M2D2: Maximum-Mean-Discrepancy Decoder for Temporal Localization of Epileptic Brain Activities

Alireza Amirshahi, Anthony Thomas, Amir Aminifar, *Senior Member, IEEE*, Tajana Rosing, *Fellow, IEEE*, and David Atienza, *Fellow, IEEE*

***Abstract*—Recent years have seen growing interest in leveraging deep learning models for monitoring epilepsy patients based on electroencephalographic (EEG) signals. However, these approaches often exhibit poor generalization when applied outside of the setting in which training data was collected. Furthermore, manual labeling of EEG signals is a time-consuming process requiring expert analysis, making fine-tuning patient-specific models to new settings a costly proposition. In this work, we propose the Maximum-Mean-Discrepancy Decoder (M2D2) for automatic temporal localization and labeling of seizures in long EEG recordings to assist medical experts. We show that M2D2 achieves 76.0% and 70.4% of F1-score for temporal localization when evaluated on EEG data gathered in a different clinical setting than the training data. The results demonstrate that M2D2 yields substantially higher generalization performance than other state-of-the-art deep learning-based approaches.**

***Index Terms*—Maximum mean discrepancy, temporal localization, epileptic seizure, non-invasive EEG.**

I. Introduction

EPILEPSY is a chronic neurological disorder characterized by persistent seizures and affects over 70 million people worldwide [1]. The root causes of epilepsy and broadly effective treatments remain the subject of ongoing investigations. Gathering data on the frequency and duration of seizures is an important component of this research and informs both clinical diagnosis on an individual level and a broader understanding of the condition as a whole. In particular, epileptic seizures are known to be associated with particular patterns in an electroencephalogram (EEG). Neurologists can inspect EEG recordings to determine the timing and frequency of seizures to develop a detailed understanding of this condition, in line with the recent trends in precision medicine. However, this process is time-consuming for medical professionals and requires hospital stays by patients.

In recent years, deep learning (DL) models have emerged as a state-of-the-art technique thanks to their ability to automatically learn useful features for discriminating seizures from regular brain activity. These models are typically trained on a large database of EEG signals collected from epileptic patients in a clinical setting, and hand-labeled by experts. One typically wishes that such models are useful beyond the immediate setting in which they were trained. That is, a model trained on one set of patients should continue to deliver high accuracy when applied to data gathered from a different set of patients in a different setting.

The most basic approach to satisfy this goal is to use deep learning methods in which one simply applies a pre-trained model to a new patient [2], [3]. However, the precise manifestation of seizures in EEG signals varies on a person-to-person basis, and existing deep learning approaches generally need to fine-tune models to target a new set of patients [4], [5], [6], [7]. Because these approaches assume access to at least some labelled EEG data for each new patient, they can typically achieve high-accuracy. However, this necessitates acquiring new labeled data for every new patient, which, in turn, requires a costly process of collecting and manually annotating a large volume of EEG data.

Our goal in this work is to reduce the burden of this process. We propose a new deep learning-based technique for approximate temporal localization of seizures in long EEG recordings. Our approach takes as input a long EEG signal, and returns a time stamp t such that a seizure occurred within $t \pm \Delta$ minutes. Thus, the expert only needs to search an interval of 2Δ minutes, instead of the entire signal. The parameter Δ controls the tradeoff between the volume of data to be annotated, and the fraction

Manuscript received 1 December 2021; revised 13 August 2022; accepted 13 September 2022. Date of publication 22 September 2022; date of current version 5 January 2023. This work supported in part by the ML-Edge Swiss National Science Foundation (NSF) Research Project under Grant 200020182009/1, in part by the PEDESITE Swiss NSF Sinergia Project under Grant SCRSII5 193813/1, in part by the RESoRT Fondation Botnar Project under Grant REG-19-019, in part by the WASP Program of the Knut and Alice Wallenberg Foundation, in part by CRISP one of six centers in JUMP, in part by DARPA through SRC Program, and in part by NSF under Grants 2003279, 1911095, 1826967, 2100237, 2112167, GRC TASK 3021.001, and GRC TASK 2942.001. *(Corresponding author: Alireza Amirshahi.)*

Alireza Amirshahi and David Atienza are with the Embedded Systems Laboratory (ESL), Institute of Electrical and Micro Engineering, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland (e-mail: alireza.amirshahi@epfl.ch; david.atienza@epfl.ch).

