**COLLEGE OF ENGINEERING**

**Preliminary Exam Report:**

**Unsupervised Feature Learning in Electroencephalography (EEG)**

Nabila Shawki

Email: nabila.shawki@temple.edu

Department of Electrical and Computer Engineering

College of Engineering, Temple University

1947 North 12th Street, Philadelphia, Pennsylvania 19122

February 25, 2020

A close up of a logo

Description automatically generated

Examining Committee:

Dr. Joseph Picone, Committee Chair, Department of Electrical and Computer Engineering, College of Engineering, Temple University

Dr. Eduard Dragut, Committee Member, Department of Computer and Information Sciences, College of Science and Technology, Temple University

Dr. Albert Kim, Committee Member, Department of Electrical and Computer Engineering, College of Engineering, Temple University

Dr. Iyad Obeid, Committee Member, Department of Electrical and Computer Engineering, College of Engineering, Temple University

**Executive Summary**

The main goal of this document is to review unsupervised feature learning and assess its relevance to the seizure detection problem in electroencephalography (EEG). We identify challenges involved with extracting features from sequential time-series data and explore how deep learning techniques can be applied to this problem. Unsupervised feature learning has been used extensively in speech recognition systems and been shown to achieve a 41% absolute word error rate (WER) reduction compared to filter bank energies in mismatched domains of clean and noisy speech. We discuss the efficacy of adapting these techniques to the seizure detection problem by replacing our standard linear frequency cepstral coefficient (LFCC) with an unsupervised feature extraction module. We will study autoencoders (AE) and their variations such as convolutional autoencoders and variational autoencoders (VAE). Since our ultimate goal is to detect and classify critical events in the EEG signal, we also review popular classification techniques for EEG applications.

The first paper, titled *Deep Convolution Neural Network and Autoencoders-Based Unsupervised Feature Learning of EEG Signals*, provides an overview of a basic autoencoder. The authors’ proposed model, based on a convolutional neural network (CNN), and referred to by the acronym AE-CDNN, was introduced and evaluated on two publicly available seizure databases. They extracted features of varying dimensions (4 – 128 features) from EEG segments that were 4096 samples in duration and applied several common classification techniques to those segments containing seizures. Features extracted using AE‑CDNN resulted in an error rate of 8% when the dimension of the feature vector was 32. AE-CDNN achieved a reduction in error rate that averaged approximately 30% relative over features of the same dimension derived using principal component analysis (PCA) and sparse random projection (SRP).

The second paper, titled *Extracting Domain Invariant Features by Unsupervised Learning for Robust Automatic Speech Recognition*, introduces a model, Factorized Hierarchical Variational Autoencoder (FHVAE), that can separate sequence-independent or segment-level features from sequence-dependent features. The authors showed that sequence-independent features are domain invariant and contain rich linguistic information. The authors trained an automatic speech recognition (ASR) system with sequence-independent features extracted from clean speech data from the Aurora-4 database. To prove domain invariance, they extracted FHVAE features from noisy speech data from Aurora-4 and CHiME-4 and evaluated an ASR system trained on clean feature data. This system was able to achieve an average WER of 24.30% on Aurora-4 and 60.39% on CHiME-4. FHVAE features reduced the WER by 41% relative compared to filter bank energies and 16% compared to features extracted using a standard VAE.

The final paper, titled *Deep learning for electroencephalogram (EEG) classification tasks: A review*, is a review paper in which the authors discussed popular deep learning architectures for a variety of EEG applications. They also explored preprocessing methods and other feature selection techniques. The authors reviewed over 90 papers and concluded that for, EEG classification tasks, convolutional neural networks (CNN), recurrent neural networks (RNN), and deep belief networks (DBN) yield better performance than stacked autoencoders and multi-layer perceptrons (MLP). For seizure detection, the authors recommended CNNs and RNNs as one study was able to reach an accuracy of 99% with a CNN and another achieved an accuracy of 100% with an RNN. Both studies used a seizure dataset from the University of Bonn.

Unsupervised feature extraction is a promising approach for seizure detection because (1) we are limited in the amount of annotated data we can generate, and (2) it allows us to train across a much larger portion of the TUH EEG Corpus. We believe the AE-CDNN model can improve the performance of our best system on long seizure events (e.g., seizures lasting more than 60 seconds). FHVAE can extract domain and patient-invariant features, which will be useful for noise and artifact reduction. By optimizing the deep learning architecture and the feature extraction process, we should be able to achieve clinically acceptable performance on seizure detection and positively impact critical care in neurology.

**Table of Contents**

1. Introduction 1

2. Unsupervised Feature Extraction using Deep Learning Techniques  3

3. A Fusion Model Integrating Autoencoders and CNNs (AE-CDNN)  4

4. Factorized Hierarchical Variational Autoencoder (FHVAE)  8

5. Applications of Deep Learning to EEG Event Classification  12

5.1. EEG Classification Tasks with Deep Learning  13

5.2. Preprocessing Methods  13

5.3. Input Formulations  14

5.4. Deep Learning Architectures  14

6. Discussion  15

7. Conclusions  16

8. References  16

# Introduction

Electroencephalography (EEG) is a powerful clinical monitoring tool that the experts use for detecting numerous brain-related activities. It is extensively used in various branches of research, including neuroscience, neural engineering, and biomedical engineering (Craik et al., 2019). EEG is attractive because it offers several advantages for studying brain dynamics and assessing neurophysiological functions. EEG provides a high temporal resolution that enables an expert to study cognitive processing in millisecond precision. It is also noninvasive, portable, and feasible (Light et al., 2010; Haufe et al., 2013).

EEG is being used not only for detecting many diseases, including epilepsy, Alzheimer’s, and schizophrenia, but also for commercial applications involving brain-computer interfaces (BCIs) (Light et al., 2010; Craik et al., 2019). Epilepsy is a significant application area since epilepsy is one of the most common brain diseases. It is a chronic neurological disorder that affects 2.4 million people around the world each year (Lin et al., 2016). It induces seizures, which can cause short-term cognitive dysfunctions such as losing consciousness and long-term problems such as inflicting irreversible damage to the brain (Wen & Zhang, 2018). Many patients can experience falls, submersion injuries, fractures, burns, and motor vehicle accidents, which can create additional injuries (Williamson et al., 2012). Traditionally, trained neurologists manually interpret an EEG signal to detect evidence of seizure activity. This process is extremely time consuming and requires a significant amount of expertise. Clinically certified experts are in short supply, especially in many developing countries (Thodoroff et al., 2016). These are the reasons which have motivated the experts to find automatic seizure detection algorithms that would make epilepsy treatment more efficient, affordable and reliable.

Over the last few years, many EEG datasets have become publicly available to support machine learning research on automatic interpretation of EEGs. Moreover, the world is now enjoying the advantages of machine learning techniques (ML) in numerous fields. Among these techniques, deep learning (DL) is a branch that has revolutionized the fields of image classification, speech processing, and computer vision. A DL network is essentially a multi-layer neural network that offers automatic and iterative optimization of its parameters, which lessens the burden of gathering a vast amount of prior knowledge about a problem. Previously, simple neural networks (NNs) used to suffer from long runtimes and poor convergence due to the vanishing gradient problem (Bengio et al., 1994).

