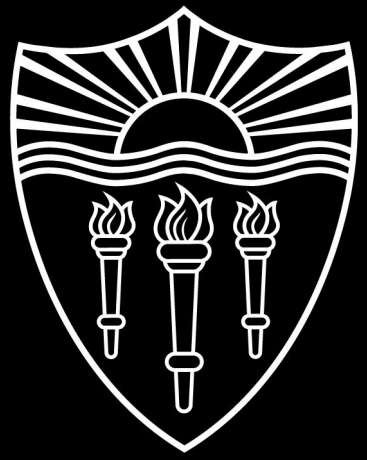
Unsupervised Multivariate Time-Series Transformers for Seizure Identification on EEG

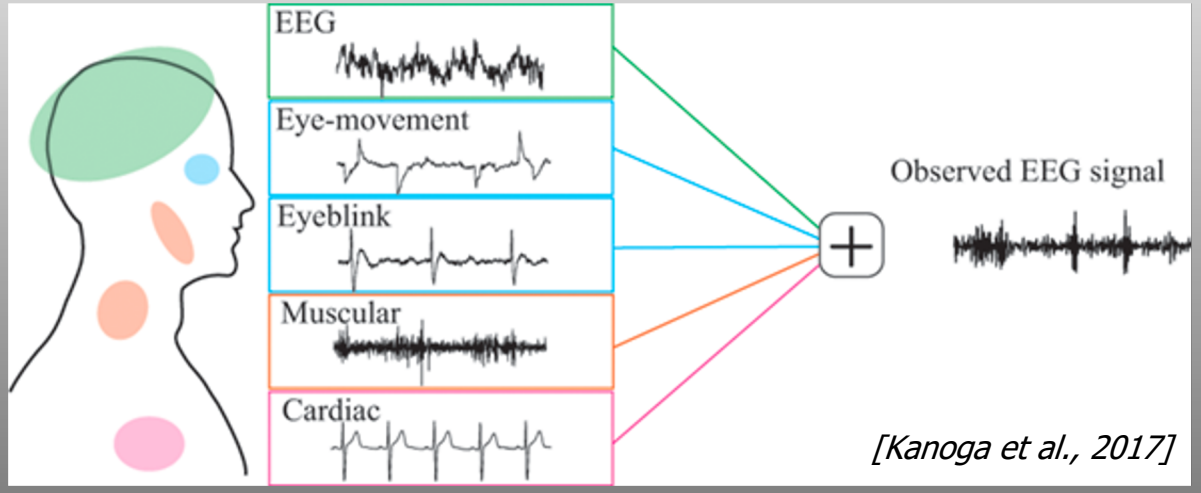


Dominique Duncan
Assistant Professor of Neurology, Neuroscience, and Biomedical Engineering
Laboratory of Neuro Imaging
USC Stevens Neuroimaging and Informatics Institute
University of Southern California

USC Mark and Mary Stevens Neuroimaging and Informatics Institute

Motivation for Automated Seizure Identification

Epilepsy is one of the most common neurological disorders.
Successful identification of early seizures can initiate antiepileptogenic intervention and therapies.
Electroencephalogram (EEG) recordings are commonly used for seizure identification yet are known to contain many artifacts.
Seizure identification via manual inspection is laborious and difficult, motivating automated methods for aiding experts.

