Noise-Resilient and Interpretable Epileptic Seizure Detection

Anthony Hitchcock Thomas, Amir Aminifar, David Atienza
Embedded Systems Laboratory (ESL)
Swiss Federal Institute of Technology Lausanne (EPFL)
{ahthomas, amir.aminifar, david.atienza}@epfl.ch

***Abstract*—Deep convolutional neural networks have recently emerged as a state-of-the art tool in detection of seizures. Such models offer the ability to extract complex nonlinear representations of an electroencephalogram (EEG) signal which can improve accuracy over methods relying on hand-crafted features. However, neural networks are susceptible to confounding artifacts commonly present in EEG signals and are notoriously difficult to interpret. In this work, we present a neural-network based algorithm for seizure detection which leverages recent advances in information theory to construct a signal representation containing the minimal amount of information necessary to discriminate between seizure and normal brain activity. We show our approach automatically learns representations that ignore common signal artifacts and which encode medically relevant information from the raw signal.**

I. Introduction

Epilepsy is a common neurological disorder affecting over 65 million people worldwide[1]. Epilepsy is characterized by chronic seizures which typically range in duration from a few seconds to several minutes and may induce loss of memory or consciousness [1]. While a variety of treatment options exist, they are not effective in a substantial fraction of patients [1], [2]. In such cases, one avenue to reduce mortality risk and improve quality of life is real-time monitoring by wearable devices which can detect seizure events and issue a warning to caregivers. Such devices continuously gather electroencephalogram (EEG) readings from patients and apply a signal processing algorithm in an attempt to classify a reading as corresponding to a seizure or normal brain activity.

Substantial research attention has been devoted both to the engineering problem of designing such devices and the algorithmic problem of developing signal processing algorithms capable of reliably detecting seizure events [3]–[6]. Prior work on detection algorithms generally falls into two categories: algorithms which use domain expertise to hand select "features" to extract from signals, and methods which attempt to automatically extract a good representation of the raw signal. In the former category, the literature has proposed a wide variety of features such as various entropy measures [7]–[9] and features of the power spectral density [10], [11]. Feature extraction based algorithms may be desirable because they can explicitly take advantage of established medical research when constructing features and may thus generally be considered more "biologically plausible" or "interpretable." On the other hand, feature extraction typically discards information from the signal which may be relevant for prediction. Accordingly, there has been growing research interest in methods which automatically extract a "good" representation of the raw signal.

For instance, [12] proposed a method based on the Karhunen-Loéve Transform (KLT) that represents the raw signal in an alternate basis where classification based on a similarity search can be efficiently performed without the need for feature extraction. More recently, there has been growing interest in the use of deep neural networks (DNNs) for seizure detection and EEG processing more generally [13]–[15]. DNNs are particularly appealing because they can learn a very rich space of transformations to the raw signal and are excellent for capturing "hierarchical features" which are composed of motifs occurring at different resolutions [16]. Our focus in this work is on neural network based methods for detection.

While neural networks are a powerful class of models, they suffer from a number of shortcomings which complicate their application to seizure detection. In particular, EEG recordings are typically noisy and include numerous artifacts, such as eye blinks or short changes in visual attention, which perturb the raw signal but may be unrelated to the outcome of interest (e.g. presence or absence of a seizure) [15]. The presence of nuisances can increase the number of training examples required to learn model parameters, and complicates interpretation of the signal representation since it may incorporate both relevant and irrelevant aspects of the signal [17]. Artifact removal has been extensively studied in the literature and is regarded as an essential step in the analysis pipeline [18], [19]. However, these procedures generally must be applied as pre-processing steps which complicates analysis and risks discarding useful information from the signal along with artifacts.