Fortunately, recent advances in special purpose processors known as Graphical Processing Units (GPUs) (Lecun et al., 2015) have increased the popularity of deep learning algorithms. GPUs have allowed researchers to train extremely sophisticated networks. GPUs have significantly reduced computation time and alleviated many of the hardware bottleneck issues. A few examples of popular DL algorithms include Convolutional Neural Networks (CNN), Deep Belief Networks (DBN), Recurrent Neural Networks (RNN), Stacked Autoencoders (SAE), Multi-layer Perceptron (MLP), Long Short-Term Memory Networks (LSTM) and many hybrid architectures consisting of combinations of these techniques (Craik et al., 2019).

One can classify machine learning or deep learning algorithms into two broad categories depending on how they learn: supervised or unsupervised learning. The first category, supervised learning, is the most popular approach where one needs to supply labeled data for training a classifier (Portugal et al., 2018). For example, in seizure detection, a DL algorithm needs the data to be annotated with specific labels such as “ictal” or “tonic seizure” so that it can learn to differentiate between an ictal event and a regular brain activity. We refer to the process of creating these labels manually by visual inspection of the data as annotation or transcription. For EEG classification tasks, researchers often prefer calculated features to raw signal data (Craik et al., 2019). Extracting such features is itself a challenge (Lin et al., 2016).

The second category, unsupervised learning, refers to a process where the algorithm learns the underlying features of the provided data without annotations. This can result in improved performance because the algorithm can be trained from a much larger pool of data. Manual annotation of data is expensive and time­­­‑consuming (Yuan et al., 2017). Unsupervised learning has attracted the attention of many DL researchers because it can significantly lower the cost of creating new applications. The machine learning process can be completely automated.

Autoencoders are a type of deep learning algorithm that can extract the latent low-level features from a high-level representation. In recent years, autoencoders and their variations have attracted significant attention in various research areas such as automatic speech recognition (Hsu & Glass, 2018) and cancer detection (Abraham & Nair, 2018).

A typical EEG classification pipeline for any task has three steps: artifact removal, feature extraction, and classification. For example, using traditional machine learning techniques, one can remove different artifacts using Independent Component Analysis (ICA). Then, from the clean data, one can hand-engineer different features or use Principle Component Analysis (PCA), a traditional unsupervised technique, to reduce data dimensionality. Finally, the data can be classified using popular supervised learning algorithms such as a Support Vector Machine (SVM) or Linear Discriminant Analysis (LDA) (Craik et al., 2019).

In this document, we discuss unsupervised feature extraction techniques for sequential signal processing. We are interested in deep learning classification algorithms as well. We have selected three papers:

* Wen, T., & Zhang, Z. (2018). Deep Convolution Neural Network and Autoencoders-Based Unsupervised Feature Learning of EEG Signals. *IEEE Access*, 6, 25399–25410. *https://doi.org/10.1109/ACCESS.2018.2833746*
* Hsu, W. N., & Glass, J. (2018). Extracting Domain Invariant Features by Unsupervised Learning for Robust Automatic Speech Recognition. *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5614–5618). Calgary, AB, Canada: IEEE. *https://doi.org/10.1109/ICASSP.2018.8462037*.
* Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering*, 16(3), 31001. *https://doi.org/10.1088/1741-2552/ab0ab5.*

In Wen and Zhang (2018), a CNN and an autoencoder-based model called AE-CDNN is introduced. This model uses CNN’s main components, convolutional layers, and maxpooling layers, to encode features from an EEG segment. AE-CDNN was able to achieve improved accuracy in seizure detection. The authors stated that the features of AE-CDNN “could speed up the convergence and reduce the training time of classifiers.” These features also achieved better accuracy on two seizure databases compared to traditional unsupervised techniques.

Hsu and Glass (2018) applied a Factorized Hierarchical Variational Autoencoder or FHVAE for extracting sequence-dependent and sequence-independent features for a domain-invariant automatic speech recognition (ASR) system. The ASR system was trained on features extracted from clean speech sequences and evaluated on both clean and noisy features. The features from the FHVAE system outperformed traditional filter bank (FBank) features and a variational autoencoder (VAE) features. The field of speech recognition was chosen for investigation because annotation has been a tremendous bottleneck in the development of the extremely large datasets needed to train complex systems (compounded by the need for these systems in many languages) (Shah et al., 2018).

In Craik et al. (2019), the authors reviewed various deep learning trends in several research areas where automated EEG classification is used extensively. They investigated 90 studies that used deep learning. They compared and contrasted major components of the systems including artifact removal, signal representations used as inputs to the deep learning networks and the deep learning architectures. The authors also recommended particular deep learning architectures for each task based on their findings.

Finally, we will discuss how these studies inform our future research directions. We will analyze the claimed superiority of unsupervised feature extraction over two approaches: hand-engineered features and raw signal input. We will discuss the impact these findings will have on our future research.

# Unsupervised Feature Extraction using Deep Learning Techniques

Unsupervised learning has been identified as one of the significant trends in deep learning (Lecun, 2019). Unsupervised learning methods are used to extract useful features from unlabeled data, remove redundant information from the input, and preserve the unique aspects of the input data. In Figure 1, a block diagram for a typical unsupervised learning and classification is shown. The light green dashed box shows the training phase where the unlabeled data is fed to the network. The blue dashed box shows the testing phase where the features are extracted. Those features are used in the classifier to verify the validity of the network (Wen & Zhang, 2018). One can also fine-tune an unsupervised learning system to act as a classifier after extracting the features (Lin et al., 2016; Yuan et al., 2017). Since an EEG signal is a multichannel time series, this document will briefly explore why using unsupervised learning on such data is beneficial. A detailed discussion of this can be found in Längkvist et al. (2014).

Most real-world data, such as language, vision, and motion, has a temporal context. Physiological signals such as speech, vision and EEG are examples of such signals. Many traditional approaches such as autoregressive models (Hyndman & Athanasopoulos, 2018) and Linear Dynamical Systems (LDS) (Cheng & Sabes, 2006) have been used to estimate parameters which are then incorporated into machine learning-based classifiers. However, sometimes these approaches cannot extract high performance features from complex, noisy time series data. In some cases when the dynamics of the data are either complicated or unknown, these methods are unable to model complex nonlinear behavior. Developing domain-specific techniques for modeling such complex time series data is expensive, time-consuming and requires significant amounts of subject matter expertise.

Time series data is much different from static data as its data points are taken from a continuous, real-valued process at certain intervals. This process is known as sampling, and the resultant data can contain noise and be highly dimensional. The data itself does not contain enough information to support discovery and modeling of the underlying process. This situation is often observed in electronic sensors (Peace et al., 2006) and financial data (Fama, 1965).

Another critical characteristic of time series data, particularly from stochastic or chaotic systems, is that an identical input can provide different outputs at different times. Further, many time series, particularly those corresponding to physical systems, are nonstationary, which mean that simple statistical measures such as the mean and variance change with time. Unsupervised feature learning with deep networks has been successful in applications involving static data representations. Unsupervised learning needs to be adapted to problems involving nonstationary signals (Längkvist et al., 2014).

A close up of a map

Description automatically generated

Figure 1. An overview of unsupervised learning  
(Wen & Zhang, 2018)

Restricted Boltzmann Machines (RBM), autoencoders, and their variations are common unsupervised deep learning techniques that can be coupled with powerful machine learning approaches such as deep LSTM and CNN networks to achieve high performance classification. This document focuses on autoencoders and their variations.