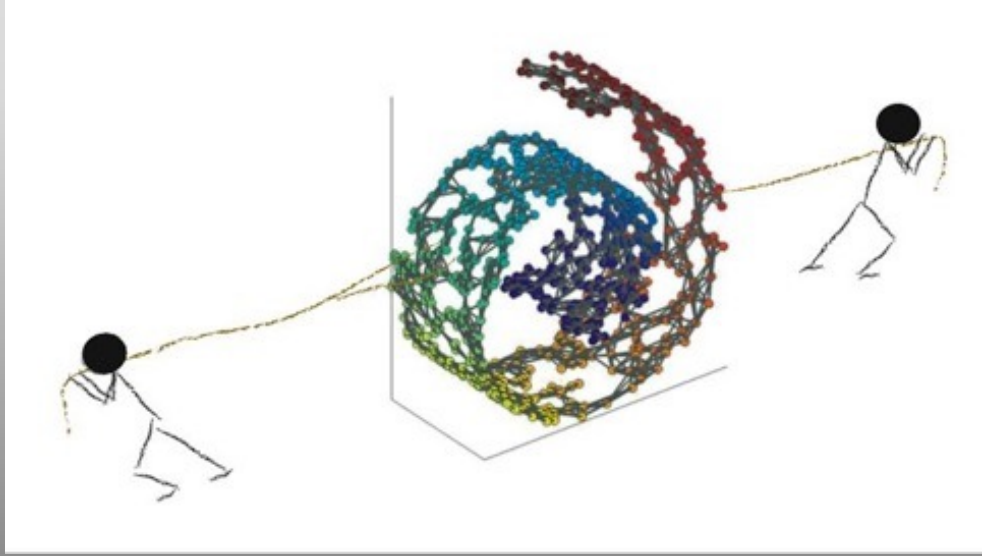


Manifold Learning Approach

Reduce the number of features while retaining the maximum amount of information

Compute a kernel with a specially-tailored distance measure

Employ manifold learning technique to recover the underlying information

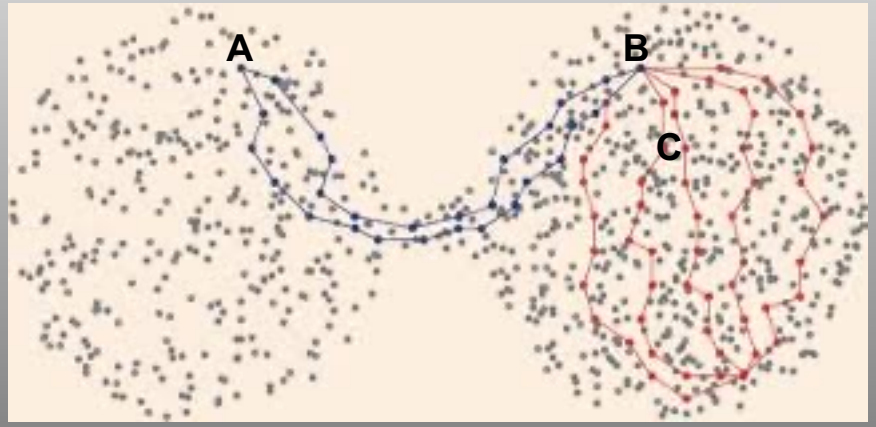


Diffusion Maps

Eigenfunctions of Markov matrices are used to construct coordinates that generate efficient representations of complex geometric structures

- **Affinity Matrix, K:**
  - $K_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\epsilon}}$, where ϵ is a kernel scale and x_i are data points.
- **Diffusion Operator, P:** Transition probabilities between data points.
  - Constructed by normalizing K: $P = D^{-1}K$, where D is a diagonal matrix.
- **Eigenvalue Decomposition:** $P\psi_k = \lambda_k \psi_k$.
  - λ_k and ψ_k are the eigenvalues and eigenvectors of P respectively.

- Compute pairwise distances between data points.
- Construct affinity matrix *K* using Gaussian kernel.
- Formulate diffusion operator *P*.
- Perform eigenvalue decomposition on *P*.
- Embed data points using:
  $y_i=[\lambda_1\psi_1(x_i),\lambda_2\psi_2(x_i),...]$ $y_i=[\lambda_1\psi_1(x_i),\lambda_2\psi_2(x_i),...]$



Benefit of Diffusion Maps over Related Methods

Nonlinearity

- PCA, for example, looks for directions of maximal variance in a linear manner, whereas diffusion maps can uncover more intricate structures and relationships that may be missed by PCA

Geometric Representation

- Embedding respects the intrinsic geometry of the data manifold

Locality Preservation

- Emphasis on local relationships between data points (i.e., important in understanding local interactions, such as how a seizure initiates and propagates through neighboring brain regions)

Interpretability

- Simulates a random walk on the data's manifold – can provide a more intuitive understanding of the data's structure and how different regions communicate

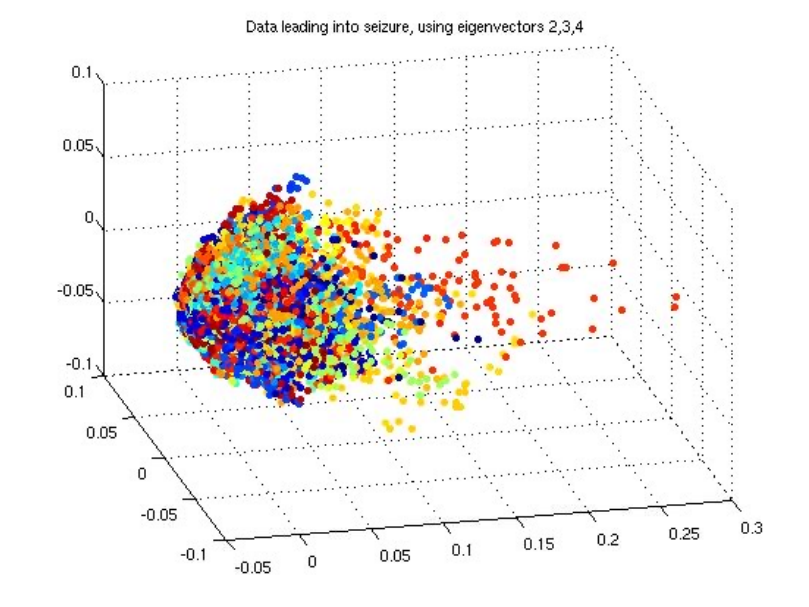
Robustness to Noise

- Especially robust to noise when the noise is not uniformly distributed across the data space