Interpreting the signal representation extracted by a DNN is challenging since the internal representation is a complex non-convex function of the raw signal values. While interpretability of DNN models has emerged as an important topic of research in the general machine learning community in recent years, relatively little work has addressed this issue to date in the context of seizure detection. Prior work in [15], [20] demonstrated that convolutional neural networks encode frequency domain information from the raw signal, which is well known

[1] This work has been supported by the DeepHealth Project (GA No. 825111), by the MyPreHealth research project (Hasler Foundation project No. 16073), and by the Swiss NSF ML-Edge Project (GA No. 182009).

to be medically significant [21]. However, their analysis is restricted to networks used to classify gestures from EEG and relies on computing pairwise correlations between frequency domain features and output activations of convolution filters.

In this work we present a deep-learning-based algorithm for seizure detection which mitigates the issues discussed above. We leverage recent advances in probabilistic interpretations of deep learning models to construct a minimally sufficient representation of the signal for discriminating between seizure and normal brain activity. Building on recent work in the Information Theory community, we demonstrate that this formulation automatically learns to ignore irrelevant signal artifacts eliminating the need for artifact removal during pre-processing. Furthermore, we analyze the latent representations learned by our approach and demonstrate they encode signal features known from medical research to be relevant for seizure detection. We emphasize that our goal in this work is not to improve on state-of-the-art accuracy for seizure detection. Rather, we wish to illustrate that adopting an information theoretic perspective on the problem provides advantages over conventional approaches without sacrificing accuracy.

II. Problem Statement and Solution

A. Notation

Bold uppercase symbols denote matrices while bold lowercase symbols denote vectors. Standard font uppercase symbols denote random variables and standard font lowercase symbols denote scalars. We denote the EEG recording for a patient by $\{(\mathbf{x}_t, y_t))\}_{t=1}^{T}$ where t indexes samples (time), T is the total number of samples, $\mathbf{x}_t \in \mathbb{R}^k$ is the raw k channel EEG recording, and $y_t \in \{0,1\}$ is a binary variable equal to one if time t corresponds to a seizure. The y_t values are obtained by hand annotation by an expert but may be subject to some degree of imprecision (up to a few seconds). As is common in EEG analysis [13]–[15], we group individual EEG samples into short time windows of length l denoted $\{(\mathbf{X}_w, y_w)\}_{w=1}^{W}$ where $W = \lceil \frac{T}{l} \rceil$ and $\mathbf{X}_w \in \mathbb{R}^{l \times k}$. We declare $y_w = 1$ if *any* $\mathbf{x}_t \in \mathbf{X}_w$ corresponds to a seizure. We denote by $p(X,Y)$ the joint distribution over $\mathbf{X}_w$ and y_w.

B. Problem Statement

As noted above, a significant issue in analysis of EEG signals is the presence of artifacts which confound the signal. We therefore seek a representation $\mathbf{z}_w \in \mathbb{R}^k$ for $\mathbf{X}_w$ that preserves only the information in $\mathbf{X}_w$ which is *relevant* for discriminating between seizure and normal brain activity. This problem is known as the "Information Bottleneck" and can be written formally as [22]–[24]:

$$\max_{p(Z\,|\,X)} I(Z\,;\,Y) \text{ s.t. } I(Z\,;\,X) \leq \delta \qquad (1)$$

which leads to the following Lagrangian:

$$\mathcal{L}_{IB}(p(Z|X)) = I(Z;Y) - \beta I(Z;X) \qquad (2)$$

Where $I(A;B)$ denotes the mutual information between random variables A and B.