Most unsupervised deep learning methods are comprised of two components: an encoder and a decoder (Masci et al., 2011). In Figure 2, a typical autoencoder is shown. It has an input layer, a hidden layer, and an output layer. The machine learning algorithm used to train the system adjusts the weights of the hidden layer such that it can replicate the input data at the output layer. The hidden layer learns the structure of the signal and delivers the desired feature representation.

Let the input data be Then, the features, , can be obtained from the hidden layer using the encoding function:

A close up of text on a white background

Description automatically generated

Figure 2. A simple autoencoder  
(Wen & Zhang, 2018).

. (1)

Here, is the weight matrix that connects the input and the hidden layers, is the bias vector, and is the activation function. The decoding process can be formulated as follows:

(2)

The weight matrix between the hidden layer and the output layers is and the bias vector is The goal is to recreate the training data at the decoder. The objective function to be minimized is defined as:

 . (3)

In the next two sections, we will study two unsupervised deep learning models built upon basic ideas of autoencoders. They used CNN and LSTM layers to capture the low-level features from time-series data.

# A Fusion Model Integrating Autoencoders and CNNs (AE-CDNN)

It is not trivial to train an autoencoder network in practice. As discussed earlier, it is difficult to extract important information from an EEG signal or any sequential data considering their high dimensionality. CNNs are one popular approach for dealing with sequential signals with a high dimensionality. They preserve the local features by receptive field and parameter sharing (Wen & Zhang, 2018).

The local features are collected by local connectivity which means that the neurons are only connected to a subset of the input matrix. This is different from a fully connected neural network where all neurons are connected to the given matrix. The hyperparameter called receptive field ensures the preservation of local features. We normally set the height and width of the local region via filter size when we parameterize a convolutional layer. These neurons are also connected spatially to all adjacent neurons (Fei-Fei, Johnson, & Yeung, 2016).

With parameter sharing, the convolutional layers control the number of parameters in a single layer. The concept is that similar weights of one slice of a layer can be reused across all the slices of that layer. Let, a slice of a convolutional layer be and a layer has a total of these 2-D slices. Using the parameter sharing scheme, all slice will have the same weights and bias. This significantly reduces the number of parameters for each layer in CNN. Using the filter size and the parameter sharing, only the unique parameters are then applied for each slice (Fei-Fei, Johnson, & Yeung, 2016).

Figure 3 shows the encoding and decoding process for AE-CDNN. The encoder iteratively applies several kernel functions and down-sampling to collect a variety of local features. The decoder reconstructs the sample using deconvolution and upsampling. As expected, the authors used two components of a CNN: convolution layers and pooling layers. They have taken advantage of the pooling layer which can downsample and extract useful features from the sample input. In the decoder, AE-CDNN uses deconvolution and upsampling layers iteratively to restore the input data.

At the encoder layer, the feature map at the kth position, , is collected using the equation described below:

 . (4)

Here, and are the weight and bias vector of the convolution kernel. A maxpooling layer is applied which uses a sliding window and collects the maximum values within that window. All of the feature maps are collected in a feature vector of dimension . Next, a reshape operation maps feature vectors to a one-dimensional array, of length . Then a fully connected layer synthesizes the information from all pooled features. The final feature vector is:

 . (5)

Here, is the weight and is the bias vector of the fully connected layer.

At the decoder end, the feature vector, , is reconstructed into . Then, it is reshaped to make pooling layers. Then each of the layers is iteratively deconvoluted and upsampled to restore the original input signal, , into an output signal, . The authors used two loss functions with an Adam optimizer (Ba et al., 2014). The first loss function is based on equation (3):

 . (6)

The model that uses this function is called AE-CDNN-L1.

The second loss function is:

 , (7)

and was introduced to reduce the impact of on the difference of each sample when the input is large. The model developed using this loss function is called AE-CDNN-L2.

A screenshot of a computer

Description automatically generated

Figure 3. An AE-CDNN model

The authors applied this technique to two datasets. The first is a public dataset that was published by Andrzejak et al. (2001). In this dataset, there are five subsets denoted A-E. Each subset has 100 EEG recordings of duration 23.6 secs (4,096 samples @ 173.5 samples/sec). They were recorded with 16-bit resolution. The A and B subsets do not contain any seizures as they were recorded from five healthy volunteers. The C and D subsets do not have seizure data either, but they were collected from five epileptic patients. Lastly, the E subset contains seizure data.

The second dataset was a subset of the CHB-MIT Corpus (Shoeb, 2009). This EEG dataset contains EEG recordings from 23 children with refractory epilepsy. Each recording lasts for one hour in duration and has 23 channels (921,600 samples @ 256 samples/sec). They were recorded with 16-bit resolution. The authors selected one channel from 23 channels for all these recordings based on variance. The reason behind selecting channels based on variance is when epilepsy occurs, the signal fluctuates significantly. This fluctuation is represented by a larger variance across the signal. Based on this rule, they extracted 200 EEG samples with seizures and 200 without seizures. Each of these signals has a duration of 4,096 samples, which matches the sample length from the first dataset.

The study used a sigmoid function as their activation function, and the output range was within 0 to 1. For ensuring that their proposed loss functions would provide valid results, they normalized their data using minimum and maximum values over all dimensions. The deep network based on the AE-CDNN model is shown in Figure 4. While the network depth, encoding, and decoding process remain the same, the authors varied the number of extracted features (*m*) to analyze the learning ability of their model. After extracting the features, they used several common classifiers to detect seizures in their datasets. The classifiers and their parameters are shown in Table 1. The ‘−’ indicates that they used the default parameters.

A screenshot of a cell phone

Description automatically generated

Figure 4. The AE-CDNN model (Wen & Zhang, 2018)

Figure 5 compares the accuracies of AE-CDNN-L1 and AE-CDNN-L2 as a function of the feature dimensions for the two datasets. Both models achieve superior accuracies when . The models performed more consistently as *m* increased for dataset 1. For dataset 2, the accuracies vary across the feature dimension. The highest accuracy (92.9%) for dataset 1 was achieved with AE-CDNN-L1. The authors reported that after 10-fold cross-validation they were able to achieve accuracies higher than 90% for both datasets for .

The author compared the performance of their unsupervised learning model to that of existing dimension reduction methods. They considered Principle Component Analysis (PCA), which can extract low dimensional features by a linear transformation of high dimensional data (Ahmad et al., 2008). The second method is Random Projection (Bingham & Manilla, 2001), that constructs a Lipschitz mapping to reduce dimensions with high probabilities. Sparse Random Projection (SRP) (Li et al., 2006) uses this idea and reduces dimensionality by projecting the original input space using a sparse random matrix. Figure 6 compares the average accuracies of the proposed methods to traditional unsupervised learning techniques. Compared to AE-CDNN models, PCA and SRP’s accuracies are more stable with respect to the feature dimension as their accuracies did not vary as the number of features increased. On the other hand, when *m* was greater than 4, both AE-CDNN models performed better than the existing traditional dimension reduction methods.

|  |  |
| --- | --- |
| **Method** | **Parameter details** |
| *k*-NN (k-Nearest Neighbors) | *k* =3 |
| SVM1 (Support Vector Machine, linear kernel) | kernel=“linear” |
| SVM2 (radial basis function kernel) | kernel=“rbf” |
| DT (Decision Tree) | max\_depth=5 |
| RF (Random Forest) | max\_depth=5, n\_estimators=10, max\_features=1 |
| MLP (Multilayer perceptron) | − |
| ADB (AdaBoost Algortihm) | − |
| GaussianNB (Gauss Bayesian Classification) | − |