Anthony Thomas and Tajana Rosing are with the Department of Computer Science and Engineering, University of California, San Diego, CA 92093 USA (e-mail: ahthomas@eng.ucsd.edu; tajana@ucsd.edu).

Amir Aminifar is with the Department of Electrical and Information Technology, Lund University, SE-221 00 Lund, Sweden (e-mail: amir.aminifar@eit.lth.se).

Digital Object Identifier 10.1109/JBHI.2022.3208780

of seizures which are identified. Our primary contribution is to show that our approach is able to localize seizures more precisely than existing work when evaluated on a completely new data set.

Our approach is called the "Maximum-Mean-Discrepancy Decoder" (M2D2). At a high level, one may think of a deep learning model as consisting of an encoder, which yields a derived representation of an input signal, and a decoder, which infers a class label from the derived representation. Intuitively, the representations derived by the encoder should have high intra-class similarity but low inter-class similarity. That is, the regions of the signal corresponding to seizures should all map to "similar" representations, and moreover, these representations should be "different" from those of non-seizure regions. We observe that this intuition is made precise by the notion of "maximum-mean-discrepancy" (MMD) from statistics. The MMD defines a general notion of similarity between samples from two probability distributions and, intuitively, works by measuring the similarity of points within and between each sample - just as we seek to do here.

Building on this observation, we train the decoder portion of our network to localize seizures based only on the empirical MMD between a candidate seizure region and the rest of the signal. We hypothesize that, by only giving the decoder access to the MMD, the encoder will produce representations that tend to have high intra-class and low inter-class similarity, and that, as a consequence, will exhibit better generalization than conventional architectures in which the decoder can directly access much more information about the input signals. To the best of our knowledge, we are the first to consider the use of MMD as a layer *within* a supervised deep neural network. We show that M2D2 leads to improved generalization performance, compared to the state-of-the-art techniques, when evaluating our model on a dataset collected in an entirely different clinical setting.

Furthermore, seizures vary widely in length from only a few seconds to over several minutes [8]. Prior work has fixed the length of the candidate seizure region at the average length of a seizure [9]; however, this may miss short or long seizures. In the proposed work, thanks to the M2D2 architecture, we are able to use a range of possible values for the candidate region length to address this issue.

The contributions of our work are summarized as follows:

- To the best of our knowledge, we are the first to evaluate the temporal seizure localization on a dataset different from the training dataset. This setting is more reflective of the real-world scenario where the models are applied beyond the immediate clinical setting in which they are trained.
- We use MMD computation as a layer implemented within a deep neural network. This layer enables the model to learn features based on not only the current input but also the distribution of the adjacent windows and the entire signal.
- In this work, the candidate seizure region is not fixed at a single length. Instead, a range of possible sizes for the candidate region of seizure is considered, and the network is trained to choose the best length.

The rest of this article is organized as follows. In Section II, we review the background in EEG signal analysis, MMD and Variational Information Bottleneck. Furthermore, the related works in seizure temporal localization are investigated. In Section III, we describe our proposed model, the M2D2 framework and the details of the architecture. Also, the training and back-propagation process of M2D2 is studied. In Section IV, the experimental setup is discussed, and then, in Section V the results are shown. Next, in Section VI, we discussed the results in different points of view. Finally, in Section VIII, we summarize the main conclusions of this work.

II. Background and Related Work

In the following section, we provide the necessary technical background on EEG analysis and the statistical techniques used in M2D2.