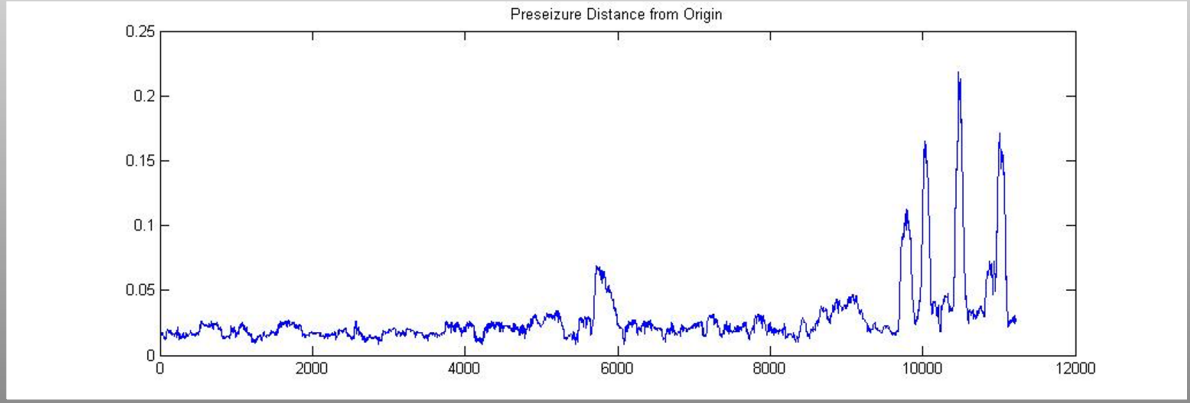
Data Clustering

- Can highlight regions of high data density and barriers between them

Diffusion Maps Example



Duncan et al., Math. Biosci. Eng., 2013



Unsupervised Diffusion Component Analysis

Algorithm

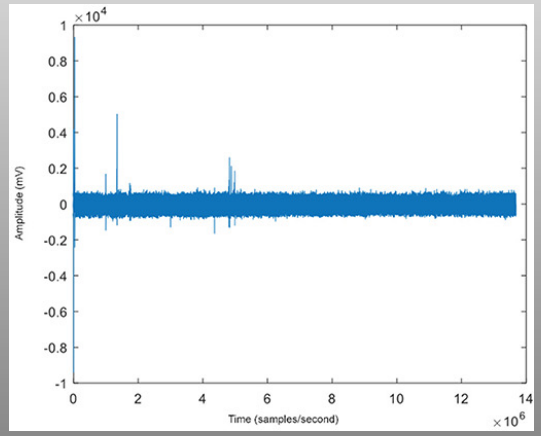
1: Obtain EEG data of patient post TBI with n electrode contacts,
2: Create m-second windows of data, overlapping by 50%,
3: Reshape each small submatrix into a vector; place each vector side by side to form a matrix,
4: Compute histograms (along matrix columns) using 20 bins,
5: Calculate the Earth Mover's Distance between consecutive feature vectors,
6: To reduce the chance of bias, introduce a random shuffle in the columns of the matrix and apply a random projection,
7: Apply the Discrete Cosine Transform,
8: Calculate local covariance matrices for overlapping windows,
9: Compute the eigenvalue decomposition to obtain eigenvalues and corresponding eigenvectors,
10: Calculate inverse covariance matrices to calculate the Mahalanobis Distance,
11: Use the median of all pairwise distances of the data matrix to choose epsilon, the Gaussian kernel scale,
12: Compute the affinity matrix and build a Gaussian kernel according to (5),
13: Normalize the kernel by a diagonal density matrix and employ eigenvalue decomposition to obtain the eigenvalues and eigenvectors,
14: Consider all possible combinations of 3 or 4 eigenvectors for the embeddings; compute the center of mass for each embedding as well as the variance of the embedded points to determine the optimal embedding as the one with the largest spread among the embedded points.

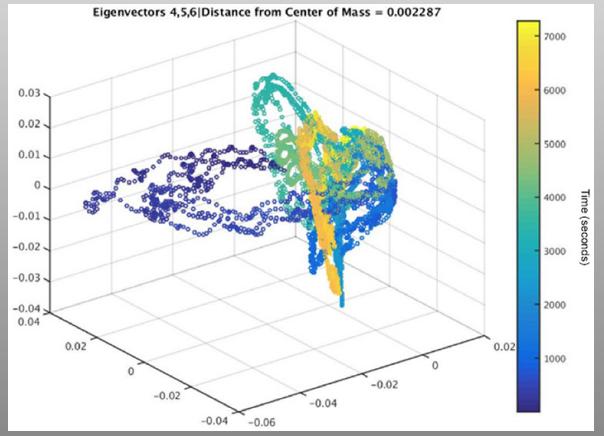
Duncan et al., Discrete Continuous Dynamical Systems, 2019

Unsupervised Diffusion Component Analysis

Cross-correlation between segments is calculated to ensure minimal variance to ensure similar behavior between the channels that were being analyzed.

The Mahalanobis distance is applied to inverse covariance matrices that are computed using the SVD to identify outliers; the combination of the Mahalanobis distance and inverse covariance matrices has previously been shown to be a successful tool for denoising data.





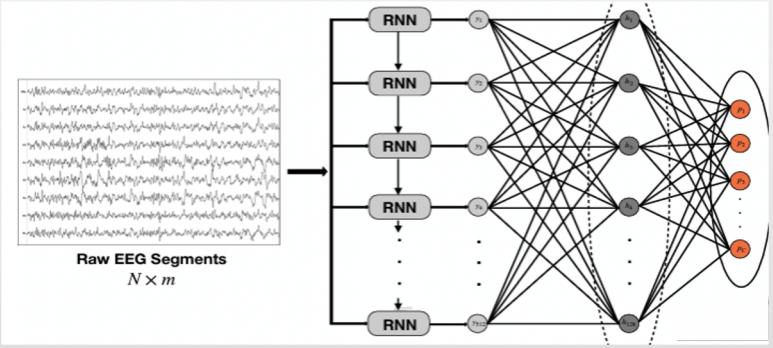
Duncan et al., Discrete Continuous Dynamical Systems, 2019

Motivation for Unsupervised Learning

Supervised methods have been extensively explored.