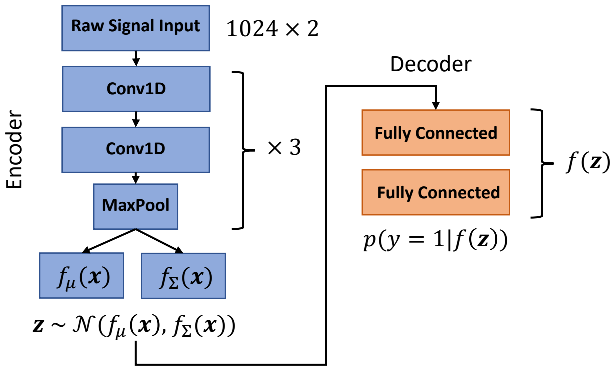
Raw Signal Input
1024 × 2
Encoder
Conv1D
Conv1D
MaxPool
× 3
$f_\mu(\boldsymbol{x})$
$f_\Sigma(\boldsymbol{x})$
$\boldsymbol{z} \sim \mathcal{N}(f_\mu(\boldsymbol{x}), f_\Sigma(\boldsymbol{x}))$
Decoder
Fully Connected
Fully Connected
$f(\boldsymbol{z})$
$p(y = 1|f(\boldsymbol{z}))$

Fig. 1. **Variational Information Bottleneck Network Architecture**

We may think of the IB problem as a two stage process: a representation (or "codeword") $\mathbf{z}$ is sampled from an encoder distribution $p(Z|X)$ over which the maximization is performed, and an outcome $\hat{y}$ is sampled from a corresponding "decoder" distribution $p(Y|Z)$ [22].

C. Variational Bound

Unfortunately, computing: $I(X;Z)$ and $I(Z;Y)$ is - in general - computationally intractable, and 2 can only be solved exactly in the case that X, Y are discrete or jointly Gaussian [24]. However, Equation 2 can be estimated empirically from sample data by minimizing the following variational upper bound [24]–[26]:

$$\mathcal{L}_{IB} \leq \hat{\mathcal{L}}_{IB}(\theta, \phi) = \frac{1}{W} \sum_{w} \mathbb{E}_{\mathbf{z} \sim q}[-\log q_\theta(y_n \,|\, \mathbf{z}_n)] \qquad (3)$$
$$+ \beta D_{KL}(q_\phi(Z|\mathbf{x}_n) \,||\, p(Z))$$

Where $D_{KL}(q \,||\, p)$ is the Kullback-Leibler (KL) divergence between p and q. This formulation is known as the "Variational Information Bottleneck" (VIB). An example is presented in Figure 1. The "encoder network" - shown in blue - takes the original signal as input and computes the mean and covariance matrix of a multivariate Gaussian distribution. A codeword $\mathbf{z}$ is then sampled from this distribution and used as input by the decoder network - shown in orange - which computes the probability that the (binary) outcome was a 1. Figure 1 shows the network architecture used in this work, but the method is generic with respect to the method used to compute f_μ and f_Σ.

Intuitively, the first term in Equation 3 tries to maximize the accuracy of the prediction. The "encoder" distribution $q_\phi(Z|X)$ tries to map similar input signals to similar regions of the latent space, while the "prior" $p(Z)$ acts as a regularizer which encourages the codewords to be distributed over a high-dimensional ball and helps prevent overfitting [27]. The Lagrange multiplier β controls the tradeoff and must be set as a *hyperparameter*. In general, the bound in Equation 3 will not be tight. A limitation of the method is that a rigorous assessment of the bound is difficult and empirical methods (e.g. cross-validation) must be used instead.

D. *Invariance to Signal Artifacts*

The encodings $\mathbf{z}$ obtained by training a neural network to minimize Equation 2 have the useful property of being maximally invariant to "nuisance transformations" to the raw signal [25], [27]. Let N be a random variable such that $I(N;X) > 0$ but $I(N;Y) = 0$; that is N induces some perturbation to the input signal but is unrelated to the outcome of interest. A common example in the context of EEG analysis is ocular artifacts (eye blinks). Then amongst all representations of the signal, the $\mathbf{z}$ obtained by training a neural network to minimize Equation 2 minimizes $I(Z;N)$ - in other words, the representation is maximally invariant to nuisances. A proof is given in [27]. In Section III-C we demonstrate empirically that training a DNN to minimize Equation 3 allows the signal representation to automatically ignore information about artifacts resulting from eye blinks.

