another method could be splitting the data before training into subgroups based on other attributes and comparing the latent spaces that are created. Especially training one model on recordings by participants with obstructive sleep apnea (OSA) and another model on participants without any sleep disorders could reveal differences in their EEG features. Before these experiments are possible, more methodological studies on using VAE with sleep EEG data need to be conducted.

9. Conclusion

We can conclude, that this paper is preliminary work that explored the possibilities of VAEs in sleep research, opening up several new research directions in the future. This study contributes an addition to traditional machine learning-assisted sleep research in the following ways: i) by introducing a method for generating realistic artificial EEG; ii) by showing potential of providing in-depth understanding of sleep EEG and sleep staging through XAI, and; iii) by creating the foundation for attempting unsupervised sleep staging through clustering in the latent space. Our findings are relevant for the field of sleep research and health information systems in general because we show how a VAE can act as a generative and

interpretable model for EEG data. Generating realistic EEG sequences is not only relevant for sleep research but can also be used as a method in various medical domains, and as such, apply to a variety of health information systems issues. We hope that introducing XAI in sleep research could increase the acceptance and usage of AI systems by sleep professionals in the hospital and beyond.

10. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 965417. We thank Sigríður Sigurðardóttir for the manual review of the generated EEG sequences. The first two authors Luka Biedebach and Matias Rusanen contributed equally and share the first authorship. The senior author of this paper is Anna Sigridur Islind.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* (Software available from tensorflow.org)

Alemi, A. A., Poole, B., Fische, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2018). Fixing a broken ELBO. *35th International Conference on Machine Learning, ICML 2018*, *1*, 245–265.

Arnardottir, E. S., Islind, A. S., & Óskarsdóttir, M. (2021). The future of sleep measurements: A review and perspective. *Sleep medicine clinics*, *16*(3), 447–464.

Asperti, A., & Trentin, M. (2020). Balancing reconstruction error and Kullback-Leibler divergence in variational autoencoders. *IEEE Access*, *8*(1), 199440–199448.

Bashivan, P., Rish, I., Yeasin, M., & Codella, N. (2015). Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Lloyd, R. M., Quan, S. F., ... Vaughn, B. V. (2018). *AASM Manual for the Scoring of Sleep and Associated Events* (Tech. Rep.). Amer. Acad. Sleep Med., Darien, IL, USA: American Academy of Sleep Medicine.

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature News*, *538*(7623), 20.

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P. L., Favaro, P., Roth, C., ... Faraci, F. D. (2019). Automated sleep scoring: A review of the latest approaches. *Sleep Medicine Reviews*, *48*, 101204.

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., & Carin, L. (2019). Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*.

Gerlings, J., Shollo, A., & Constantiou, I. (2020). Reviewing the need for explainable artificial intelligence (xai). *arXiv preprint arXiv:2012.01007*.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1) (No. 2). MIT Press Cambridge.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... others (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Lerchner, A. (2017). β-VAE: Learning basic visual concepts with a constrained variational framework. In *Iclr* (Vol. 44).

Himanen, S.-L., & Hasan, J. (2000). Limitations of rechtschaffen and kales. *Sleep medicine reviews*, *4*(2), 149–167.

Kahng, M., Thorat, N., Chau, D. H. P., Viégas, F. B., & Wattenberg, M. (2019). GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 310–320.

Kallinikos, J. (2002). Reopening the black box of technology artifacts and human agency.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *CoRR*, *abs/1412.6980*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, *12*(4), 307–392.

Korkalainen, H., Leppanen, T., Aakko, J., Nikkonen, S., Kainulainen, S., Leino, A., ... Toyras, J. (2019). Accurate Deep Learning-Based Sleep Staging in a Clinical Population with Suspected Obstructive Sleep Apnea. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 2073–2081.

Krishna, G., Co, T., Carnahan, M., & Tewfik, A. H. (2020). Constrained Variational Autoencoder

for improving EEG based Speech Recognition Systems.

Krishna, G., Tran, C., Carnahan, M., & Tewfik, A. (2020). *Constrained variational autoencoder for improving eeg based speech recognition systems.* arXiv.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Kuznetsov, V., Moskalenko, V., Gribanov, D., & Zolotykh, N. Y. (2021). Interpretable feature generation in ecg using a variational autoencoder. *Frontiers in genetics*, *12*.

Li, X., Zhao, Z., Song, D., Zhang, Y., Pan, J., Wu, L., … Wang, D. (2020). Latent factor decoding of multi-channel eeg for emotion recognition through autoencoder-like neural networks. *Frontiers in Neuroscience*, *14*.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1).

Loaiza-Ganem, G., & Cunningham, J. P. (2019). The continuous bernoulli: Fixing a pervasive error in variational autoencoders. *Advances in Neural Information Processing Systems*, *32*(NeurIPS), 1–11.

Najdi, S., Gharbali, A. A., & Fonseca, J. M. (2017). Feature transformation based on stacked sparse autoencoders for sleep stage classification. In *Doctoral conference on computing, electrical and industrial systems* (pp. 191–200).

Orlikowski, W. J., Iacono, C. S., et al. (2001). Desperately seeking the "it" in it research–a call to theorizing the it artifact. *Information systems research*, *12*(2), 121–134.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., & Igel, C. (2021). U-Sleep: resilient high-frequency sleep staging. *npj Digital Medicine*, *4*(1), 1–12.

Phan, H., & Mikkelsen, K. (2021). *Automatic Sleep Staging: Recent Development, Challenges, and Future Directions.*

Prabhudesai, K. S., Collins, L. M., & Mainsah, B. O. (2019). Automated feature learning using deep convolutional auto-encoder neural network for clustering electroencephalograms into sleep stages. In *2019 9th international ieee/embs conference on neural engineering (ner)* (pp. 937–940).

Razavi, A., Van den Oord, A., & Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, *32*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Rybkin, O., Daniilidis, K., & Levine, S. (2020). Simple and effective VAE training with calibrated decoders.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Schmitz, L., Sveinbjarnarson, B. F., Gunnarsson, G. N., Davidsson, O. A., Davidsson, T. B., Arnardóttir, E. S., … Islind, A. S. (2022). Towards a digital sleep diary standard. In *Americas conference on information systems.*

Shaban-Nejad, A., Michalowski, M., Brownstein, J., & Buckeridge, D. (2021). Guest Editorial Explainable AI: Towards Fairness, Accountability, Transparency and Trust in Healthcare. *IEEE Journal of Biomedical and Health Informatics*, *25*(7), 2374–2375.

Song, T., Sun, J., Chen, B., Peng, W., & Song, J. (2019). Latent space expanded variational autoencoder for sentence generation. *IEEE Access*, *7*, 144618–144627.

Spinner, T., Körner, J., Görtler, J., & Deussen, O. (2018). Towards an interpretable latent space : an intuitive comparison of autoencoders with variational autoencoders. In *Proceedings of the workshop on visualization for ai explainability 2018 (visxai).*

Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the turing test* (pp. 23–65). Springer.

Way, G. P., & Greene, C. S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In *Pacific symposium on biocomputing 2018: Proceedings of the pacific symposium* (pp. 80–91).

Wei, R., Zhang, X., Wang, J., & Dang, X. (2018). The research of sleep staging based on single-lead electrocardiogram and deep neural network. *Biomedical engineering letters*, *8*(1), 87–93.

Xie, Y., Gao, G., & Chen, X. (2019). Outlining the design space of explainable intelligent systems for medical diagnosis. *arXiv preprint arXiv:1902.06019*.