- Spatiotemporal feature extraction
- More automation via deep learning models (CNNs, RNNs, …)

[Faghihpirayesh et al., 2021]

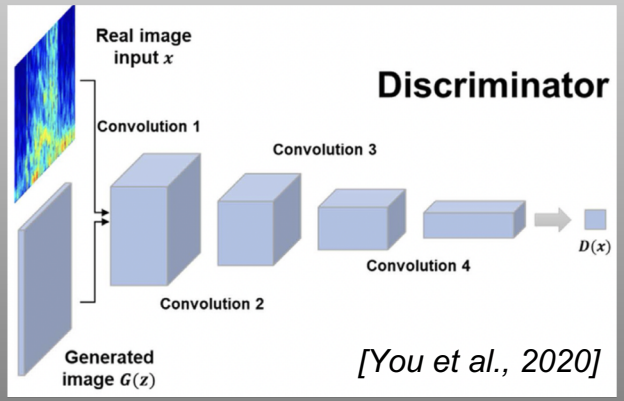


Stochastic nature of EEG limits access to seizure annotations that are:

- Large enough to combat imbalance towards non-seizure data
- Consistent across different experts

Unsupervised deep models have been limited.

- GANs that require manual feature extraction
- CNNs that are not tailored for multivariate time-series data



[You et al., 2020]

Unsupervised Multivariate Time-Series Transformers for Seizure Identification on EEG

Ilkay Yıldız Potter (*BioSensics LLC*), George Zerveas & Carsten Eickhoff (*Brown University, Computer Science Department*), Dominique Duncan

First *unsupervised* transformer-based model for seizure identification on *raw* EEG

Pose seizure identification as an anomaly detection problem

- Train an autoencoder involving a transformer encoder via an unsupervised loss function on non-seizure signals
- Incorporate a novel masking strategy uniquely designed for modeling multivariate time-series (MVTS) data
- Identify seizures via reconstruction errors at inference time

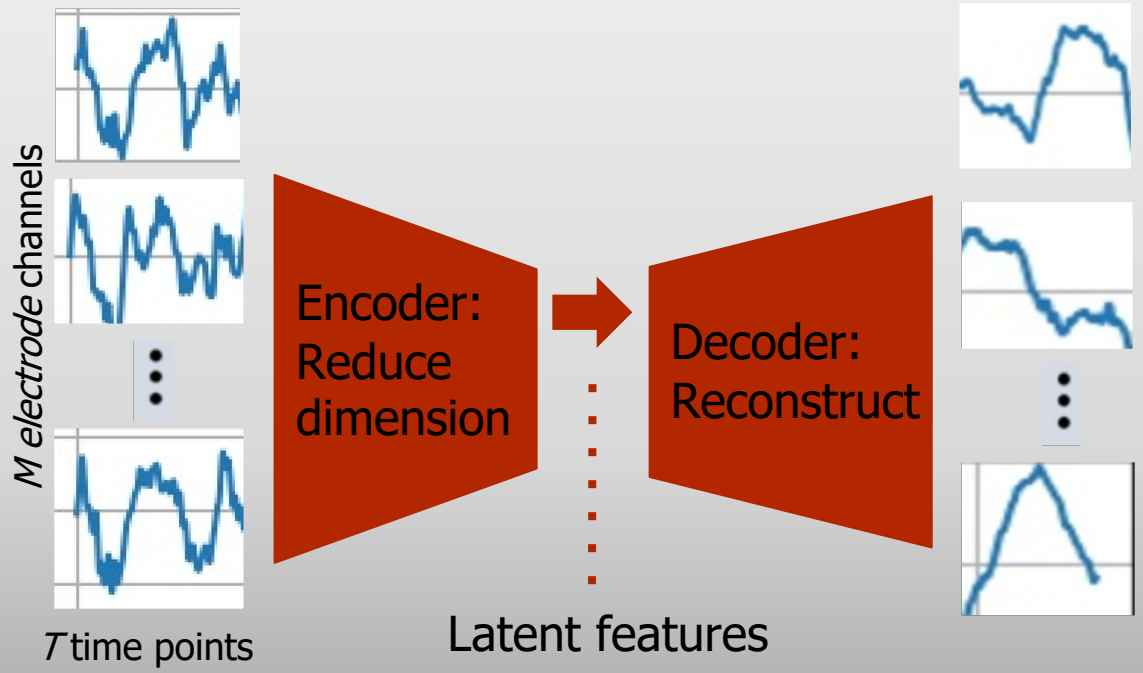
Evaluated on three publicly-available benchmark EEG datasets

- Outperform supervised learning counterparts
- Particular benefit for learning from highly imbalanced data

USC Mark and Mary Stevens Neuroimaging and Informatics Institute

Seizure Identification as Anomaly Detection

Multivariate EEG signal (extracted as sliding windows)



