Unsupervised Multivariate Time-Series Transformers for Seizure Identification on EEG

***Abstract*—Epilepsy is one of the most common neurological disorders, typically observed via epileptic seizures. Seizure identification is commonly monitored through Electroencephalogram (EEG) recordings due to their routine and low expense collection. The stochastic nature of EEG makes seizure identification via manual inspections performed by highly-trained experts a tedious endeavor, motivating the use of automated identification. The literature on automated identification focuses mostly on supervised learning methods requiring expert labels indicating EEG segments that contain seizures, which are difficult to obtain. Motivated by these observations, we pose seizure identification as an unsupervised anomaly detection problem. To this end, we employ the first fully-unsupervised transformer-based model for seizure identification on raw EEG. We train an autoencoder involving a transformer encoder via an unsupervised loss function, incorporating a novel masking strategy uniquely designed for multivariate time-series data such as EEG. Training employs EEG recordings that do not contain any seizures, while seizures are identified with respect to mean reconstruction errors at inference time. We evaluate our method on three publicly available benchmark EEG datasets for distinguishing seizure vs. non-seizure windows. Our method leads to significantly *better seizure identification performance than the supervised learning counterparts*, by up to 16% recall, 9% accuracy, and 9% Area under the Receiver Operating Characteristics Curve (AUC), establishing a particular benefit on highly imbalanced data.**

***Index Terms*—Epilepsy, Seizure, EEG, Unsupervised Learning, Time-series Transformer**

I. Introduction

Epilepsy is one of the most common neurological disorders, affecting over 70 million people worldwide [1]. Epilepsy patients typically suffer from seizures, involving uncontrolled jerking movements or momentary losses of awareness due to abnormal excessive or synchronous activities in the brain [2]. The degraded quality of life for patients strongly motivates early seizure identification, as early seizures have been shown to be prognostic markers for later epileptogenic development. Successful identification of early seizures can initiate antiepileptogenic intervention and therapies that can remarkably improve the quality of life for patients and their caregivers. To this end, electroencephalogram (EEG) recordings received particular attention for seizure identification [3], due to their routine and low expense collection compared to, e.g., neuroimaging. Seizures on EEG are defined as generalized spike-wave discharges at three per second or faster, and clearly evolving discharges of any type that reach a frequency of four per second or faster.

Despite their volume and rich information content, EEG recordings are known to contain many artifacts due to movement, physiological activity such as perspiration, and measurement hardware [4], [5]. The stochastic nature of clinically-acquired EEG makes seizure identification via manual inspection laborious and difficult, leading to significant variability across clinical labels of different experts [6]. This challenge motivated the recent literature to focus on automated identification of epileptic seizures on EEG as a promising complement to manual inspection. The literature on automated EEG seizure identification is extensive (c.f. Section II), focusing mostly on *supervised* machine learning methods using both manual feature extraction [7]–[10], as well as deep neural networks (DNNs) without manual feature extraction [11]–[13].

Despite their success, supervised methods require expert labels indicating EEG segments that contain seizures, while obtaining large and consistently-labeled EEG datasets is unfavourable due to the stochastic nature of EEG [6]. Difficulty of label collection also leads to severely imbalanced EEG datasets, in which the number of non-seizure recordings significantly exceeds the number of seizure recordings; this poses a further challenge for supervised learning that is prone to overfitting towards dominant class predictions [14].

Unsupervised machine learning methods that do not rely on labeled data have not yet been widely explored. A few methods employed traditional shallow models for unsupervised seizure identification on both raw EEG [15], as well as spatio-temporal features extracted from EEG [16]–[18]. To the best of our knowledge, unsupervised DNN methods for EEG seizure identification have been limited to a couple of recent works, requiring feature extraction prior to training [19] or employing convolutional DNN architectures that are not tailored for multivariate time-series data such as EEG [20].

We propose a *fully-unsupervised* deep learning approach that can identify seizures on *raw* EEG recordings. To this end, we make the following contributions:

- We employ the first unsupervised transformer-based model for seizure identification on raw EEG, inspired by recent advances in multivariate time-series analysis [21].
- We pose seizure identification as an anomaly detection problem. To this end, we train an autoencoder involving a transformer encoder via an unsupervised loss function, incorporating a novel masking strategy uniquely designed for modeling multivariate time-series data such as EEG. As training employs EEG recordings that *do not* contain seizures, seizures are identified via mean reconstruction errors at inference time.
- We extensively validate the seizure identification performance of our method on three publicly available benchmark EEG datasets. Our method can successfully

distinguish between non-seizure vs. seizure windows, with up to *0.94 Area under the Receiver Operating Characteristics Curve (AUC)*. Moreover, our unsupervised anomaly detection approach leads to significantly *better seizure identification performance than the supervised learning counterparts*, by up to 16% recall, 9% accuracy, and 9% AUC, establishing a particular benefit for learning from highly imbalanced data.

II. Related Work

The literature on automated seizure identification on EEG is vast; we refer the reader to the review by [22] for more details. A significant body of works focus on extracting spatio-temporal features from EEG via, e.g., wavelet transformations [7], [23], local mean decomposition [6], Fourier transformations [8], [10], and power spectra [9]. Extracted features are used to train *supervised* machine learning methods, including support vector machines and neural networks, to identify whether a given EEG contains a seizure in a binary classification setting.