A. EEG Analysis

We here provide a brief overview of electroencephalography as it pertains to our work [10]. EEG analysis records a time-series of electrical impulses generated by the brain. The particular spatiotemporal patterns of these impulses are generally held to be related to brain activity at a particular moment in time. For instance, specific waveforms in the EEG can be associated with everyday activities like blinking or chewing. Similarly, certain atypical neurological conditions, e.g., the seizures associated with epilepsy, manifest in EEG recordings making their analysis an important diagnostic tool. The waveforms in an EEG are generated by measuring the voltage difference between pairs of electrodes distributed over the scalp. The readings produced by each such pair are called a channel. Thus, an EEG contains a spatial and temporal component, both of which are typically relevant for analysis.

EEG recordings typically contain a multitude of artifacts which present a significant complication for analysis. Artifacts may arise from natural causes–common examples being muscular activities like chewing or blinking and changes in conductance from sweat–or non-natural causes–a common example being jostled or disconnected electrodes. Artifact removal is an essential component of EEG analysis and is typically performed as a pre-processing step [11], [12]. Furthermore, while seizures (or ictal EEG) have some common trends, their precise manifestation is different across patients [13], making reliable decision-making in the presence of such artifacts and heterogeneity challenges.

B. Maximum-Mean-Discrepancy

The maximum mean discrepancy is a metric on the space of probability distributions [14]. Intuitively, the MMD works by representing a pair of distribution as points in a high-dimensional feature space and then measuring the distance be the two representative points. More formally, let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a real, continuous, positive-definite kernel function with an associated

reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. Let p be a probability measure supported on $\mathcal{X}$. For our purposes, we may assume $\mathcal{X}$ is a Euclidean space. The kernel mean embedding (KME) of p is defined as $\mu_p = \int_{\mathcal{X}} k(\cdot, x)\, dp(x)$ [15]. Given a pair of measures p and q supported on $\mathcal{X}$, the MMD is simply the distance between their respective KMEs:

$$\text{MMD}^2(p, q) = \|\mu_p - \mu_q\|^2_{\mathcal{H}}.$$

Given samples $\mathcal{P} = \{x_1, \ldots, x_n\}$ and $\mathcal{Q} = \{y_1, \ldots, y_m\}$ drawn i.i.d. from p and q respectively, the MMD can be estimated empirically as [14]:

$$\widehat{\text{MMD}}^2(p, q) = \frac{1}{n^2} \sum_{x,x' \in \mathcal{P}} k(x, x') + \frac{1}{m^2} \sum_{y,y' \in \mathcal{Q}} k(y, y') - \frac{2}{nm} \sum_{x \in \mathcal{P}, y \in \mathcal{Q}} k(x, y). \quad (1)$$

Intuitively, when the sampled points have a high intra-distribution similarity (measured by the first two terms) and a low inter-distribution similarity (measured by the third term), the MMD will be large. Throughout the remainder of the work, we will work with the squared-MMD, which suffices for our purposes.

C. Variational Information Bottleneck

In principle, the MMD can be applied directly to the raw signal values. However, in practice, performance is often improved by obtaining a lower-dimensional representation of the signal that compresses away uninformative short-term fluctuations. To do so, we here leverage a technique from Information Theory known as "Information Bottleneck" (IB). Given a pair of correlated random variables X and Y, the IB problem is to obtain a compressed representation Z of X that contains the minimal amount of information needed to predict Y [16]. Assuming X and Y are described by a distribution $p(X, Y)$, the IB problem can be formalized as solving:

$$p^*(Z|X) = \underset{p(Z|X)}{\text{argmax}}\, I(Z; Y) \text{ s.t. } I(Z; X) \leq \gamma,$$

where $I(A; B)$ is the mutual information between a pair of random variables A and B. Sampling from $p^*(Z|X)$ can be seen as an encoding process which takes an input $x \in \mathcal{X}$ and maps it to a codeword $z \in \mathcal{Z}$. The objective $I(Z; Y)$ ensures the codewords are informative about the outcome of interest Y, and the constraint $I(Z; X) \leq \gamma$ restricts the information the codewords convey about the original signal. Given the encoder $p(Z|X)$, a corresponding "decoder" distribution $p(Y|Z)$, can be computed analytically.