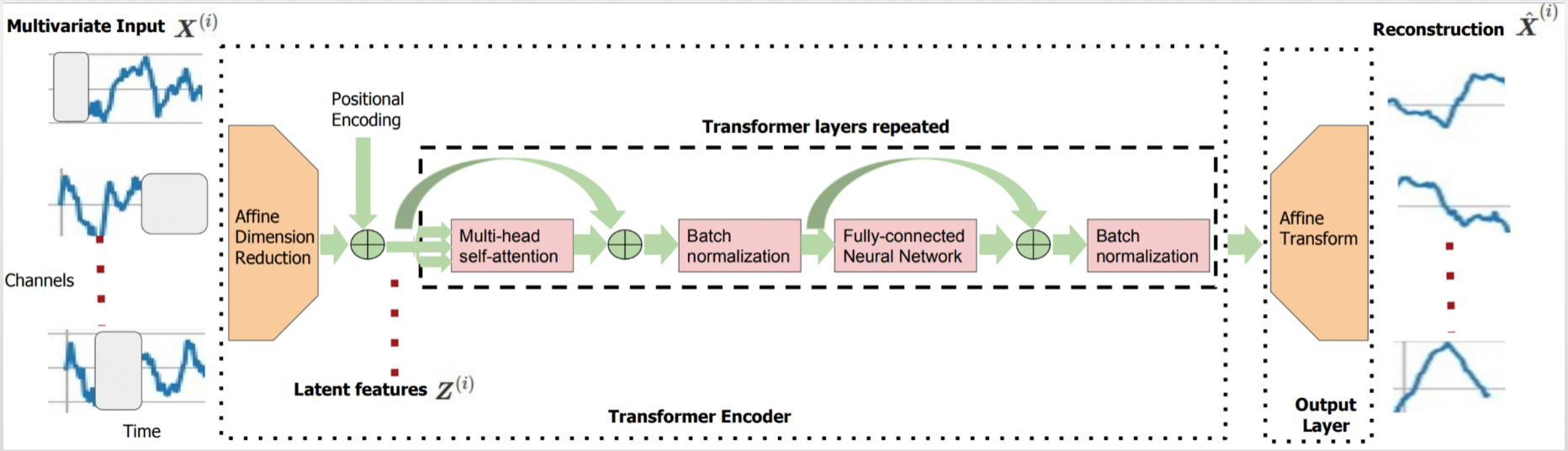
An autoencoder to learn latent features that generate EEG

Train on non-seizure signals (e.g. from healthy subjects)

Seizure windows belong to a different distribution: larger reconstruction errors!

Use the average error over channels and time as the seizure probability score of a window

Autoencoder Architecture and Training Objective



A transformer encoder and a linear decoder

- Encoder architecture by Vaswani et al. (2017), with fully-trainable positional encoding and batch normalization [*Zerveas et al. (2021)*].

A training objective designed for MVTS data [*Zerveas et al. (2021)*]

$$\frac{1}{|\mathcal{M}|} \sum_{(t,m)\in\mathcal{M}} (X_{t,m}^{(i)} - \hat{X}_{t,m}^{(i)})^2$$

- Mask a proportion of each window (set $\mathcal{M}$) and use for optimizing the reconstruction objective
- Alternate between masked and unmasked sequences, with lengths distributed geometrically

Datasets

MIT-BCH *[Shoeb et al., 2009]*

- Scalp, 38 channels

UPenn and MayoClinic, 2014

- Intracranial, 72 channels

TUH *[Obeid and Picone, 2016]*

- Scalp, 38 channels

Preprocessing

- Unify sampling rate with respect to the smallest
- Bandpass filter with range 0.5-50 Hz to eliminate powerline noise
- Extract sliding windows with 50% overlap and at length of shortest seizure
  - MIT: 13,600 non-seizure and 963 seizure windows
  - UPenn: 14,329 non-seizure and 1307 seizure windows
  - TUH: 54,264 non-seizure and 2826 seizure windows

- **Severely imbalanced towards non-seizure**

Experiment Setup

Stratified partition of all windows, 60% training, 20% validation, 20% testing

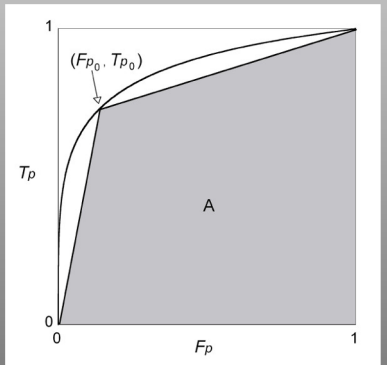
Unsupervised Methods (Train on non-seizure windows only)

- Deep models
  - Proposed autoencoder
  - Convolutional variational autoencoder *[Yildiz et al., 2022]*
- Shallow models
  - t-SNE dimension reduction *[Van der Maaten and Hinton, 2008]* => K-means

Supervised Methods (Train on both window types)

- Deep models optimized with cross-entropy
  - Transformer encoder => linear classification layer
  - Unsupervised pre-training => fine-tuning *[Zerveas et al., 2021]*
- Shallow models
  - XGBoost *[Chen and Guestrin, 2016]*
  - ROCKET *[Dempster et al. 2020]*