III. Experimental Analysis

A. *Data*

We use a publicly available dataset provided by the "PhysioNet" EEG database [28]. The dataset consists of surface EEG recordings gathered from 10 patients with drug resistant epilepsy. We here use two EEG channels located at the left and right temple to be consistent with the environment faced by developers of embedded systems for seizure detection [5]. For each patient, we randomly sample two seizures to serve as test data and one to serve as validation data for hyperparameter tuning. The remainder are reserved for training. We discard two patients with less than four seizures. To account for variability due to the sampled test set we repeat each experiment 25 times and report mean values.

Consistent with prior work in the domain, we partition seizure events into windows of eight seconds in length and advance the window by one-half second at each training observation. Because the vast majority of the signal corresponds to normal brain activity, we randomly sample eight second windows of normal brain activity to obtain a training set with balanced positive and negative examples. While the exact number of training and testing examples varies depending on the randomly sampling, the train and test set both consist of approximately $5,000$ observations while the validation set contains approximately $2,500$ observations.

To mitigate bias, we resample the "non-seizure" windows every three epochs during training. We rescale the data in each channel to have zero mean and unit variance but do not pre-process the signal in any other way. We use the "Adam" optimizer [29], and train until validation cost fails to decrease for thirty successive epochs or until 600 total epochs. At the conclusion of training, the weights which yield the lowest validation cost are retained.

B. *Model Architecture and Training*

As noted above, we parameterize the encoder distribution as:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{z}|f_\mu(\mathbf{x}), f_\Sigma(\mathbf{x})) = f_\mu(\mathbf{x}) + f_\Sigma(\mathbf{x}) \odot \epsilon \qquad (4)$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\odot$ denotes element-wise multiplication. Intuitively, we use a deep neural network to compute the mean and variance of a factorized (e.g. diagonal) Gaussian distribution and then sample codewords from this distribution. Thus, our network is inherently stochastic in nature. We emphasize that this stochasticity is essential to the properties of the representation [30].

A diagram of our network architecture is shown in Figure 1. The network consists of three identical 1-dimensional convolution blocks. Each block consists of two convolution layers each with eight filters using a kernel size of three and a stride of one followed by a 1-dimensional max pooling layer. Following the convolution layers, the mean and (diagonal) log-covariance matrix of the Gaussian are computed using a fully connected layer with a linear activation. The decoder is a simple multi-layer-perceptron with 2 fully connected layers. All models are implemented using TensorFlow/Keras and use the Adam optimizer [29], [31]. We tune hyperparameters for learning rate, decay, and L2-regularization using grid search.

C. *Invariance to Nuisance Factors*

We here investigate the practical ability of the VIB to learn signal representations which are invariant to EEG signal artifacts. In practice the VIB bound will not be tight and there is expected to be "leakage" of irrelevant information about X to Z. We here use a test proposed in [25] to measure the invariance of the extracted representations to nuisances. As a case study, we consider invariance to perturbations caused by eye blinks but emphasize that the method is agnostic to the specific type of nuisance which need not be specified.

We consider two neural networks trained for seizure detection. The first is our VIB model as described in section III-B. The second is a baseline model which uses the same architecture but replaces the stochastic layer of the encoder with a standard fully connected layer. Accordingly, we train the baseline model only to minimize the cross-entropy loss. We here fix the dimension of the latent space at $d = 4$ as we found little difference in results between models with d ranging from 2 to 16.

Our hypothesis is that the representations learned by the VIB model convey less information about signal artifacts - in this case whether or not the subject is blinking - than do those of the standard neural network. To test this, we extract the latent representation for each window (e.g. the $\mathbf{z}_n$) and train a secondary classifier - a random forest - to discriminate between windows which contain a blink and those which do not. While our data is not labeled with blinks, we obtain "ground-truth" data by applying a publicly available MATLAB blink-detection algorithm to the raw signal [32]. If the latent representations convey information about whether or not the subject is blinking then we would expect better than chance performance on this task.