In practice, one typically has access to a set of samples $\{(x_i, y_i)\}_{i=1}^n$ drawn i.i.d. from $p(X, Y)$ which is unknown. The problem is intractable in this case, and so a common approach is to instead assume a parametric form for the encoder $q_\phi(Z|X)$, and decoder $q_\theta(Y|Z)$, and to then minimize a "variational" upper bound [17], [18]:

$$\hat{\mathcal{L}}_{IB}(\theta, \phi) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_z[-\log q_\theta(y_j \mid z_j)] + \beta D_{KL}(q_\phi(Z|x_j) \,||\, p(Z)),$$

where $D_{\text{KL}}(A \,\|\, B)$ is the KL-divergence between A and B. In practice, the encoder and decoder distributions are typically parameterized using neural networks [18], [19], [20]. From a practical perspective, the VIB is useful, because the learned representations enjoy robustness to certain types of signal artifacts [17], [20] which may improve the resilience of seizure detection algorithms [21].

D. Related Work

Algorithmic approaches for detecting and localizing seizures in EEG recordings have been extensively studied in the literature. Earlier work focused on methods for extracting hand-crafted features from EEG signals which are then used as input to learning algorithms like logistic regression models and support-vector-machines [13], [22], [23], [24], [25], [26].

More recently, there has been an increasing trend toward deep learning-based methods which obviate the need for feature extraction and typically lead to higher accuracy using convolutional neural networks (CNN), and EEG signals [27], [28], [29]. In [30] Long Short-Term Memory (LSTM) modules are used with the CNN to improve the seizure detection accuracy. In [31] a self-learning method is used to pre-train a Graph Neural Network (GNN) for the seizure detection and seizure type classification task. It is shown that by using the pre-training, the seizure detection performance can increase F1-score by 4.3%. In [32], we use the knowledge distillation technique to detect seizures using only ECG signals, while the teacher model uses multi-modal ECG and EEG signals. Using high-accurate individual ECG signal alleviate the signal acquisition in real-life scenarios. Of particular note here is our prior work in [21] used CNN with the VIB to detect seizures from EEG recordings. However, they use a simple decoder architecture that does not incorporate the MMD as we do here.

There has also been prior interest in using MMD or similar techniques to localize seizures. The work in [9] uses a similar approach that imputes the location of a seizure by finding a window of samples that maximizes the sum-of-squared Euclidean distances between samples in the window and the remainder of the signal. This is similar to computing the MMD with a linear kernel, and is subsumed by the more general kernel based MMD. Our approach is loosely motivated by the work in [33], [34]. This work assumes that a seizure represents a change in the behavior of an unknown underlying (time-dependent) density describing an EEG signal, and uses the MMD to obtain the "change point" that partitions the signal into two maximally different distributions. Unlike our approach, this work is entirely unsupervised and considers only simple, hand-crafted features of the signal when computing the MMD.