Seizure vs. non-seizure classification by thresholding at elbow of ROC curve



Results Against Unsupervised Baselines

Dataset	Method	Precision	Recall	Accuracy	AUC
MIT	**Unsupervised Transformer**	$\mathbf{0.98 \pm 0.003}$	$\mathbf{0.9 \pm 0.006}$	$\mathbf{0.87 \pm 0.006}$	$\mathbf{0.94 \pm 0.023}$
	Unsupervised K-means	0.33 ± 0.008	0.5 ± 0.009	0.5 ± 0.009	0.59 ± 0.041
	Unsupervised VAE	0.97 ± 0.003	0.75 ± 0.008	0.61 ± 0.009	0.61 ± 0.041
	Supervised XGBoost	0.98 ± 0.003	0.8 ± 0.007	0.8 ± 0.007	0.88 ± 0.031
	Supervised ROCKET	0.98 ± 0.003	0.74 ± 0.008	0.78 ± 0.008	0.86 ± 0.032
	Supervised Transformer	0.98 ± 0.003	0.83 ± 0.007	0.83 ± 0.007	0.88 ± 0.031
	Pre-trained 50% Supervised Transformer	0.97 ± 0.003	0.72 ± 0.008	0.63 ± 0.009	0.66 ± 0.021
	Pre-trained 100% Supervised Transformer	$\mathit{0.99 \pm 0.002}$	$\mathit{0.98 \pm 0.003}$	$\mathit{0.94 \pm 0.005}$	$\mathit{0.97 \pm 0.017}$
UPenn	**Unsupervised Transformer**	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.76 \pm 0.013}$	$\mathbf{0.68 \pm 0.014}$	$\mathbf{0.73 \pm 0.027}$
	Unsupervised K-means	0.33 ± 0.014	0.5 ± 0.015	0.5 ± 0.015	0.56 ± 0.028
	Unsupervised VAE	0.8 ± 0.012	0.5 ± 0.015	0.49 ± 0.015	0.47 ± 0.027
	Supervised XGBoost	0.87 ± 0.01	0.62 ± 0.015	0.6 ± 0.015	0.65 ± 0.028
	Supervised ROCKET	0.87 ± 0.01	0.67 ± 0.014	0.62 ± 0.015	0.67 ± 0.028
	Supervised Transformer	0.87 ± 0.01	0.69 ± 0.014	0.62 ± 0.015	0.64 ± 0.028
	Pre-trained 50% Supervised Transformer	0.86 ± 0.011	0.77 ± 0.013	0.63 ± 0.015	0.64 ± 0.032
	Pre-trained 100% Supervised Transformer	$\mathit{0.92 \pm 0.008}$	$\mathit{0.85 \pm 0.011}$	$\mathit{0.82 \pm 0.012}$	$\mathit{0.89 \pm 0.02}$

Dramatic improvement against other unsupervised methods

- outperform state-of-the-art deep learning counterpart VAE by up to 33% AUC
- K-means classifies all windows as non-seizure

** Our model in bold, best model in italics*

Benefit Against Supervised Baselines

Dataset	Method	Precision	Recall	Accuracy	AUC
MIT	**Unsupervised Transformer**	**0.98 ± 0.003**	**0.9 ± 0.006**	**0.87 ± 0.006**	**0.94 ± 0.023**
	Unsupervised K-means	0.33 ± 0.008	0.5 ± 0.009	0.5 ± 0.009	0.59 ± 0.041
	Unsupervised VAE	0.97 ± 0.003	0.75 ± 0.008	0.61 ± 0.009	0.61 ± 0.041
	Supervised XGBoost	0.98 ± 0.003	0.8 ± 0.007	0.8 ± 0.007	0.88 ± 0.031
	Supervised ROCKET	0.98 ± 0.003	0.74 ± 0.008	0.78 ± 0.008	0.86 ± 0.032
	Supervised Transformer	0.98 ± 0.003	0.83 ± 0.007	0.83 ± 0.007	0.88 ± 0.031
	Pre-trained 50% Supervised Transformer	0.97 ± 0.003	0.72 ± 0.008	0.63 ± 0.009	0.66 ± 0.021
	Pre-trained 100% Supervised Transformer	*0.99 ± 0.002*	*0.98 ± 0.003*	*0.94 ± 0.005*	*0.97 ± 0.017*
UPenn	**Unsupervised Transformer**	**0.88 ± 0.01**	**0.76 ± 0.013**	**0.68 ± 0.014**	**0.73 ± 0.027**
	Unsupervised K-means	0.33 ± 0.014	0.5 ± 0.015	0.5 ± 0.015	0.56 ± 0.028
	Unsupervised VAE	0.8 ± 0.012	0.5 ± 0.015	0.49 ± 0.015	0.47 ± 0.027
	Supervised XGBoost	0.87 ± 0.01	0.62 ± 0.015	0.6 ± 0.015	0.65 ± 0.028
	Supervised ROCKET	0.87 ± 0.01	0.67 ± 0.014	0.62 ± 0.015	0.67 ± 0.028
	Supervised Transformer	0.87 ± 0.01	0.69 ± 0.014	0.62 ± 0.015	0.64 ± 0.028
	Pre-trained 50% Supervised Transformer	0.86 ± 0.011	0.77 ± 0.013	0.63 ± 0.015	0.64 ± 0.032
	Pre-trained 100% Supervised Transformer	*0.92 ± 0.008*	*0.85 ± 0.011*	*0.82 ± 0.012*	*0.89 ± 0.02*

Despite the lack of seizure labels during training:

- better than all supervised & 50% fine-tuned models by up to 9% AUC
- better precision-recall balance

- The most expensive transformer (unsupervised pre-training => supervised fine-tuning with *all* training labels) is naturally better, albeit by a smaller margin against unsupervised learning via our method.