Results are presented in Table I. The first column presents test accuracy on the primary task of seizure detection while the second presents test accuracy on the secondary task of blink detection. We use a student-t test to determine whether the observed differences in mean are statistically significant [33]. Table I reports p-values from a test of the null hypothesis that the observed difference between the baseline model performance and the VIB model is due to chance against the alternative that it is not.

Results are consistent with our hypothesis that minimizing the VIB cost function yields representations which automatically ignore "nuisance" information in the raw signal. We find that blink-detection models trained on the latent features extracted from VIB models reduce accuracy on the blink detection task by up to 12.3% relative to the baseline, and reject the hypothesis that this difference is due to chance. Furthermore, we find that accuracy on the nuisance task decreases as β increases which is consistent with the interpretation of β as controlling flow of information into the latent representation. We find *no statistically significant difference* in accuracy on the primary task of seizure detection between any of the VIB and baseline models. We regard these results as evidence that the VIB cost function can be used to reduce the confounding effects of nuisances without significantly reducing model accuracy.

TABLE I
ACCURACY ON BLINK DETECTION TASK

	Accuracy			
Model	**Seizures**	**p-value**	**Blinks**	**p-value**
Baseline	0.849		0.674	
$\beta = 1 \times 10^{-7}$	0.841	0.510	0.625	0.000
$\beta = 1 \times 10^{-5}$	0.831	0.197	0.611	0.000
$\beta = 1 \times 10^{-3}$	0.835	0.256	0.606	0.000
$\beta = 1 \times 10^{-2}$	0.832	0.222	0.591	0.000
Hypothesis	$H_a : \mu_{vib} \neq \mu_b$		$H_a : \mu_{vib} < \mu_b$	

Notes: Table reports test-set accuracy on the primary task of seizure detection and the "nuisance" task of blink detection. A value of 1.0 would indicate perfect accuracy. Each reported value is the mean of 25 runs of the experiment. Reported p-values are for a test of the null hypothesis that the observed difference is due to chance against the indicated alternative.

D. Analysis of Latent Features

The preceding discussion provides an empirical verification of our claim that the VIB formulation learns a representation of the signal containing only the information needed to discriminate between seizure and non-seizure. In this section, we address the problem of *interpreting* that information.

The power in several EEG frequency bands has long been known to be significant for characterizing seizures [21]. These bands are known as: delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (15-30 Hz), and gamma (30-80 Hz) [34]. We now demonstrate that these features of the signal are correlated with the representation obtained by the neural network.

We first extract the latent representation for all (training) signal windows in our database and stack them into a matrix $\mathbf{Z} \in \mathbb{R}^{W \times d}$ where W denotes the total number of windows and d denotes the dimension of the latent space. For each window and each EEG channel, we then extract the relative power in the five frequency bands stated above and stack them into a matrix $\mathbf{F} \in \mathbb{R}^{W \times 5k}$. Our goal is to estimate the correlation between $\mathbf{F}$ and $\mathbf{Z}$.

Table II presents results from a "canonical correlation analysis" (CCA) of $\mathbf{F}$ and $\mathbf{Z}$. CCA is a statistical method which computes linear combinations of features from two sets of variables which are maximially correlated and is well suited to estimating the *linear* relationship between two sets of variables. As in Table I, each value represents the mean across all 25 runs of the experiment. Each number represents the strength of the correlation between a linear combination of features in $\mathbf{F}$ and a linear combination of features in $\mathbf{Z}$. A value of ± 1.0 would indicate a perfect linear relationship while a value of 0.0 would indicate no linear relationship.

We find that the strongest canonical correlates have correlation coefficients between 0.50 and 0.61 which implies a fairly strong linear relationship between the latent variables and frequency domain features of the raw signal. This can be interpreted as indicating that the latent representation contains a significant amount of information about the power features in the raw signal. We note that CCA can only capture *linear* relationships between two variables and the true mutual information between $\mathbf{F}$ and $\mathbf{Z}$ may be higher.