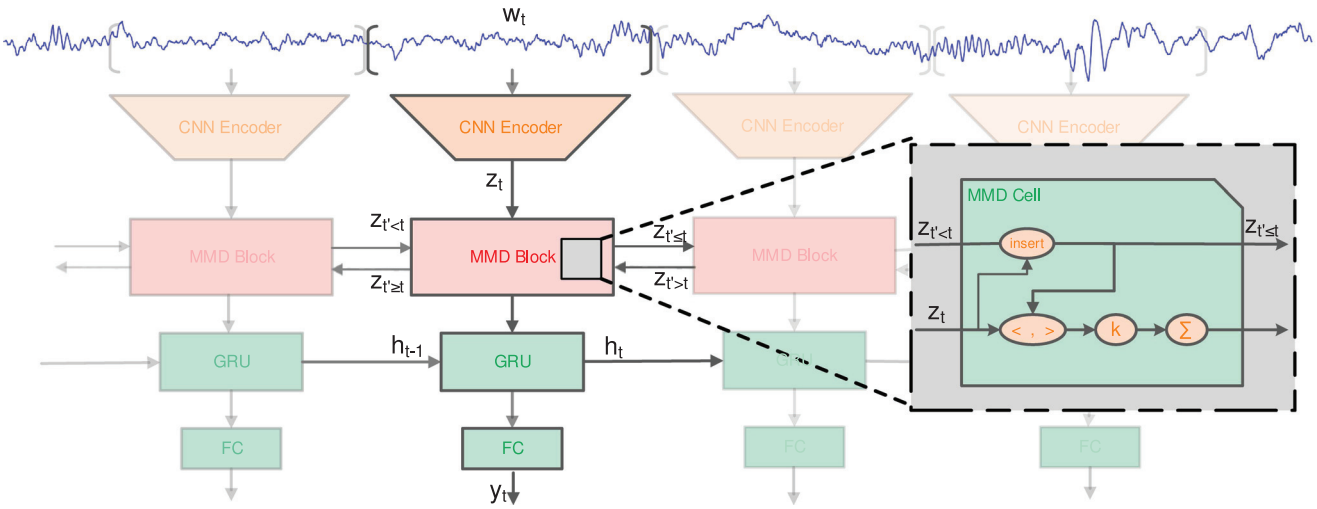


Fig. 1. Unrolled representation of the M2D2 architecture. The structure has a CNN as an encoder to extract codewords from the input signal. The decoder consists of the MMD layer, GRU, and a fully connected layer to produce the output for every signal window.

Moreover, these works do not present any systematic evaluation of their methods on a broad sample of EEG data. An important contribution of our work is to undertake the first extensive empirical evaluation of MMD in seizure temporal localization problems.

III. Methodology

A. Problem Formulation

Let $\{x_1, \ldots, x_T\} : x_i \in \mathcal{X} \subseteq \mathbb{R}^n$ be the raw samples from a n-channel EEG recording, where $t \in [T]$ indexes time. In our setting, the total length of the recordings is around one hour and the x_i are sampled at a rate of 256 Hz, although neither of these parameters are of particular importance. We partition the recording into a set of non-overlapping windows each consisting of r samples, which we denote by $\{\mathbf{w}_1, \ldots, \mathbf{w}_L\} : \mathbf{w}_i \in \mathcal{X}^{n \times r}$. We take $r = 1024$ corresponding to a length of 4 seconds. Let $[\mathbf{w}_i, \mathbf{w}_j]$, where $j \geq i$, be the interval of windows corresponding to a seizure event. Our goal is to identify any $i^* \in [i, j]$.

B. M2D2 Framework

The inputs to our model are the $\mathbf{w}_i$ formed by grouping together a set of r contiguous readings of the raw signal. These inputs are then encoded to a lower-dimensional representation $\mathbf{z} \in \mathbb{R}^d (d \ll nr)$ using the VIB method described above. We define q_ϕ to be a multivariate Gaussian whose mean and covariance matrix are parameterized using one-dimensional convolutional neural networks. That is, $q_\phi(Z|X = x) = \mathcal{N}(\mu(x), \Sigma(x))$, where $\mu(x)$ and $\Sigma(x)$ are computed using a 1D CNN. The output of our model is a scalar value $\hat{y} \in [0, 1]$ which is the probability that a particular $\mathbf{w}$ contains a seizure. The entire architecture is trained end to end to minimize the binary cross-entropy between the fitted values $\hat{y}$ and the ground truth.

Our primary novelty comes in our definition of the decoder. Let $\mathcal{Z} = \{z_1, \ldots, z_L\}$ be the compressed representations for each of the input windows in our training data. Our approach groups together a set of m adjacent z_i into a candidate ictal (seizure) region, which we denote $\mathcal{P}$. We then compute the empirical maximum mean discrepancy between $\mathcal{P}$ and the remainder of the signal $\mathcal{Q}$. The resulting vector of distances is used as a set of input features for the decoder. More formally, let define $\delta_t = \widehat{\text{MMD}}^2(\mathcal{P}, \mathcal{Q})$. The decoder can then be described as a function $f_\theta(\delta_1, \delta_2, \ldots, \delta_L)$ that returns a value $\hat{y} \in [0, 1]$ corresponding to the probability that $\mathbf{w}_t$ contains a seizure. Thus, the decoder has access only to the MMD between each candidate region and the remainder of the signal. Intuitively, the MMD output quantifies how different the distribution of the samples in $\mathcal{P}$ is from the distribution of the remaining samples. We describe the architecture of our decoder in more detail in the following section.