** Our model in bold, best model in italics*

Results on TUH

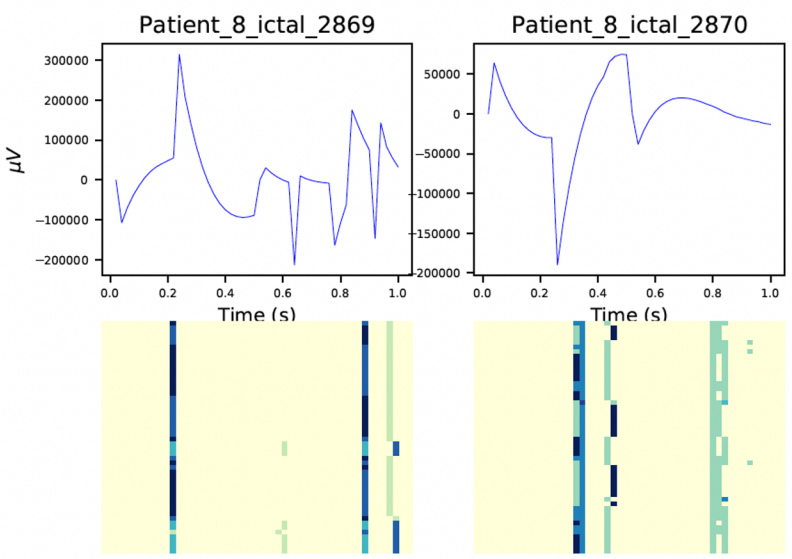
TUH	**Unsupervised Transformer**	$\mathbf{0.92 \pm 0.005}$	$\mathbf{0.57 \pm 0.009}$	$\mathbf{0.61 \pm 0.009}$	$\mathbf{0.57 \pm 0.013}$
	Unsupervised K-means	0.17 ± 0.007	0.5 ± 0.009	0.35 ± 0.008	0.57 ± 0.013
	Unsupervised VAE	$\mathit{0.93 \pm 0.005}$	$\mathit{0.86 \pm 0.006}$	$\mathit{0.83 \pm 0.007}$	$\mathit{0.86 \pm 0.009}$
	Supervised XGBoost	0.93 ± 0.005	0.73 ± 0.008	0.71 ± 0.008	0.78 ± 0.011
	Supervised ROCKET	0.93 ± 0.005	0.7 ± 0.008	0.66 ± 0.008	0.74 ± 0.012
	Supervised Transformer	0.92 ± 0.005	0.37 ± 0.009	0.54 ± 0.009	0.52 ± 0.012
	Pre-trained 50% Supervised Transformer	0.94 ± 0.005	0.61 ± 0.009	0.75 ± 0.008	0.71 ± 0.025
	Pre-trained 100% Supervised Transformer	0.93 ± 0.005	0.66 ± 0.008	0.7 ± 0.008	0.72 ± 0.012

TUH is particularly challenging (a compilation of several EEG databases collected over years)

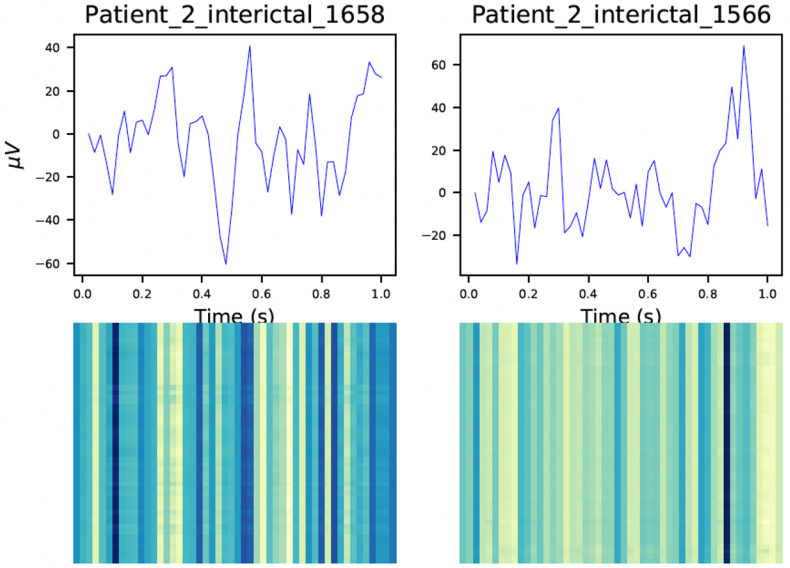
- Our unsupervised transformer still fares better than the purely supervised counterpart
- Unsupervised VAE is the best, motivating unsupervised learning as we propose.