Deep neural network (DNN)-based supervised seizure identification methods have lately dominated the literature [24] and obviated the need for manual feature extraction. DNN methods further improved in combination with recurrent neural networks to aid time-series modeling [25], adversarial training to generalize identification across patients [11], autoencoder-based feature extraction [26], [27], and attention mechanisms to improve predictions and interpretability [12].

In recent years, self-attention modules have become an integral part of DNN methods employed in machine vision [28], natural language processing [29], and time-series modeling [21]; the resulting DNN architectures are termed as *transformers*. Transformer architectures have been very recently applied for various identification tasks on EEG, including, e.g., sleep-stage classification , human-computer interface-based action recognition, and seizure identification [13], [30]. These methods employ unsupervised pre-training prior to supervised training on ground-truth expert labels pertaining to the identification task. The unsupervised pre-training objective involves different augmentations of the same EEG segment and trains the transformer by maximizing the similarity of different augmentations of the same segment, while simultaneously minimizing the similarity with different segments.

All in all, the literature on automated seizure identification often focuses on supervised machine learning methods. Despite their success, these methods require expert labels indicating EEG segments that contain seizures, which are difficult to obtain due to the stochastic nature of EEG [6]. Meanwhile, unsupervised machine learning methods that do not rely on labeled data have not yet been widely explored. A few methods applied shallow models for unsupervised seizure identification, including K-means, hierarchical clustering, and Gaussian mixture models, on both raw EEG [15], as well as spatio-temporal features extracted from EEG [16], [17].

Recently, a couple of unsupervised DNN methods for seizure identification on EEG have been proposed. You et al. (2020) preprocess EEGs to extract time-frequency spectrogram images and train a generative adversarial network (GAN) [31] on the spectrograms that do not contain seizures. For each spectrogram at testing time, they have to search for the latent GAN input that leads to the smallest loss value and use the corresponding generated spectrogram for seizure identification. As training involves non-seizure activity, test spectrograms that significantly differ from the spectrograms generated by the GAN are successfully identified to contain seizures. Yıldız et al. (2022) train a convolutional variational autoencoder (VAE) over raw EEG, employing an objective tailored for suppressing EEG artifacts. Unlike You et al. (2020), they identify seizures with respect to the reconstruction errors at inference time.

We differ from the existing works by applying the first *fully-unsupervised transformer-based* model on *raw* EEG. Our architecture and training objective are particularly designed for multivariate time-series analysis and do not require a sophisticated minimax optimization such as GAN training. The fundamental benefit of a transformer encoder over other DNN architectures is that self-attention can selectively highlight important input features and sequence segments, without relying on sequence-aligned convolutions or slow recurrent modules [32]; we also experimentally demonstrate this advantage against the state-of-the-art VAE architecture in Section IV-E.

III. Problem Formulation

We consider a dataset of N EEG recordings, each collected from M electrode channels and consisting of T time points. Formally, we denote each EEG recording by $\boldsymbol{X}^{(i)} \in \mathbb{R}^{T \times M}$, for $i \in [1, \ldots, N]$. Our aim is to design an unsupervised method that does not rely on ground-truth expert labels during learning and can identify the existence of seizures in a given EEG recording. To this end, we employ an autoencoder architecture involving a transformer network encoder that is uniquely designed for multivariate time-series data [21], such as EEG. We note that our method naturally generalizes to EEG recordings comprising different numbers of time points and channels (see our preprocessing setup in Section IV-B).

A. Multivariate Time-Series Transformer

Our autoencoder architecture is based on a transformer encoder and is depicted in Figure 1: the model learns to extract and transform latent features from a given EEG recording in order to reconstruct the stochastically-masked input [21]. Formally, the transformer encoder network receives a recording $\boldsymbol{X}^{(i)} \in \mathbb{R}^{T \times M}$, $i \in [1, \ldots, N]$, and extracts latent features $\boldsymbol{Z}^{(i)} \in \mathbb{R}^{T \times D}$. The output layer applies an affine transformation on $\boldsymbol{Z}^{(i)}$ to reconstruct the recording as $\hat{\boldsymbol{X}}^{(i)} \in \mathbb{R}^{T \times M}$.

1) Transformer Encoder: Transformer encoder operations begin with projecting a recording $\boldsymbol{X}^{(i)}$ from M dimensions to D dimensions via a trainable affine transformation $\boldsymbol{P} \in \mathbb{R}^{M \times D}$. To preserve the ordering information of the input sequence, a fully-trainable positional encoding $\boldsymbol{E} \in \mathbb{R}^{T \times D}$ is added for each input. The resulting latent features extracted from each recording are thus: $\boldsymbol{Z}^{(i)} = \boldsymbol{X}^{(i)}\boldsymbol{P} + \boldsymbol{E}$.

Dimensional projection and positional encoding are followed by the successive application of several transformer layers.