Table 1: A summary of classifier methods

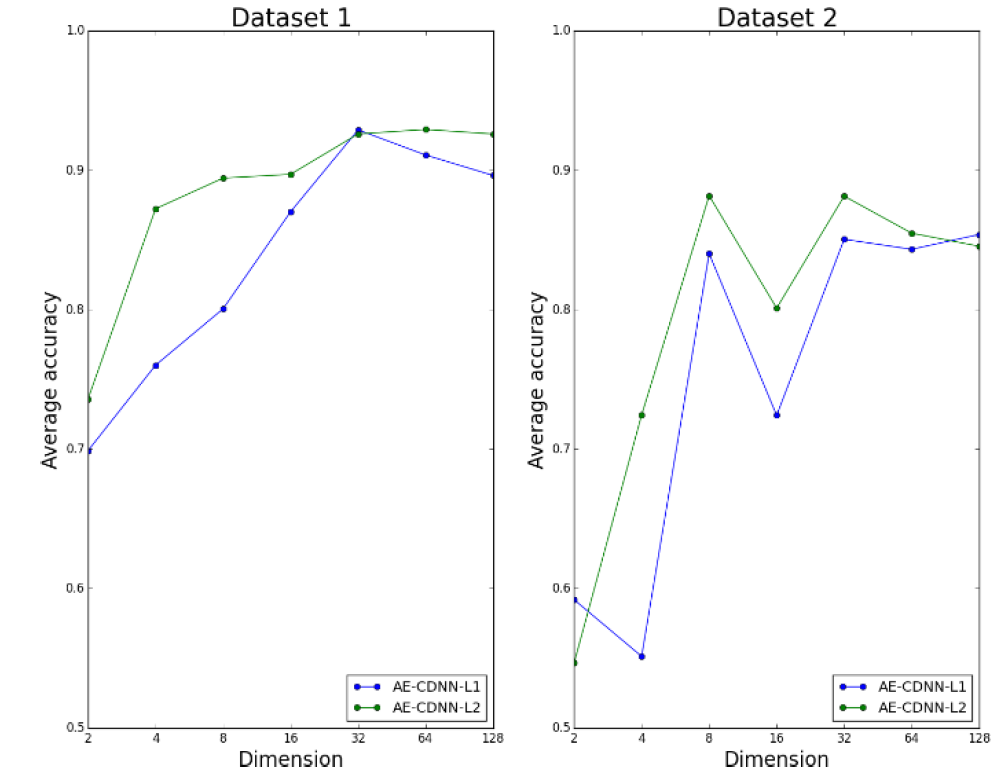


Figure 5. Comparison of AE-CDNN-L1 to AE-CDNN-L2 (Wen & Zhang, 2018)

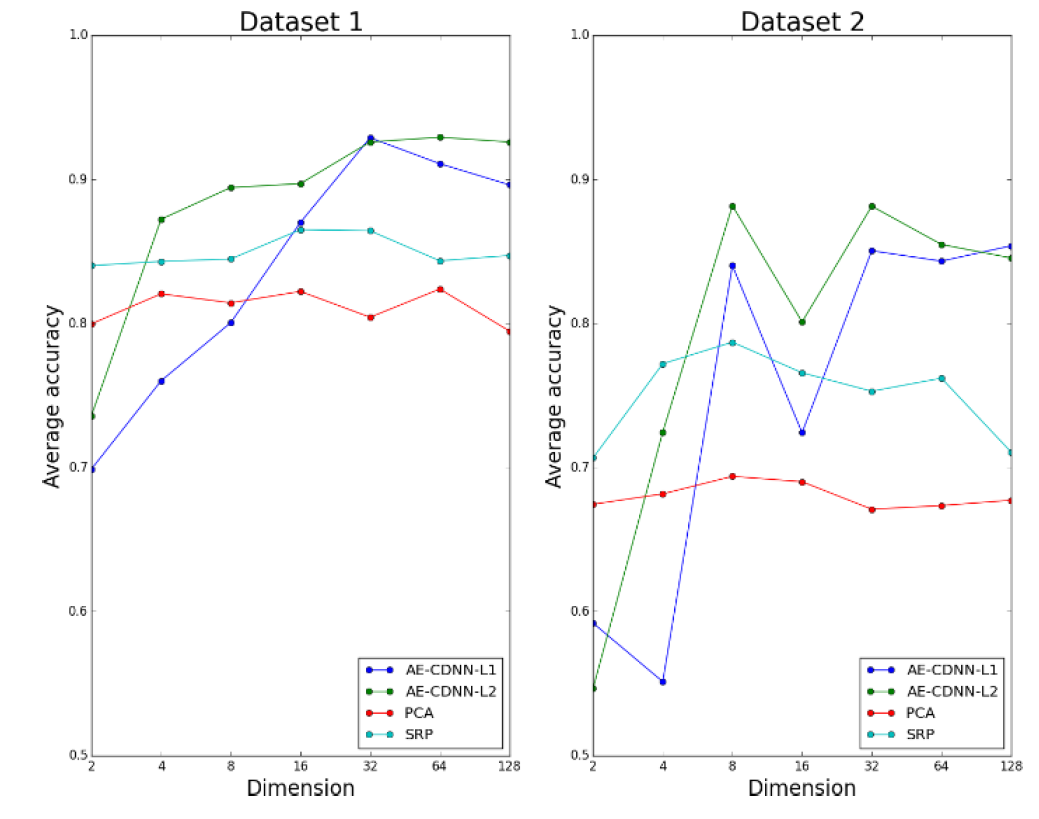


Figure 6. Comparison of AE-CDNN Performance to PCA and SRP (Wen & Zhang, 2018)

Finally, the authors compared performance to other published studies. They compared their results to other studies by 10-fold cross-validation and achieved comparable results. Moreover, they showed that their conclusions held for two different datasets when most studies focused on only one dataset.

Though the AE-CDNN model was shown to be useful for extracting features from a segment of EEG data, it was not evaluated on sequential data. The authors only did classification of one segment selected from one channel. For a commercially feasible device, one must be able to classify a sequence of EEG data in real time. Also, it must be noted because of this limitation, this study did not use sensitivity and specificity as their performance metrics. These are two very important metrics that measure false alarms, which is a very important issue in automated seizure detection (Gadhoumi et al., 2016). Nevertheless, this model provides a promising baseline that can be used to support further investigations.

# Factorized Hierarchical Variational Autoencoder (FHVAE)

Recently the application of factorized hierarchical learning has gained the attention of researchers developing variational autoencoders. When an autoencoder can extract the lower-dimensional features from a given input, a variational autoencoder can learn the prior distribution of latent space (Tschannen et al., 2018). A factorized hierarchical variational autoencoder (FHVAE) is a type of deep probabilistic generative model that learns both an inference model and a generative model. These models can infer the latent variables from the input. However, it is not easy to figure out the associated factors in those latent variables. As discussed in Section 2, understanding the underlying features of sequential data is difficult and relatively underexplored. Learning the behavior of a dataset based on individual factors is important to achieve domain-invariant recognition. FHVAE attempts to learn disentangled and interpretable representations from sequential data in an unsupervised manner (Hsu et al., 2017).

The FHVAE model extracts the sequence-dependent and sequence-independent priors in a factorized hierarchical graphical model. The authors used this model to extract domain-invariant features for an automatic speech recognition (ASR) system (Hsu & Glass, 2018).