C. M2D2 Architecture

The architecture of M2D2 is summarized in Fig. 1. As described above, we first encode $\mathbf{w}_t$ to a codeword $\mathbf{z}_t$ which is generated by sampling from $q_\phi(Z|x)$. A complication arises because computing δ_t requires knowledge of all the $\mathbf{z}_i$–some of which occur in the future. To address this issue, we compute δ_t using $\mathbf{z}_i$ in two passes. In the first "forward" pass, we have access to all $\mathbf{z}_{t'}$ for $t' \leq t$. In the second "backward" pass, we have the analogous quantities for $t' > t$. Any particular δ_t can then be easily obtained in a streaming fashion by computing kernel evaluations between the $\mathbf{z}_i$ and summing up these values. Thereby we avoid the need to store all the individual kernel evaluations.

Fig. 1 presents the MMD cell. As an input, the cell takes $\mathbf{z}_t$ in each time step. The cell has also an *state* to remember $\mathbf{z}$ values over time. The memory unit is updated in every time step by inserting a copy of $\mathbf{z}_t$ inside *state* to prepare $z_{t' \leq t}$ ($z_{t' \geq t}$ during the backward pass) for the next time step. Given these stored $\mathbf{z}_i$ values, the MMD cell can compute any particular δ_t.

As mentioned in Section I, seizures vary in length from only a few seconds to over several minutes, meaning that there is not a single value of m (number of adjacent windows in the candidate region) that is generally appropriate. In the M2D2 architecture, we are able to use a range of possible values for m to address this issue. Technically, the MMD block computes δ_t^m for various lengths m and allows the decoder to determine the best combination of these values. This non-linear combination depends on not only the current $\mathbf{z}_t$ but also all the adjacent $\mathbf{z}_{\mathbf{t}'}$. Therefore, we use a GRU module in the decoder to find the combination of δ_t^m based on the output of the MMD block in all time steps t'. We use a GRU in preference to a simple RNN to avoid the vanishing and exploding gradient problem [35]. We use a bi-directional GRU because we need information from both $t' \geq t$ and $t' < t$. The output of the GRU layer goes to a simple fully-connected (FC) layer, followed by a linear layer with a sigmoid activation (logistic regression) to predict the output. We show the results of an ablation study in Section VI-A to study the effect of every component of M2D2 on the model's performance.

D. Training

The proposed M2D2 framework can be trained end-to-end via back-propagation using standard methods based on stochastic gradient descent. To show how gradients are computed for the MMD layer, let θ be the parameters of the last layer in the encoder. Then, the gradient of δ_t with respect to θ is given by:

$$\frac{\partial \delta_t}{\partial \theta} = \frac{\partial \delta_t}{\partial z_t} \cdot \frac{\partial z_t}{\partial \theta}, \tag{2}$$

$$\frac{\partial \delta_t}{\partial z_t} = \frac{2}{m^2} \sum_{i=t+1}^{t+m-1} k'(\mathbf{z}_i, \mathbf{z}_t) + \frac{1}{m^2} k'(\mathbf{z}_t, \mathbf{z}_t) - \frac{2}{mL} \left(\sum_{i=t+1}^{t+m-1} k'(\mathbf{z}_i, \mathbf{z}_t) + \sum_{j=1}^{L} k'(\mathbf{z}_t, \mathbf{z}_j) \right) \tag{3}$$

where k' is the derivative of the kernel function. A detailed derivation can be found in the appendix.

As (3) shows, the gradient is obtained without multiplication through the time steps, which addresses the problem of vanishing and exploding gradients. A minor issue is that the gradient involves a sum over a large number of terms in every time step. This may cause the gradient to become large in absolute magnitude, which leads to large fluctuations in the weights. To address this problem, we add a penalty, defined as $\lambda(\|z\|_2 - 1)$ for $\lambda > 0$ a tunable parameter, that helps to control the magnitude of k.