** Our model in bold, best model in italics*

Benefit of Attention within Transformer



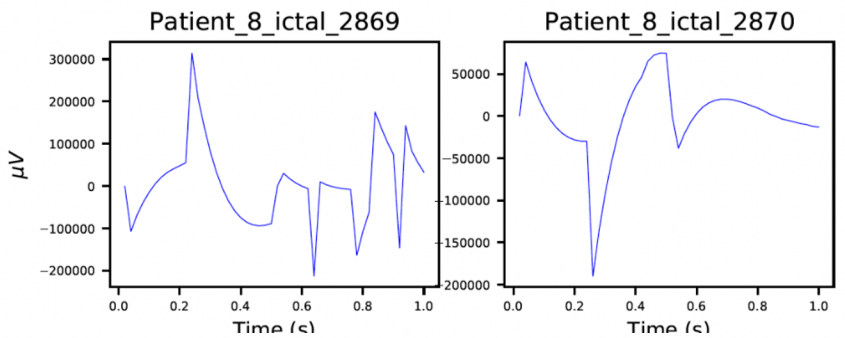
(a) Correctly Identified Seizure Windows (True Positive)



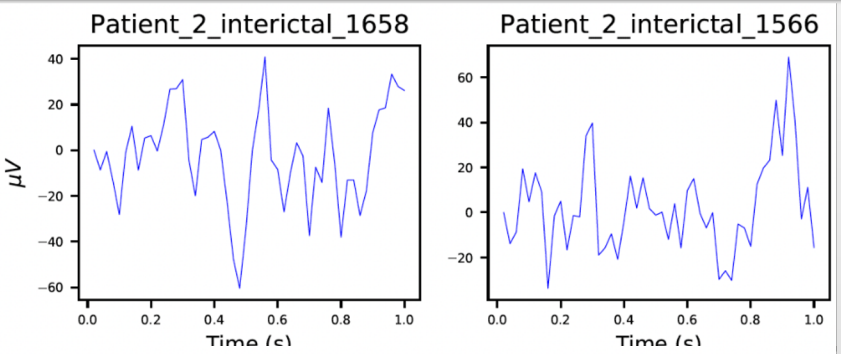
(b) Correctly Identified Non-seizure Windows (True Negative)

- Can successfully learn to pay more attention to seizure patterns including high-frequency spikes and waves evolving with large amplitudes
- When seizure is deemed to exist, patterns of focused attention, containing only few time points with large weights, aiding explainability

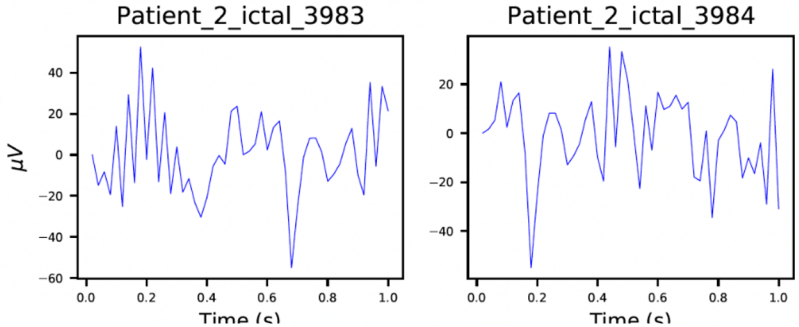
Analysis of Example Predictions



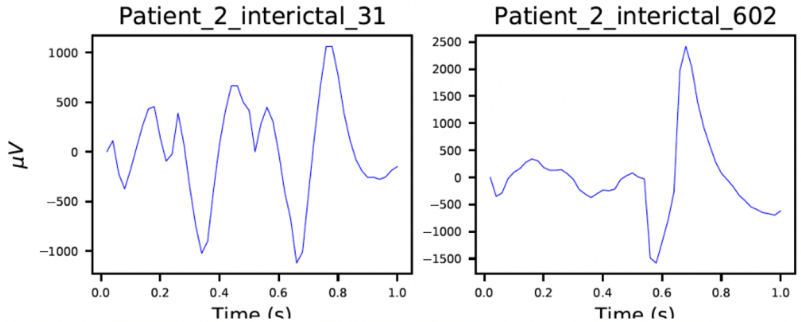
(a) Correctly Identified Seizure Windows (True Positive)



(b) Correctly Identified Non-seizure Windows (True Negative)



(c) Falsely Identified Seizure Windows (False Negative)



(d) Falsely Identified Non-seizure Windows (False Positive)

- Seizure patterns cannot be easily identified using only amplitude/frequency, motivating a more sophisticated approach such as ours
  - non-seizure windows in (d) have a larger amplitude range than the seizure windows (c)
  - seizure windows in (c) contain similar spikes to the non-seizure windows in (b) w.r.t. amplitude and frequency

Conclusion

Summary:

Unsupervised method for seizure identification on raw EEG
Train an autoencoder involving a transformer encoder to reconstruct stochastically-masked EEG recordings
Seizures are identified based on higher reconstruction errors
Even outperform state-of-the-art supervised methods requiring expert labels

Potential impacts:

Can alleviate the burden on clinical experts regarding laborious and difficult EEG inspections to provide labels indicating segments that contain seizures
Can aid availability of seizure diagnoses for the wider public, especially in areas where access to well-trained healthcare professionals is limited

Thank you!

This work was supported by the National Institutes of Health (NIH) National Institute of Neurological Disorders and Stroke (NINDS) grant R01NS111744.

Lab Website: https://sites.usc.edu/duncanlab/

Email: duncand@usc.edu

USC Mark and Mary Stevens Neuroimaging and Informatics Institute