In a speech generation process, the fundamental frequency (F0), volume, etc. are the sequence-level attributes that impact an utterance. These are similar within a sequence but different between sequences. On the other hand, spectral contours can be defined as a segment-level attribute which can be similar within and between utterances. FHVAE takes advantage of this idea and attempts to extract segment-level and sequence-level attributes separately and utilize them for automatic speech recognition.

The FHVAE model imposes sequence-dependent priors and sequence-independent priors on different sets of latent variables. Let , , denote a dataset that has i.i.d. sequences. Here, is a sequence that contains segments.

A sequence that has segments is assumed to be generated using a random process that involves the latent variables , , and as follows:

1. An *s-vector* is drawn from a prior distribution N ;
2. i.i.d. *latent segment variables* and *latent sequence variables*  are drawn from a sequence-independent prior N and a sequence-dependent prior N respectively;
3. i.i.d. speech segments are drawn from a condition distribution N , whose mean and diagonal variance are parametrized by neural networks.

The joint probability for a sequence can be formulated as:

. (8)

Here, is assumed to be a summarization of sequence-level attributes for a sequence, and encodes the sequence level attributes for a segment that are similar within an utterance. Finally, encodes the segment-level features. Together, and provide enough information to generate a segment.

FHVAE then proposes an inference model that attempts to infer the actual posterior :

*.* (9)

In this equation, the inferred values of and depend on the corresponding segment or the posteriors:

N (10)

N . (11)

These posteriors are assumed to have a diagonal Gaussian distribution. As before, the means and variances are parameterized using neural networks. However, is an isotropic Gaussian which is modeled as N is a trainable lookup table of the posterior mean of for each training sequence (Hsu et al. 2017).

The sequence-level attributes are encoded by and when each utterance is treated as a sequence. The segment-level features are encoded with which preserves the residual linguistic information, which is invariant to the segment-level variations and are the features for a domain invariant ASR feature.

Like all generative models, FHVAE must maximize the marginal likelihood of the observed dataset. As the actual posterior is intractable, FHVAE introduces a *discriminative segment variational lower bound* with a weighting parameter that encourages discriminability and does not let FHVAE degenerate into a VAE if is same for all utterances:

(12)

The authors used their proposed method on two datasets, Aurora-4 and CHiMe-4. Aurora-4 includes two microphone types and six artificially added noise types. There are four conditions: (A) clean, (B) channel, (C) noisy, and (D) channel + noisy. Here, A is a matched condition where the other three are mismatched conditions. They divided this dataset into training and evaluation sets. The CHiME-4 dataset contains naturally noisy data that were recorded from a distance. They also trained a VAE model for comparison purposes. The ASR system was trained with set A from Aurora-4 and was evaluated on the CHiME-4 development set.

The Seq2Seq-VAE and Seq2Seq-FHVAE models used LSTM layers. While the VAE used a negative variational lower bound with an L2 regularization with a weight of 10-4, the FHVAE used the negative discriminative segment variational lower bound, α. The VAE contained 64 dimensions in its latent space. The FHVAE contained two 32-dimension latent spaces.

These models do not use the raw speech data directly for unsupervised feature generation. Instead, the data in each frame is presented as 80-dimensional filter bank (FBank) energies. The authors used the Computational Network Toolkit (CNTK) for initial acoustic model training (an HMM-GMM model based on Kaldi). For the classification, a three-layer LSTM network was used with a cross-entropy criterion and trained using a truncated-backpropagation-through-time (BPTT) algorithm.

The authors presented the results of several experiments. They worked with three types of feature representations: (1) FBank, (2) a latent variable, extracted using the VAE model, and (3) a latent segment variable, extracted from the FHVAE model. They conducted six experiments on Aurora-4. The results for these experiments are shown in Table 2. In experiment no. 1, the authors used three feature sets for establishing a baseline. For the matched case A, the FBank feature performed the best. The Word Error Rate (WER) was 3.21%, which was the best for this condition across all features. However, for all other cases, it decreased significantly. The average WER was 65.64%. When the features from the VAE analysis were used, the average WER was 44.79%. The same features were used without the discriminative training weight (). In this case, the average WER was 40.31%. Finally, they used the FHVAE feature generated with and achieved an average WER of 26.58%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Settings** | | | | | | **Avg. WER (%)** |
| Exp. Index | Feature | No. Layers in Encoder/Decoder | No. Units |  | Seq. Label |
| 1 | FBank | - | - | - | - | 65.64 |
|  | 1/1 | 256/256 | - | - | 44.79 |
|  | 1/1 | 512/256 | - | - | 40.31 |
|  | 1/1 | 256/256 | 10 | uttid | 26.58 |
| 2 |  | 1/1 | 256/256 | 10 | uttid | 26.58 |
|  | 2/2 | 256/256 | 10 | uttid | 25.54 |
|  | 3/3 | 256/256 | 10 | uttid | 24.30 |
| 3 |  | 1/1 | 128/128 | 10 | uttid | 34.66 |
|  | 1/1 | 256/256 | 10 | uttid | 26.58 |
|  | 1/1 | 512/512 | 10 | uttid | 26.97 |
| 4 |  | 1/1 | 256/256 | 0 | uttid | 33.30 |
|  | 1/1 | 256/256 | 5 | uttid | 30.55 |
|  | 1/1 | 256/256 | 10 | uttid | 26.58 |
|  | 1/1 | 256/256 | 15 | uttid | 29.92 |
|  | 1/1 | 256/256 | 20 | uttid | 32.64 |
| 5 |  | 1/1 | 256/256 | 10 | uttid | 26.58 |
|  | 1/1 | 256/256 | 10 | noise | 32.27 |
|  | 1/1 | 256/256 | 10 | speaker | 34.95 |
| 6 |  | 1/1 | 256/256 | 10 | uttid | 26.58 |
|  | 1/1 | 256/256 | 10 | uttid | 43.61 |

Table 2. WER on the Aurora-4 Corpus

In experiment no. 2, the authors varied the number of layers. The system performance increased as they increased the number of layers. The best performance across all the experiments was a WER of 24.30%. This was obtained using three layers in both encoder and decoder, which is WER = 24.30%. However, they did not further investigate this combination for any other variations for Aurora-4.

In experiment no. 3, they changed the number of units in the FHVAE layers. In this case, 256/256 was the best combination, which is the same combination as that used in the baseline experiment. Again, the discriminative training weight () plays a huge role in robust, domain invariant feature learning.

In experiment no. 4, the authors varied the value of from 0 to 20. The performance of the ASR increased as the value of increased because the system can factorize the sequence and segment level information more. Then, as increased, the performance decreased as it prevents the system from learning by inhibiting its capacity for understanding the lower level features.

In experiment no. 5, the sequence labels were varied for supervised classification. In this case, as well, the best performance was the same as the baseline experiment. Using the noise label or the speaker ID did not increase the performance. The FHVAE models used in this experiment set were trained with speaker-specific and noise-specific priors by concatenating sequences with the same speaker or noise label. However, the information regarding speaker or noise gets discarded while extracting domain invariant features in the segment-level. This is expected as the speaker’s attributes and the noise are sequence-level information.

In experiment no. 6, the authors compared and a combination of and (the s-vector). Here, introduced sequence-level attributes to the model. As there were differences between the prior distribution in the testing and training set, the performance deteriorated with the inclusion of Both the s-vector and the features of the VAE, , carry information that are different between sequences like noise, speaker identity, and room geometry, which are not domain invariant. As expected, the performance for decreased (43.61%) and it is comparable to the performance of features (44.79%). In Table 3, the author showed their results for CHiME 4. In experiment no. 1, the features from FHVAE performed worse than FBank and VAE features by a small margin when it was tested on the clean dataset. But it outperformed the others when it was used on the noisy set. In experiment 2, the WER decreased as the number of layers was increased.