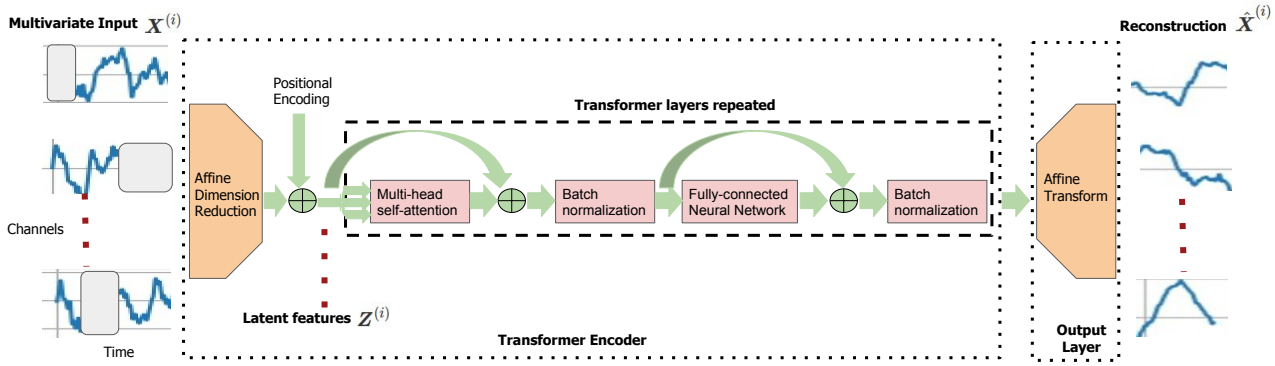


Figure 1: Our autoencoder architecture. The transformer encoder network receives a recording $\boldsymbol{X}^{(i)}$, and extracts latent features $\boldsymbol{Z}^{(i)}$. The output layer applies an affine transformation on $\boldsymbol{Z}^{(i)}$ to reconstruct the recording as $\hat{\boldsymbol{X}}^{(i)} \in \mathbb{R}^{T \times M}$. During training, a proportion of each channel is masked by setting the input values at masked time points (shaded in gray) to 0.

Each transformer layer consists of a multi-headed self-attention (MSA) module, a stochastic dropout operation $\tilde{d}$ [33], batch normalization (Norm) [34], and a fully-connected network (FCN) consisting of two linear layers separated by a GELU [35] activation, a non-linearity designed to be used in combination with dropout and batch normalization. Formally, latent features are updated by each transformer layer via:

$$\begin{aligned} \boldsymbol{Z}^{(i)} &\leftarrow \text{Norm}\left(\tilde{d}\left(\text{MSA}(\boldsymbol{Z}^{(i)})\right) + \boldsymbol{Z}^{(i)}\right), \\ \boldsymbol{Z}^{(i)} &\leftarrow \text{Norm}\left(\tilde{d}\left(\text{FCN}(\boldsymbol{Z}^{(i)})\right) + \boldsymbol{Z}^{(i)}\right). \end{aligned} \tag{1}$$

The summation of each latent feature with its transformation is a *skip connection* that aids generalization [36], along with batch normalization that has been shown improvement against layer normalization for multivariate time-series analysis [21].

2) Multi-headed Self-attention Module: An MSA module is designed to assign selective importance to latent features extracted for each time point by the preceding layers of the encoder [32]. Particularly, MSA contains trainable parameters that capture the similarity between input features at different time points via their query, key, and value representations. Multiple attention heads enable adaptations to long-term dependencies and capture relevance between segments of multivariate data, without prior bias based on position [21].

Formally, at each time point t, the output representation is computed via a weighted sum over the value vectors $\boldsymbol{z}^{(i)}_{v,t'} \in \mathbb{R}^{D_v}$, $t' \in [1, \ldots, T]$, where the importance weight assigned to the value vector at time t' is computed as a dot-product similarity between its corresponding key vector $\boldsymbol{z}^{(i)}_{k,t'} \in \mathbb{R}^{D_q}$ and a query vector $\boldsymbol{z}^{(i)}_{q,t} \in \mathbb{R}^{D_q}$ at time t. As a result, given a latent feature $\boldsymbol{Z}^{(i)}$, a query $\boldsymbol{Z}^{(i)}_q = [\boldsymbol{z}^{(i)}_{q,1}; \ldots; \boldsymbol{z}^{(i)}_{q,T}] \in \mathbb{R}^{T \times D_q}$, a key $\boldsymbol{Z}^{(i)}_k = [\boldsymbol{z}^{(i)}_{k,1}; \ldots; \boldsymbol{z}^{(i)}_{k,T}] \in \mathbb{R}^{T \times D_q}$, and a value $\boldsymbol{Z}^{(i)}_v = [\boldsymbol{z}^{(i)}_{v,1}; \ldots; \boldsymbol{z}^{(i)}_{v,T}] \in \mathbb{R}^{T \times D_v}$ are computed by applying three different trainable affine transformations on $\boldsymbol{Z}^{(i)}$. The self-attention output for a single attention head (SA) is then computed via a scaled dot-product:

$$\text{SA}(\boldsymbol{Z}^{(i)}) = \text{softmax}\left(\frac{\boldsymbol{Z}^{(i)}_q \boldsymbol{Z}^{(i)\top}_k}{\sqrt{D_q}}\right) \boldsymbol{Z}^{(i)}_v, \tag{2}$$

where softmax converts the similarity scores to a probability distribution over the input sequence of length T. This operation is performed in parallel for each of the H attention heads (each with its own trainable transformations). The resulting outputs $\text{SA}_h \in \mathbb{R}^{T \times D_v}$, $h \in [1, \ldots, H]$ are first concatenated and finally aggregated into a single representation through a trainable linear transformation $\boldsymbol{W}_A \in \mathbb{R}^{HD_v \times D}$:

$$\text{MSA}(\boldsymbol{Z}^{(i)}) = [\text{SA}_1(\boldsymbol{Z}^{(i)})\ \text{SA}_2(\boldsymbol{Z}^{(i)}) \ldots \text{SA}_H(\boldsymbol{Z}^{(i)})]\boldsymbol{W}_A. \tag{3}$$

B. Reconstruction-Based Loss Function

We aim for the transformer model to extract discriminative latent features that govern the generation of EEG recordings, i.e., to model the input data distribution. To this end, we corrupt each input sample by a novel masking strategy that is uniquely designed for modeling multivariate time-series data such as EEG [21]. We train the transformer model via a loss function that minimizes the error between the original (unmasked) recording $\boldsymbol{X}^{(i)}$ and the corresponding reconstruction $\hat{\boldsymbol{X}}^{(i)}$.