We regard this as tentative evidence that our model is extracting signal features known to be medically relevant. While this accords with medical research and prior work [15], [20], a noteworthy limitation of this approach is that it indicates only an *empirical* relationship and cannot be used to conclude the effect is causal. We leave this matter for future work.

TABLE II
CANONICAL CORRELATIONS

	Value of β			
	1×10^{-7}	1×10^{-5}	1×10^{-3}	1×10^{-2}
Correlate 1	0.604	0.611	0.586	0.505
Correlate 2	0.349	0.298	0.283	0.175
Correlate 3	0.084	0.066	0.060	0.060
Correlate 4	0.033	0.034	0.030	0.034

Notes: Each value indicates the correlation coefficient between the canonical correlates of $\mathbf{F}$ and $\mathbf{Z}$. A value of ± 1.0 would indicate a perfect (linear) relationship while a value of 0 would indicate no linear relationship. Reported values are the average of 25 runs.

IV. CONCLUSION

In this work we have adopted an information theoretic view of seizure prediction and demonstrated how to construct signal representations containing only the information *relevant* for discriminating between seizure and non-seizure. We have demonstrated empirically that these signal representations are able to automatically "ignore" information about ocular artifacts in EEG signals. We have further demonstrated that the signal representations obtained by our method encode features of the raw signal known to be medically relevant, thus validating their interpretability.

References

[1] F. Tang, A. Hartz, and B. Bauer, “Drug-resistant epilepsy: multiple hypotheses, few answers,” *Frontiers in Neurology*, vol. 8, p. 301, 2017.

[2] M. Brodie, S. Barry, G. Bamagous, J. Norrie, and P. Kwan, “Patterns of treatment response in newly diagnosed epilepsy,” *Neurology*, vol. 78, no. 20, pp. 1548–1554, 2012.

[3] U. R. Acharya, S. V. Sree, G. Swapna, R. J. Martis, and J. S. Suri, “Automated EEG analysis of epilepsy: a review,” *Knowledge-Based Systems*, vol. 45, pp. 147–165, 2013.

[4] S. Nasehi and H. Pourghassem, “Seizure detection algorithms based on analysis of EEG and ECG signals: a survey,” *Neurophysiology*, vol. 44, no. 2, pp. 174–186, 2012.

[5] D. Sopic, A. Aminifar, and D. Atienza, “e-glass: A wearable system for real-time detection of epileptic seizures,” *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2018.

[6] T. N. Alotaiby, S. A. Alshebeili, T. Alshawi, I. Ahmad, and F. E. A. El-Samie, “EEG seizure detection and prediction algorithms: a survey,” *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 183, 2014.

[7] B. Litt and J. Echauz, “Prediction of epileptic seizures,” *The Lancet Neurology*, vol. 1, no. 1, pp. 22–30, 2002.

[8] P. Celka and P. Colditz, “A computer-aided detection of EEG seizures in infants: a singular-spectrum approach and performance comparison,” *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 5, pp. 455–462, 2002.

[9] N. Kannathal, M. L. Choo, U. R. Acharya, and P. Sadasivan, “Entropies for detection of epilepsy in EEG,” *Computer Methods and Programs in Biomedicine*, vol. 80, no. 3, pp. 187–194, 2005.

[10] A. Shoeb, A. Kharbouch, J. Soegaard, S. Schachter, and J. Guttag, “A machine-learning algorithm for detecting seizure termination in scalp EEG,” *Epilepsy & Behavior*, vol. 22, pp. S36–S43, 2011.

[11] S. B. Wilson, M. L. Scheuer, R. G. Emerson, and A. J. Gabor, “Seizure detection: evaluation of the reveal algorithm,” *Clinical Neurophysiology*, vol. 115, no. 10, pp. 2280–2291, 2004.