FHVAE has been proven to work for domain invariant speech recognition. Seizure detection tasks can also benefit from this model. FHVAE can be used for extracting domain-invariant EEG features for seizure detection.

# Applications of Deep Learning to EEG Event Classification

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Settings** | | | | | | **Avg. WER (%)** | |
| Exp. Index | Feature | No. Layers in Encoder/Decoder | No. Units |  | Seq. Label | Clean | Noisy |
| 1 | FBank | - | - | - | - | 19.37 | 87.69 |
|  |  | 1/1 | 512/256 | - | - | 19.47 | 73.95 |
|  |  | 1/1 | 256/256 | 10 | uttid | 19.57 | 67.94 |
| 2 |  | 1/1 | 256/256 | 10 | uttid | 19.57 | 67.94 |
|  | 2/2 | 256/256 | 10 | uttid | 19.73 | 62.44 |
|  | 3/3 | 256/256 | 10 | uttid | 19.52 | 60.39 |

Table 3. WER for features of CHiME-4 dataset from multiple experiments.

After feature extraction, the next step is classification. In this section, we will discuss recent practices in classifying EEG tasks. Craik et al. (2019) have discussed the usage of deep learning algorithms for classifying brain activities with EEG data. They approached the problem systematically to find which EEG classification tasks have employed DL, what kinds of input formulations have been used for training, and if there is a specific DL network suitable for specific types of tasks. The authors performed a literature review on the publications available on the Web of Science and PubMed. They employed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) method to identify the papers that have used DL on EEG tasks on 22 December 2018. Initially, they found 349 papers from which they selected 90 papers using a few well-chosen criteria:

* *Electroencephalography only* – The search excluded the studies which presented EEG with other physiological signals or videos, such as electrooculography and electromyography.
* *Task classification* – The authors focused only on human brain activity related studies. They excluded power analysis, non-human studies, and feature selection without any end classification.
* *Deep Learning* – The review defined deep learning as a neural network with at least two hidden layers. This criterion might have discarded any studies that might have used any traditional machine learning algorithms along with the deep learning architectures.
* *Time* – The review only included papers from the last five years.

The 90 papers that were selected were then reviewed. The authors organized these papers into the following categories: task information, preprocessing, input formulation, and deep learning strategies. This document will discuss their findings with a primary focus on the task of seizure detection.

## EEG Classification Tasks with Deep Learning

Several EEG classification research areas have explored deep learning. The authors divided these tasks into six major groups:

* *Emotion Recognition Tasks:* A total of 14 studies (16%) applied deep learning methods for emotion recognition tasks. These studies are an integral part of brain-computer interfaces. For these studies, subjects watch videos that are associated with certain emotions. The subjects take an emotion self-assessment test. Finally, the results from these tests and the original emotion class are mapped into a pair of valence and arousal values.
* *Motor Imagery Tasks:* A total of 22 studies (24%) are also related to BCIs to understand a user’s intended movement. A user thinks about a specific movement using their limbs or tongue for collecting the data for the classifier.
* *Mental Workload Tasks:* The authors reported that 14 papers (16%) of the studies dealt with classifying mental workload tasks. For these studies, the data acquisition procedure involves collecting EEG data while a person accomplishes mental tasks. Examples of these tasks include driving simulation studies, live pilot studies, and responsibility tasks. Stress monitoring and BMI (brain-machine interfaces) performance monitoring are two fields where these studies are useful.
* *Seizure Detection Tasks:* Seizure detection was discussed in 12 papers (14%). EEG signals of patients with epilepsy were recorded so that their brain activity can be observed during a seizure event. In order to create a control class, EEG data from non-epileptic patients were also recorded.
* *Sleep Stage Scoring Tasks:* Only 8 papers (9%) addressed sleep stage scoring tasks. EEG data were recorded overnight and annotated by experts. They classified the stages into 1, 2, 3, 4, and rapid eye movement stage. The intentions behind these studies include reduction of the dependence on manual interpretation by trained experts and a better understanding of the sleep stages of a patient.
* *Event Related Potential Tasks:* Only 9 papers (10%) addressed the classification of event related potentials. EEG signals were recorded while the subjects watched a rapid sequence of pictures or letters and attempted to focus on a specific indicator. When that specific letter/image appears, a response can be seen in the EEG data, and it is recorded as a P300 response. These EEG signals are cleaner with minimum artifacts and a high signal-to-noise ratio. These studies are helpful for understanding non-verbal communication.

Finally, 11 papers (13%) of the studies do not fall into these categories. These additional studies address the following topics: Alzheimer’s classification, bullying indices detection, depression, gait pattern classification, gender classification, detection of abnormal EEG, and transcranial stimulus treatment effectiveness. Table 4 summarizes the percentages of studies for each task.

## Preprocessing Methods

A disadvantage of the EEG signal is that it is noisy. The electrodes pick up other electrical physiological signals, such as an electromyogram (EMG) signal associated with eye blinks and muscle movements from the neck. When a patient moves, motion artifacts can also be introduced. Artifact removal from EEG data has been widely studied. Among the 90 papers presented in this review, 41% did not explicitly address artifact removal. The rest of the removal procedures, as shown in Table 4, fall into one of three methods:

* Manual Removal (29%) – 26 studies removed artifacts such as EMG manually. However, it is challenging to spot persistent and sparse noise in a multi-channel setting.
* Automatic Removal (8%) – independent component analysis (ICA) and discrete wavelet transform (DWT) were used to remove artifacts.
* No Removal (22%) – a significant number of papers did not use an explicit artifact removal technique.

After artifact removal, most studies limited the bandwidth of the EEG data using frequency domain filters. If the data within a specific frequency range is not necessary, one can discard it. About 50% of the studies used a low pass filter to keep data at or below 40 Hz. No studies considered if a deep learning algorithm gives comparable results if no preprocessing methods were applied.

For seizure detection, there were three studies that applied a variety of strategies to reduce or remove artifacts. Two of the studies worked with data below 40 Hz while one study worked with data below 70 Hz. We can note that unsupervised techniques like denoising autoencoders can extract useful features automatically by ignoring noise and artifacts (Qiu et al., 2018).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tasks** | **Studies**  **(%)** | **Artifact removal strategies** | | | | **Input formulations** | | |
| **Manual removal**  **(ct.)** | **Automatic removal (ct.)** | **No removal (ct.)** | **Not Addressed (ct.)** | **Signal Values (%)** | **Calculated Features (%)** | **Images (%)** |
| **Emotion Recognition** | 16% | 4 | − | 2 | 3 | 4% | 9% | 5% |
| **Motor Imagery** | 22% | 5 | 2 | 3 | 7 | 9% | 13% | 4% |
| **Mental Workload** | 16% | 3 | 2 | 1 | 4 | 4% | 9% | 5% |
| **Seizure Detection** | 14% | 1 | 1 | 1 | − | 9% | 3% | 5% |
| **Sleep Stage Scoring** | 9% | 3 | − | 1 | 3 | 5% | 4% | 1% |
| **Event Related**  **Potential** | 10% | 1 | 2 | 2 | 3 | 6% | 4% | 1% |
| **Others** | 13% | − | − | − | − | − | − | − |
| **Total** | 100% | 29% | 8% | 22% | 41% | 39% | 41% | 20% |

Table 4. Artifact removal strategies and input formulations statistics in Craik et al. (2019)

## Input Formulations

Apart from the inherent noisy characteristics, EEG data also suffer from channel crosstalk. The electrodes placed on a human scalp pick up delayed and noisy versions of the same signal observed on spatially adjacent sensors. Identifying the correct data is not simple “because of the anisotropic volume conduction characteristics in human brain tissues, skull, scalp, and hair.” The authors found that these studies relied on three different types of strategies to combat channel crosstalk: calculated features (41%), images (20%), and the signal values (39%), as shown in Table 4. Seizure detection studies mostly use raw signal data. Only approximately 25% of the papers that addressed seizure detection used calculated features.