Formally, a proportion $r \in (0, 1)$ of each channel $m \in \{1, \ldots, M\}$ in each EEG recording $\boldsymbol{X}^{(i)}$ is dynamically masked at the beginning of each training step by setting the encoder input values at chosen time points to 0. The values at each channel alternate between consecutive masked and unmasked sequences. The number of masked time points follows a geometric distribution with mean l_m, while the number of unmasked time points follows a geometric distribution with mean $l_u = \frac{1-r}{r} l_m$. This transition paradigm is also known as an M/M/1 queue, in which the number of customers in a system is geometrically distributed [37]. The resulting masking strategy encourages the transformer to attend on time points preceding and following the masked segments both in individual channels, as well as across the aligned time points in other channels to capture inter-channel dependencies, and has been found more

effective than other denoising strategies for downstream tasks, including Bernoulli masking (c.f. Table II & [21]).

Finally, the reconstruction loss for end-to-end training of our model is the mean-squared reconstruction error. Crucially, the loss is computer over only the set of *masked* time points $\mathcal{M} = \{(t,m) \,|\, \text{masked } X^{(i)}_{t,m}, t \in \{1,\dots,T\}, m \in \{1,\dots,M\}\}$:

$$\frac{1}{|\mathcal{M}|} \sum_{(t,m)\in\mathcal{M}} (X^{(i)}_{t,m} - \hat{X}^{(i)}_{t,m})^2. \tag{4}$$

C. Seizure Identification

We aim to employ the trained transformer to distinguish between EEG recordings that contain seizures and those which do not; this motivates us to pose unsupervised seizure identification as an anomaly detection problem. Thus, we train the transformer architecture on recordings that *do not* contain seizures. This allows for the learned latent features to capture non-seizure activity [19]. As the transformer is trained to model non-seizure activity, recordings with no seizures are expected to be reconstructed with low error in inference time. In contrast, EEG recordings including seizure activity come from a different distribution, and thus, the model naturally reconstructs such input recordings with a relatively larger error; we use this observation as an indicator for a seizure (c.f. Section IV-D).

We note that the exclusion of seizure recordings from the training set *does not* constitute supervision or require any special annotation, as the default states of patients and healthy individuals alike are non-seizure, whose recordings can be collected and kept separate from the recordings of seizure episodes (which we only use for evaluating our method). In real-life applications, EEG data with no seizure activity can be easily augmented with recordings from healthy individuals, which are trivially accessible compared to the ones from patients experiencing seizures.

IV. Experiments

A. Datasets

We evaluate our method on three publicly available EEG datasets collected at the: (i) Massachusetts Institute of Technology (MIT) and Boston Children's Hospital [38] (ii) University of Pennsylvania (UPenn) and Mayo Clinic [39], and (iii) Temple University Hospital of Philadelphia (TUH) [40].

The MIT dataset contains EEG recordings acquired on the scalp with 256 Hz sampling rate from a maximum of $M = 38$ channels. 198 seizure recordings were labeled w.r.t. their start and end times. The total duration of non-seizure recordings is $40,800$ seconds and seizure recordings is 2889 seconds.

The UPenn dataset contains 1-second long EEG recordings acquired intracranially at $500 - 5000$ Hz from a maximum of $M = 72$ channels. The total duration of non-seizure recordings is 7164 seconds and seizure recordings is 653 seconds.

The TUH dataset contains EEG recordings acquired on the scalp with 250 Hz sampling rate from a maximum of $M = 38$ channels. 1229 seizure recordings were labeled w.r.t. their start and end times. The total duration of non-seizure recordings is $49,922$ seconds and seizure recordings is 2600 seconds.

B. Preprocessing

EEG recordings are typically preprocessed to eliminate the powerline noise at 60 Hz [19]. We first unify the sampling rates in each dataset by downsampling to the smallest sampling rate across all recordings. Then, we filter the recordings via a 4-th order Butterworth bandpass filter with range 0.5-50 Hz.

To construct samples with the same size, we extract sliding windows over each recording, where each window contains T time points and overlaps with its consecutive window by 50%. We choose T based on the shortest seizure segment in each dataset. In doing so, $T = 1536$ for MIT, $T = 500$ for UPenn, and $T = 462$ for TUH. This process results in $13,600$ windows with non-seizure activity and 963 windows with seizure activity for MIT, $14,329$ windows with non-seizure activity and 1307 windows with seizure activity for UPenn, and $54,264$ windows with non-seizure activity and 2826 windows with seizure activity for TUH. In real-life applications, a minimum seizure window length can be decided by clinical experts, as in UPenn that directly provides 1 second-long seizure recordings.

Moreover, we aim to consistently form $T \times M$ size windows, while not disregarding any channels with potential seizure activity. Thus, to construct samples with the same number of M channels, we reuse data from other channels for the recordings that have missing data at certain channels, compared to the recording with the largest number of channels in each dataset. Again, in real-life applications, clinical experts can determine which channels to employ or discard for seizure identification. Finally, we normalize windows by subtracting the mean and dividing by the standard deviation across all windows to aid the convergence of training [34].