[12] F. Gianfelici, C. Turchetti, and P. Crippa, “A non-probabilistic recognizer of stochastic signals based on KLT,” *Signal Processing*, vol. 89, no. 4, pp. 422–437, 2009.

[13] A. Emami, N. Kunii, T. Matsuo, T. Shinozaki, K. Kawai, and H. Takahashi, “Seizure detection by convolutional neural network-based analysis of scalp electroencephalography plot images,” *NeuroImage: Clinical*, vol. 22, p. 101684, 2019.

[14] M. Zhou, C. Tian, R. Cao, B. Wang, Y. Niu, T. Hu, H. Guo, and J. Xiang, “Epileptic seizure detection based on EEG signals and CNN,” *Frontiers in neuroinformatics*, vol. 12, p. 95, 2018.

[15] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, “Deep learning with convolutional neural networks for EEG decoding and visualization,” *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[16] H. N. Mhaskar and T. Poggio, “Deep vs. shallow networks: An approximation theory perspective,” *Analysis and Applications*, vol. 14, no. 06, pp. 829–848, 2016.

[17] T. Poggio and F. Anselmi, *Visual cortex and deep networks: learning invariant representations*. MIT Press, 2016.

[18] J. A. Urigüen and B. Garcia-Zapirain, “EEG artifact removal – state-of-the-art and guidelines,” *Journal of Neural Engineering*, vol. 12, no. 3, p. 031001, 2015.

[19] X. Jiang, G.-B. Bian, and Z. Tian, “Removal of artifacts from EEG signals: a review,” *Sensors*, vol. 19, no. 5, p. 987, 2019.

[20] K. G. Hartmann, R. T. Schirrmeister, and T. Ball, “Hierarchical internal representation of spectral features in deep convolutional networks trained for EEG decoding,” *2018 6th International Conference on Brain-Computer Interface (BCI)*, pp. 1–6, 2018.

[21] J. W. Britton, L. C. Frey, J. Hopp, P. Korb, M. Koubeissi, W. Lievens, E. Pestana-Knight, and E. L. St, *Electroencephalography (EEG): An introductory text and atlas of normal and abnormal findings in adults, children, and infants*. American Epilepsy Society, Chicago, 2016.

[22] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” *arXiv preprint physics/0004057*, 2000.

[23] N. Slonim and N. Tishby, “Agglomerative information bottleneck,” *Advances in Neural Information Processing Systems*, pp. 617–623, 2000.

[24] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, “Deep variational information bottleneck,” *International Conference on Learning Representations (ICLR)*, 2017.

[25] A. Achille and S. Soatto, “Information dropout: Learning optimal representations through noisy computation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2897–2905, 2018.

[26] M. Chalk, O. Marre, and G. Tkacik, “Relevant sparse codes with variational information bottleneck,” *Advances in Neural Information Processing Systems*, pp. 1957–1965, 2016.

[27] A. Achille and S. Soatto, “Emergence of invariance and disentanglement in deep representations,” *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1947–1980, 2018.

[28] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, “Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[29] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[30] R. A. Amjad and B. C. Geiger, “Learning representations for neural network-based classification using the information bottleneck principle,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[31] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “Tensorflow: A system for large-scale machine learning,” *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

[32] K. Kleifges, N. Bigdely-Shamlo, S. E. Kerick, and K. A. Robbins, “Blinker: automated extraction of ocular indices from EEG enabling large-scale analysis,” *Frontiers in Neuroscience*, vol. 11, p. 12, 2017.

[33] J. M. Wooldridge, *Introductory econometrics: A modern approach*. Nelson Education, 2016.

[34] C. Amo, L. de Santiago, R. Barea, A. López-Dorado, and L. Boquete, “Analysis of gamma-band activity from human EEG using empirical mode decomposition,” *Sensors*, vol. 17, no. 5, p. 989, 2017.