## Deep Learning Architectures

In this section, we discuss the variety of deep learning architectures used in these studies. Table 5 shows that CNNs are the most popular design choice, as 43% of the studies used it. On the other hand, only 8% of the studies used stacked autoencoders (SAE). None of the studies that used SAE exploited convolutional or RNN layers. Instead, they used dense, hidden layers. Seizure detection tasks favored either CNNs or RNNs. None of the seizure detection studies used hybrid architectures or DBNs.

The next parameter that varies across these studies is the activation function. About 70% of studies that used convolutional layers used a rectified linear unit (ReLU) as the layer activation function. A few other notable activation functions are exponential linear unit (8%), leaky rectified linear unit (8%), and hyperbolic tangent or tanh (5%). For the fully connected internal layers, sigmoid functions seemed to be popular while for the classifier fully connected layers, softmax functions were popular.

Only three studies that used SAE discussed their activation functions. Two studies applied sigmoid functions in their AE layers and one used ReLU.

Overall, for seizure detection, the authors recommended CNNs and RNNs. There are two studies that used their networks on a shared seizure detection dataset from the University of Bonn. The study that used RNN reached 100% accuracy while the one that used CNN achieved 99% accuracy. It should be noted that this study only considered the papers that globally fine-tuned their autoencoders to classify the input data at the end. The studies that applied any kind of traditional machine learning algorithms for further classification might have been ignored by this study (Chai et al., 2016; Wen & Zhang, 2018). Also, the study only contained simple, stacked autoencoders. Autoencoders with CNN or RNN layers did not seem to be considered.

# Discussion

In this preliminary exam report, unsupervised feature learning for sequential data and deep learning architecture trends for EEG data analysis have been discussed. Unsupervised feature learning offers many advantages for machine learning system design. We discussed two unsupervised feature learning techniques that were variations of classic autoencoders. AE-CDNN used convolutional layers to extract features from segments. These features were used in conjunction with several popular classifier architectures to detect seizures. AE-CDNN can be modified to extract features from segments of EEG data and can replace the linear frequency cepstral coefficient features currently used in our best seizure detection system.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Deep Learning Algorithms** | **Tasks** | | | | | | | **Input Formulations** | | |
| **Emotion Recognition (%)** | **Motor Imagery (%)** | **Mental Workload (%)** | **Seizure Detection (%)** | **Sleep Stage Scoring (%)** | **Event Related**  **Potential (%)** | **Overall (%)** | **Signal Values (%)** | **Calculated Features (%)** | **Images (%)** |
| **CNN** | 29% | 40% | 29% | 46% | 38% | 56% | 43% | 55% | 15% | 30% |
| **DBN** | 29% | 26% | 22% | − | − | 22% | 18% | 28% | 72% | − |
| **SAE** | 8% | − | 15% | 8% | 12% | 22% | 8% | 12% | 88% |  |
| **MLPNN** | 14% | 14% | − | 8% | 12% | − | 9% | 10% | 90% |  |
| **Hybrid** | 14% | 6% | 28% | − | 38% | − | 12% | − | − |  |
| **RNN** | 6% | 14% | 6% | 38% | − | − | 10% | 36% | 44% | 20% |

Table 5. The Deep Learning Architecture strategy statistics in Craik et al. (2019).

The other model proposed for feature extraction was a factorized hierarchical variational autoencoder (FHVAE). This was a twist on a variational autoencoder to extract domain invariant features. This approach can preserve important linguistic information that is critical to building a robust ASR system. The model can extract segment-level features successfully from speech data under noisy conditions. This property of FHVAE inspired us to investigate this model closely. As discussed in Section 5.2, the EEG signal is inherently noisy, and artifact reduction plays a major role in high performance systems.

Another point to note is that ictal patterns vary widely between patients. FHVAE can analyze those patterns to identify the latent characteristics that are more common among patients. If modified properly, FHVAE can extract useful features of a seizure which can help us to build domain and patient-invariant seizure detection systems.

Finally, we presented an overview of how deep learning techniques are being applied to a variety of EEG-related tasks. We discussed preprocessing techniques, input types, and deep learning architectures. Convolutional neural network (CNN) and recurrent neural network (RNN) are more popular for EEG classification tasks than other types of deep learning architectures. We also learned about deep belief networks that are popular in mental workload, motor imager, event related potential, and emotion recognition tasks. In a shared database for emotion recognition, DEAP, DBN with power spectral density (PSD) features outperformed all other networks by achieving 89% accuracy. We will explore this model for the seizure detection task.

# Conclusions

It is challenging to analyze EEG data and extract useful features for detecting seizures. We are interested in combining unsupervised feature extraction techniques with deep learning algorithms because they can discover segment-level latent variables. These features can help improve performance, particularly for long duration seizure events (e.g., seizures lasting longer than 60 seconds). We have a vast amount of unlabeled data in the TUH EEG Corpus that we cannot use for training. Currently, we are limited to a relatively small amount of annotated data. Unsupervised techniques offer the potential for processing the entire corpus.

We can also optimize the deep learning architecture for these unsupervised features. We showed that features extracted using unsupervised deep learning algorithms for time-series data are better than hand-engineered features or features from traditional machine learning algorithms. We also explored classification algorithms for several EEG tasks. We learned about various preprocessing methods such as low pass filtering and artifact removal using manual and automatic techniques. Depending on the input types (e.g., calculated features, images, or raw signals), the network design and parameters varied significantly. CNN and recurrent neural networks performed better than other types of neural networks.

The features produced from AE-CDNN have proven to be easier to learn compared to hand-engineered features. By using default parameters of different classifiers, the features from this model were able to detect seizures with up to 95% accuracy. These features showed better performance compared to PCA and SRP. Since these features performed with more than 92% accuracy on two datasets, we believe an AE-CDNN approach has the potential to improve state of the art in seizure detection for sequential EEG data. With sliding windows, a channel can be reduced to multiple segments from which AE-CDNN can extract features.

The FHVAE model needs further investigation if it is to be successfully applied to EEG signals. FHVAE model used filter bank energies, rather than raw speech data for unsupervised feature extraction. As the seizure patterns are patient dependent, FHVAE will consider the patient information as sequence-level information and discard it. It will also be able to detect the artifacts as sequence-level features. We will then investigate if the extracted segment-level features can identify seizure events.

# References

Abraham, B., & Nair, M. S. (2018). Computer-aided classification of prostate cancer grade groups from MRI images using texture features and stacked sparse autoencoder. *Computerized Medical Imaging and Graphics*, 69, 60–68. *https://doi.org/10.1016/j.compmedimag.2018.08.006*.