C. Experiment Setup and Competing Methods

We partition all windows containing non-seizure and seizure activity into training, validation, and test sets in a stratified manner, allocating 60% for training, 20% for validation, and the remaining 20% for testing. As baseline methods, we implement shallow and deep learning models for both supervised and unsupervised settings.

1) Unsupervised Learning Methods: For our method, we employ the transformer encoder architecture proposed by Vaswani et al. (2017), with the modifications of fully-trainable positional encoding, batch normalization and the same hyperparameters suggested by Zerveas et al. (2021). We train the autoencoder over *only* non-seizure training windows using the unsupervised loss given by Eq. (4). We monitor the loss value computed over the non-seizure windows in the validation set and use the model that attains the lowest validation loss.

Following the literature on shallow unsupervised methods [41], we reduce the dimension of all EEG windows in the test set to 3 using the t-Distributed Stochastic Neighbor Embedding (t-SNE) [42] algorithm, and apply K-means clustering [43] on the resulting windows with two clusters indicating non-seizure and seizure. Moreover, as an unsupervised deep learning baseline, we train a state-of-the-art convolutional VAE [20].

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	$\mathit{0.92 \pm 0.008}$	$\mathit{0.85 \pm 0.011}$	$\mathit{0.82 \pm 0.012}$	$\mathit{0.89 \pm 0.02}$
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

Table I: Seizure identification performance metrics and confidence intervals on UPenn, MIT and TUH. We compare our transformer-based unsupervised identification method (in bold) with unsupervised methods comprising VAE and t-SNE followed by K-means clustering, as well as supervised methods comprising XGBoost, ROCKET, and the same transformer architecture trained via supervised and pre-trained supervised learning. Best performance for each dataset are in italics.

2) Supervised Learning Methods: First, we employ the same transformer encoder architecture described in Section III-A and map the latent features learned from each window to a binary prediction. In doing so, we concatenate all latent features corresponding to all time points of each window into a single vector and apply a fully-connected layer comprising a scalar output with sigmoid activation. We train the resulting architecture via cross-entropy loss over all training windows, employing the same hyperparameters found optimal by Zerveas et al. (2021). To combat overfitting due to class imbalance in supervised learning, we oversample and augment the seizure windows in training via random reversing and drifting. We monitor the F1-score computed over the validation set and use the model that attains the best validation score.

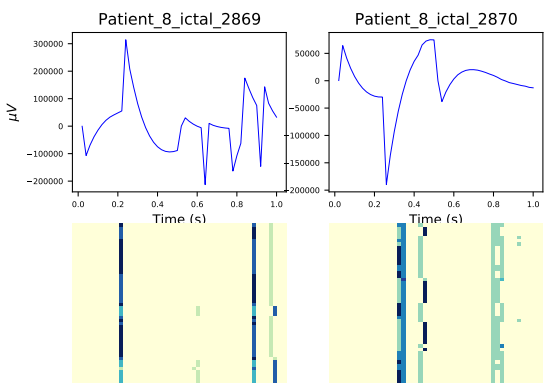
Moreover, we train state-of-the-art shallow models XGBoost [44] and ROCKET [45] over the supervised training set. XGBoost is a decision-tree classifier using gradient boosting for ensembling. ROCKET transforms time-series using 500 random convolutional kernels and uses the extracted features to train a ridge regression classifier. Ridge regression hyperparameter is varied in $[10^{-3}, 10^{3}]$ and best hyperparameter is determined w.r.t. the accuracy over the validation set.

3) Pre-trained Supervised Learning: Finally, we combine the transformer-based seizure identification methods via unsupervised pre-training and supervised fine-tuning [21]. Following the unsupervised approach described in Section IV-C1, we first pre-train the transformer encoder over non-seizure training windows. Having initialized its weights accordingly, we then fine-tune the model via both non-seizure and seizure training windows, using the same setup described in Section IV-C2.

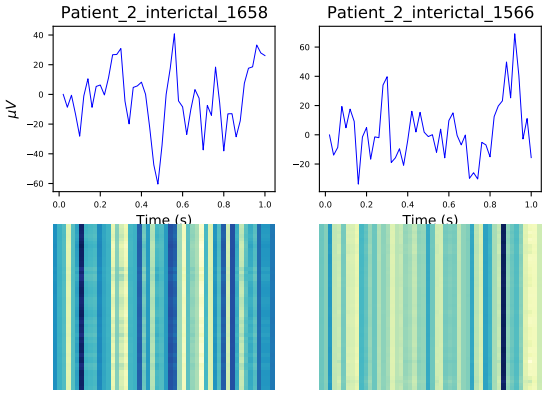
D. Evaluation Metrics

To evaluate the seizure identification performance of our approach, as well as the VAE baseline, we use the mean absolute error over the time points and electrode channels in each EEG window from the test set as the corresponding seizure prediction score. For all supervised competing methods, we use the traditional prediction score for inference.

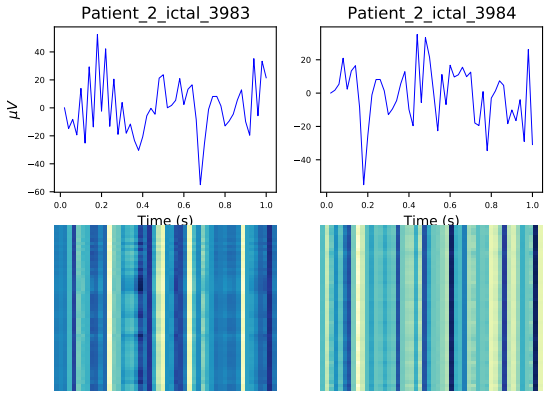
For all competing methods described in Section IV-C, we report AUC for distinguishing seizure vs. non-seizure windows in the test set. To compute binary decision metrics, we threshold the prediction score of each window at the value for which the geometric mean of recall and true negative rate is maximal [46].



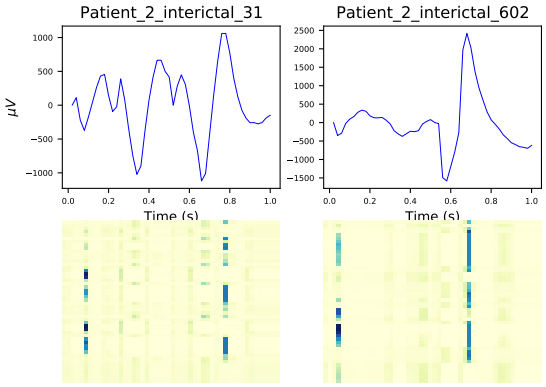
(a) Correctly Identified Seizure Windows (True Positive)



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



(c) Falsely Identified Seizure Windows (False Negative)



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

Figure 2: Example EEG windows, corresponding seizure identifications and self-attention weights on UPenn. First and third rows contain example windows of true positive, true negative, false negative, and false positive identifications, respectively. Second and fourth rows contain the corresponding self-attention weight heatmaps computed by the transformer architecture, where darker colors indicate higher importance. For each window, we visualize the channel with the largest reconstruction error.

Using the respective threshold, we calculate class-weighted precision and recall, as well as balanced accuracy for binary identification of seizure vs. non-seizure windows in the test set, considering the imbalanced distribution between the two. In real-life applications, decision thresholds may be determined by clinical experts with respect to the desired trade-off between false positives and negatives [47].

We report all metrics along with the 95% confidence intervals, which are computed as $1.96 \times \sigma_A$, where σ_A^2 is the variance for metric A. Variance for AUC is computed by:

$$\sigma_A^2 = \frac{1}{mn}\big(A(1-A)+(m-1)(P_x-A^2)+(n-1)(P_y-A^2)\big), \quad (5)$$

where $P_x = A/(2-A)$, $P_y = 2A^2/(1+A)$, and m, n are the number of seizure and non-seizure windows, respectively [48]. Variance for other metrics are computed by:

$$\sigma_A^2 = A(1-A)/(m+n). \quad (6)$$

E. Results and Discussion

1) Seizure Identification Performance: Table I shows the seizure identification performance of our transformer-based unsupervised method vs. supervised and pre-trained supervised transformers, XGBoost, ROCKET, VAE, and t-SNE followed by K-means clustering over all datasets. Our novel transformer-based anomaly detection method establishes a dramatic improvement among all unsupervised methods, by successfully distinguishing between non-seizure vs. seizure windows with up to *0.94 AUC* and outperforming its state-of-the-art deep learning counterpart VAE by up to 33% AUC on MIT. Clustering on raw EEG windows cannot capture the complex evolution of EEG and predicts all windows as non-seizure. These observations demonstrate the benefit of the transformer architecture for unsupervised anomaly detection in our setting.

Crucially, despite the lack of seizure labels during training, our unsupervised anomaly detection approach leads to significantly *better seizure identification than all purely supervised learning baselines and the pre-trained transformer fine-tuned with 50% of the training labels* over UPenn and MIT, by up to 16% recall, 9% accuracy, and 9% AUC. Moreover, unlike supervised learning, class imbalance strongly biases supervised models towards non-seizure predictions and hinders generalization over the distribution of held-out test samples.

As a result, unsupervised anomaly detection via transformers establishes a consistently better balance between precision and recall than supervised learning and further demonstrates its benefit in learning from imbalanced datasets such as ours.

The TUH dataset is particularly challenging by being a compilation of several EEG databases collected over years from patients with vast variations in demographic and medical backgrounds [40], compared to self-contained UPenn and MIT datasets collected from only 8 and 24 patients, respectively. In this case, our unsupervised transformer still fares significantly better than the purely supervised transformer, while unsupervised VAE outperforms *all* supervised learning baselines, including the pre-trained transformer. These observations further motivate unsupervised learning for our task.

As expected, the computationally expensive transformer model, which has first undergone unsupervised pre-training and then supervised fine-tuning with *all* training labels, outperforms both purely supervised as well as purely unsupervised transformer models (the latter by a smaller margin). However, our unsupervised anomaly detection method *does not* require ground-truth seizure labels during training as a crucial advantage, while still leading to successful seizure identification.

2) Seizure Identification Examples: We visualize example EEG windows from UPenn and the corresponding seizure identifications of the unsupervised transformer in the first and third rows of Figure 2, selecting the channel with the largest mean reconstruction error for each window. Agreeing with the clinical description of seizures, true seizure windows in Figure 2a contain high-frequency waves with large amplitudes [2]. Meanwhile, true non-seizure windows in Fig. 2b attain significantly less amplitude changes and spikes compared to true positive windows. Note that the seizure patterns cannot be identified w.r.t. only large amplitude or high frequency, motivating a more sophisticated approach such as ours. For instance, non-seizure windows in Fig. 2d have a larger amplitude range than the seizure windows in Figure 2c, while the seizure windows in Fig. 2c contain similar spikes to the non-seizure windows in Figure 2b w.r.t. amplitude and frequency.