Ahmad, T., Fairuz, R. A., Zakaria, F., & Isa, H. (2008). Selection of a subset of EEG channels of epileptic patient during seizure using PCA. *Proceedings of the 7th WSEAS International Conference on Signal Processing, Robotics and Automation,* 270–273. Cambridge, UK. *https://pdfs.semanticscholar.org/2b14/685ebcb545ddb4235031e467df45b10b324e.pdf.*

Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., & Elger, C. E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6), 061907. *https://doi.org/10.1103/PhysRevE.64.061907.*

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning Long-Term Dependencies with Gradient Descent is Difficult. *IEEE Transactions on Neural Networks*, 5(2), 157–166. *https://doi.org/10.1109/72.279181.*

Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 245–250. *https://doi.org/10.1145/502512.502546*.

Chai, X., Wang, Q., Zhao, Y., Liu, X., Bai, O., & Li, Y. (2016). Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition. *Computers in Biology and Medicine*, 79, 205–214. *https://doi.org/10.1016/j.compbiomed.2016.10.019.*

Cheng, S., & Sabes, P. N. (2006). Modeling sensorimotor learning with linear dynamical systems. *Neural Computation*, 18(4), 760–793. *https://doi.org/10.1162/089976606775774651.*

Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering*, 16(3), 031001. *https://doi.org/10.1088/1741-2552/ab0ab5.*

Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business*, 38(1), 34 –105. *http://www.jstor.org/stable/2350752.*

Fei-Fei, L., Johnson, J., & Yeung, S. (2016). Convolutional Neural Networks (CNNs / ConvNets). *http://cs231n.github.io/convolutional-networks/.*

Gadhoumi, K., Lina, J.-M., Mormann, F., & Gotman, J. (2016). Seizure prediction for therapeutic devices: A review. *Journal of Neuroscience Methods*, 260 (Supplement C), 270–282. *https://doi.org/10.1016/j.jneumeth.2015.06.010*.

Haufe, S., Nikulin, V. V, Müller, K.-R., & Nolte, G. (2013). A critical assessment of connectivity measures for EEG data: A simulation study. *NeuroImage*, *64*, 120–133. *https://doi.org/10.1016/j.neuroimage.2012.09.036.*

Hsu, W. N., & Glass, J. (2018). Extracting Domain Invariant Features by Unsupervised Learning for Robust Automatic Speech Recognition. *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 5614–5618. Calgary, AB, Canada: IEEE. *https://doi.org/10.1109/ICASSP.2018.8462037.*

Hsu, W. N., Zhang, Y., & Glass, J. (2017). Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data. *Advances in Neural Information Processing Systems,* 1876-1887. Long Beach, California, USA. *https://papers.nips.cc/paper/6784-unsupervised-learning-of-disentangled-and-interpretable-representations-from-sequential-data.pdf.*

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). Melbourne, Australia: OTexts. *https://otexts.com/fpp2/.*

Kingma, D. P., & Ba, J. L. (2015). Adam: a Method for Stochastic Optimization. *Proceedings of the International Conference on Learning Representations* (pp. 1–15). San Diego, California, USA. *https://arxiv.org/abs/1412.6980.*

Längkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42(1), 11–24. *https://doi.org/10.1016/j.patrec.2014.01.008.*

LeCun, Y. (2019). 1.1 Deep Learning Hardware: Past, Present, and Future. *Proceedings of the IEEE International Solid- State Circuits Conference (ISSCC),* 12–19. San Francisco, CA, USA, USA: IEEE. *https://doi.org/10.1109/ISSCC.2019.8662396.*

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, 521(7553), 436–444. *https://doi.org/10.1038/nature14539.*

Li, P., Hastie, T. J., & Church, K. W. (2006). Very sparse random projections. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 287–296). Philadelphia, PA, USA: ACM. *https://doi.org/10.1145/1150402.1150436.*

Light, G. A., Williams, L. E., Minow, F., Sprock, J., Rissling, A., Sharp, R., Braff, D. L. (2010). Electroencephalography (EEG) and event-related potentials (ERPs) with human participants. *Current Protocols in Neuroscience*, 52(1), 6–25. *https://doi.org/10.1002/0471142301.ns0625s52.*

Lin, Q., and Ye, S., and Huang, X., and Li, S., and Zhang, M., and Xue, Y., & and Chen, W.-S. (2016). Classification of Epileptic EEG Signals with Stacked Sparse Autoencoder Based on Deep Learning. *Intelligent Computing Methodologies* (pp. 802–810). Springer International Publishing. *https://doi.org/10.1007/978-3-319-42297-8\_74.*

Masci, J., Meier, U., Cireşan, D., & Schmidhuber, J. (2011). Stacked Convolutional Auto-encoders for Hierarchical Feature Extraction. In *Proceedings of the 21th International Conference on Artificial Neural Networks - Volume Part I,* 52–59. Espoo, Finland: Springer-Verlag. *http://dl.acm.org/citation.cfm?id=2029556.2029563.*

Pearce, T. C., Schiffman, S. S., & Nagle, H. T. (Eds.). (2006). *Handbook of Machine Olfaction: Electronic Nose Technology*. John Wiley & Sons. *https://doi.org/10.1002/3527601597.*

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, 97, 205–227. *https://doi.org/10.1016/j.eswa.2017.12.020.*

Qui, Y., Zhou, W., & Du, P. (2018). Denoising Sparse Autoencoder-Based Ictal EEG Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(9), 1717–1726. *https://doi.org/10.1109/TNSRE.2018.2864306.*

Shah, V., Anstotz, R., Obeid, I., & Picone, J. (2018). Adapting an Automatic Speech Recognition System to Event Classification of Electroencephalograms. I. Obeid & J. Picone (Eds.), *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (p. 1). Philadelphia, Pennsylvania, USA. *https://doi.org/10.1109/SPMB.2018.8615625*.

Shoeb, A. H., & Guttag, J. V. (2010). Application of machine learning to epileptic seizure detection. *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 975–982). Haifa, Israel. *https://doi.org/10.1016/j.jneumeth.2010.05.020.*

Thodoroff, P., Pineau, J., & Lim, A. (2016). Learning Robust Features using Deep Learning for Automatic Seizure Detection. *Proceedings of Machine Learning Research (PMLR)*, 178-190. *http://proceedings.mlr.press/v56/Thodoroff16.html.*

Tschannen, M., Bachem, O., & Lucic, M. (2018). Recent Advances in Autoencoder-Based Representation Learning. In *Neural Information Processing Systems (NeurIPS) Workshop on Bayesian Deep Learning,* 1–26. Montreal, Canada. *http://arxiv.org/abs/1812.05069*.

Wen, T., & Zhang, Z. (2018). Deep Convolution Neural Network and Autoencoders-Based Unsupervised Feature Learning of EEG Signals. *IEEE Access*, 6, 25399–25410. *https://doi.org/10.1109/ACCESS.2018.2833746.*

Williamson, J. R., Bliss, D. W., Browne, D. W., & Narayanan, J. T. (2012). Seizure prediction using EEG spatiotemporal correlation structure. *Epilepsy & Behavior*, 25(2), 230–238. *https://doi.org/10.1016/j.yebeh.2012.07.00.*

Yuan, Y., Xun, G., Jia, K., & Zhang, A. (2017). A Multi-view Deep Learning Method for Epileptic Seizure Detection Using Short-time Fourier Transform. *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 213–222). Boston, Massachusetts, USA: ACM. *https://doi.org/10.1145/3107411.3107419.*