3) Benefit of Self-Attention: We visualize the self-attention weights computed by the last encoder layer of the unsupervised transformer on example EEG windows from UPenn as 2D heatmaps in the second and fourth rows of Figure 2. For each time point along the horizontal axis of each heatmap, self-attention weights (c.f. Equation (3)) from other time points are indicated along the vertical axis. Darker heatmap colors correspond to larger weights and, thus, higher importance.

It appears that the transformer model within our unsupervised identification method can successfully learn to pay more attention to seizure patterns including high-frequency spikes and waves evolving with large amplitudes [2]. Moreover, when the model predicts the existence of seizures, it shows patterns of focused attention, containing only few time points with large weights (Figures 2a and 2d), while windows identified as non-seizure (Figures 2b and 2c) lead to much more evenly distributed attention. These observations indicate that employing a transformer architecture with self-attention can improve both performance, as well as explainability of seizure identification decisions, by underlining, e.g., spike-wave discharges that are indicative of seizures [2].

Dataset	Method	Precision	Recall	Accuracy	AUC
MIT	Geometric (Ours)	**0.98**	**0.9**	**0.87**	**0.94**
	Bernoulli	0.98	0.85	0.85	0.9
UPenn	Geometric (Ours)	**0.88**	**0.76**	**0.68**	**0.73**
	Bernoulli	0.86	0.72	0.65	0.72
TUH	Geometric (Ours)	0.92	**0.57**	**0.61**	**0.57**
	Bernoulli	**0.93**	0.4	0.59	0.54

Table II: Effect of masking strategy on seizure identification.

4) Effect of Masking Strategy: Table II shows the seizure identification performance of training with our geometric masking strategy against masking each time point independently at random with a Bernoulli distribution. Our approach of unsupervised training with geometric masking consistently leads to better performance than Bernoulli masking, demonstrating its benefit in modeling multivariate data such as EEG.

V. Conclusion

We propose a fully-unsupervised transformer-based method for seizure identification on raw EEG. Our approach involves training an autoencoder involving a transformer encoder to reconstruct stochastically-masked EEG recordings of non-seizure activity, and thus, modeling a non-seizure data distribution without any ground-truth labels. Since EEG recordings of seizures belong to a different distribution, they are identified based on the higher reconstruction errors attained at inference time. Our method can successfully distinguish between non-seizure and seizure windows and can even achieve significantly better seizure identification performance than state-of-the-art supervised time-series methods, including its purely supervised transformer-based counterpart. Generalizing our method to other applications involving anomalous activity detection on multivariate time-series data is a promising future direction.

Our unsupervised approach can significantly alleviate the burden on clinical experts regarding laborious and difficult EEG inspections to provide labels indicating segments that contain seizures. Furthermore, if automated identification performance meets clinical requirements, our method can aid availability of seizure diagnoses for the wider public, especially in areas where access to well-trained healthcare professionals is limited.

References

[1] R. D. Thijs, R. Surges, T. J. O'Brien, and J. W. Sander, "Epilepsy in adults," *The Lancet*, vol. 393, no. 10172, pp. 689–701, 2019.

[2] P. M. Vespa, V. Shrestha, N. Abend, D. Agoston, A. Au, M. J. Bell, T. P. Bleck, M. B. Blanco, J. Claassen, R. Diaz-Arrastia, *et al.*, "The epilepsy bioinformatics study for anti-epileptogenic therapy (EpiBioS4Rx) clinical biomarker: study design and protocol," *Neurobiology of Disease*, vol. 123, pp. 110–114, 2019.

[3] R. J. Staba, M. Stead, and G. A. Worrell, "Electrophysiological biomarkers of epilepsy," *Neurotherapeutics*, vol. 11, no. 2, 2014.

[4] Z. Deng, P. Xu, L. Xie, K.-S. Choi, and S. Wang, "Transductive joint-knowledge-transfer TSK-FS for recognition of epileptic EEG signals," *Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 8, pp. 1481–1494, 2018.

[5] S. Saba-Sadiya, E. Chantland, T. Alhanai, T. Liu, and M. M. Ghassemi, "Unsupervised EEG artifact detection and correction," *Frontiers in Digital Health*, vol. 2, p. 57, 2021.

[6] T. Zhang and W. Chen, "LMD based features for the automatic seizure detection of EEG signals using SVM," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 8, 2016.

[7] H. U. Amin, M. Z. Yusoff, and R. F. Ahmad, "A novel approach based on wavelet analysis and arithmetic coding for automated detection and diagnosis of epileptic seizure in EEG signals using machine learning techniques," *Biomedical Signal Processing and Control*, vol. 56, 2020.

[8] R. Ramos-Aguilar, J. A. Olvera-López, I. Olmos-Pineda, and S. Sánchez-Urrieta, "Feature extraction from EEG spectrograms for epileptic seizure detection," *Pattern Recognition Letters*, vol. 133, pp. 202–209, 2020.

[9] B. Zhu and M. Shoaran, "Unsupervised domain adaptation for cross-subject few-shot neurological symptom detection," in *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2021.

[10] V. K. Mehla, A. Singhal, P. Singh, and R. B. Pachori, "An efficient method for identification of epileptic seizures from EEG signals using fourier analysis," *Physical and Engineering Sciences in Medicine*, 2021.

[11] X. Zhang, L. Yao, M. Dong, Z. Liu, Y. Zhang, and Y. Li, "Adversarial representation learning for robust patient-independent epileptic seizure detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2852–2859, 2020.

[12] Y. Li, Y. Liu, W.-G. Cui, Y.-Z. Guo, H. Huang, and Z.-Y. Hu, "Epileptic seizure detection in EEG signals using a unified temporal-spectral squeeze-and-excitation network," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 4, pp. 782–794, 2020.

[13] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan, "Time-series representation learning via temporal and contextual contrasting," *arXiv preprint arXiv:2106.14112*, 2021.

[14] N. Islah, J. Koerner, R. Genov, T. A. Valiante, and G. O'Leary, "Machine learning with imbalanced EEG datasets using outlier-based sampling," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*. IEEE, 2020, pp. 112–115.

[15] S. Chakrabarti, A. Swetapadma, P. K. Pattnaik, and T. Samajdar, "Pediatric seizure prediction from EEG signals based on unsupervised learning techniques using various distance measures," in *2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech)*, 2017.

[16] S. Belhadj, A. Attia, A. B. Adnane, Z. Ahmed-Foitih, and A. A. Taleb, "Whole brain epileptic seizure detection using unsupervised classification," in *2016 8th International Conference on Modelling, Identification and Control (ICMIC)*. IEEE, 2016, pp. 977–982.

[17] J. Birjandtalab, M. B. Pouyan, and M. Nourani, "Unsupervised EEG analysis for automated epileptic seizure detection," in *First International Workshop on Pattern Recognition*, vol. 10011. International Society for Optics and Photonics, 2016, p. 100110M.

[18] K. Charupanit, I. Sen-Gupta, J. J. Lin, and B. A. Lopour, "Detection of anomalous high-frequency events in human intracranial EEG," *Epilepsia Open*, vol. 5, no. 2, pp. 263–273, 2020.

[19] S. You, B. H. Cho, S. Yook, J. Y. Kim, Y.-M. Shon, D.-W. Seo, and I. Y. Kim, "Unsupervised automatic seizure detection for focal-onset seizures recorded with behind-the-ear EEG using an anomaly-detecting generative adversarial network," *Computer Methods and Programs in Biomedicine*, vol. 193, p. 105472, 2020.

[20] Yıldız, R. Garner, M. Lai, and D. Duncan, "Unsupervised seizure identification on EEG," *Computer Methods and Programs in Biomedicine*, vol. 215, p. 106604, 2022.

[21] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2114–2124.

[22] P. Boonyakitanont, A. Lek-Uthai, K. Chomtho, and J. Songsiri, "A review of feature extraction and performance evaluation in epileptic seizure detection using EEG," *Biomedical Signal Processing and Control*, vol. 57, p. 101702, 2020.

[23] M. Radman, M. Moradi, A. Chaibakhsh, M. Kordestani, and M. Saif, "Multi-feature fusion approach for epileptic seizure detection from EEG signals," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3533–3543, 2020.

[24] W. Zhao, W. Zhao, W. Wang, X. Jiang, X. Zhang, Y. Peng, B. Zhang, and G. Zhang, "A novel deep neural network for robust detection of seizures using EEG signals," *Computational and Mathematical Methods in Medicine*, vol. 2020, 2020.

[25] S. Chakrabarti, A. Swetapadma, and P. K. Pattnaik, "A channel independent generalized seizure detection method for pediatric epileptic seizures," *Computer Methods and Programs in Biomedicine*, vol. 209, 2021.

[26] A. M. Abdelhameed and M. Bayoumi, "Semi-supervised EEG signals classification system for epileptic seizure detection," *Signal Processing Letters*, vol. 26, no. 12, pp. 1922–1926, 2019.

[27] L. Sun, B. Jin, H. Yang, J. Tong, C. Liu, and H. Xiong, "Unsupervised EEG feature extraction based on echo state network," *Information Sciences*, vol. 475, pp. 1–17, 2019.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[30] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," *Frontiers in Human Neuroscience*, vol. 15, p. 253, 2021.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, 2014.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*. PMLR, 2015, pp. 448–456.

[35] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," *CoRR*, vol. abs/1606.08415, 2016. [Online]. Available: http://arxiv.org/abs/1606.08415

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[37] R. G. Gallager, *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2013.

[38] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[39] UPenn and MayoClinic, "Upenn and Mayo Clinic's seizure detection challenge," 2014. [Online]. Available: https://www.kaggle.com/c/seizure-detection/

[40] I. Obeid and J. Picone, "The temple university hospital EEG data corpus," *Frontiers in Neuroscience*, vol. 10, p. 196, 2016.

[41] S. Chakrabarti, A. Swetapadma, P. K. Pattnaik, and T. Samajdar, "Pediatric seizure prediction from EEG signals based on unsupervised learning techniques using various distance measures," in *2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology*, 2017, pp. 1–5.

[42] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[43] C. M. Bishop, "Pattern recognition," *Machine learning*, vol. 128, no. 9, 2006.

[44] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[45] A. Dempster, F. Petitjean, and G. I. Webb, "ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, 2020.

[46] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[47] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *CoRR*, vol. abs/1901.03407, 2019. [Online]. Available: http://arxiv.org/abs/1901.03407

[48] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.