IV. Experiments

A. Datasets

In this work, we consider the setup of real-world and stigma-free wearable monitoring devices [36]. In such settings, in order to make monitoring devices energy efficient and visually unobtrusive, one typically only has access to a reduced set of electrodes. Thus, to be reflective of practically relevant settings, in the datasets, we consider only the electrodes F7T3 and F8T4 in the standard 10–20 system, [37], which can be easily hidden in glasses [13]. The datasets used in this work are as follows:

1) Epilepsiae [38]: This dataset is one of the largest public databases in the world for seizure disorder [39], [40]. It contains totally 4747 EEG recordings from 30 different epilepsy patients. From these recordings, 262 recordings contains at least one epileptic seizure. The data is collected from child and adolescent patients in hospitals across multiple countries. The EEG data is divided into recording sessions of up to one hour. The number of total recordings varies for each patient between 96 and 281 sessions. The total length of seizures in this database is 348 minutes. The average length of each epileptic seizure is 76.5 ± 76.8 seconds.

2) CHB-MIT [41]: This dataset consists of EEG recording for originally-labeled 23 patients sampled at a frequency of 256 Hz. The data is collected from pediatric patients at the Children's Hospital of Boston (CHB) in the United States. The recording length varies in different patients from one hour up to four hours. In total, the dataset contains 664 EEG recordings from which 129 recordings contain epileptic seizures with a total of 182 seizures. Totally the dataset has 182.2 minutes of seizure time out of 961 hours of signal. The length of seizure attacks is 60.1 ± 67.1 seconds, by average.

B. Baseline Methods

We compare our method against the following baselines, which are modeled after methods previously proposed in the literature.

1) Baseline VIB (B-VIB) [21]: Our work in [21] uses the variational information bottleneck approach described in the Background section. We consider the architecture proposed in this work. The decoder is a standard fully-connected network which consists of a single hidden layer followed by a linear layer which outputs the probability that a given window–$\mathbf{w}_i$–contains a seizure. The imputed location of the seizure is taken to be the window that maximizes this probability (e.g., has the highest $\hat{y}$).

2) Baseline MMD (B-MMD): In this baseline, first, we train a Variational Autoencoder (VAE) [42] whose encoder is identical to the B-VIB approach described above. Using the pre-trained encoder in this VAE, we extract the codewords ($\mathbf{z}_i$). After extracting the codewords for all w_i in the session, we apply MMD computation as described in (1). For all t in the signal, we compute a score δ_t, which is the MMD between a candidate seizure region and the remainder of the signal. Similar to our proposed method, this baseline uses MMD to find the seizure temporal location. However, instead of using the codeword vector as input to a decoder, B-MMD imputes the location of the seizure as the window that maximizes δ_t. Since MMD is computed separately from the CNN encoder, the output cannot back-propagate to the encoder to fine-tune the weights and parameters. Therefore, we categorize this baseline as unsupervised learning with MMD. This baseline is analogous to [34] except for the embedding part, which in B-MMD, a deep learning method is used, whereas [34] uses a pre-determined set of features.

3) Fully Convolutional Network (B-FCN) [40]: This approach reshapes the signal into a 3D array, which can be loosely interpreted as an "image," and uses a convolutional neural network to perform classification. This work applies 23-channel EEG signals to the network. Therefore, to have a fair comparison, we apply the same method to the two-channel datasets used here and retrain the network accordingly. For each window, the network returns a predicted probability that the window contains a seizure. We localize the seizure as the window maximizing this value.

4) Medically-Relevant Features (B-FET) [13]: This baseline manually extracts medically-relevant time and frequency domain features from the EEG data, and trains a Random Forest classifier on the feature space. The features consist of various entropy measures and the spectral power of the EEG signal in specific frequency bands. The entropy measures in this baseline include suggested features in [43], [44], such as sample entropy, permutation entropy, and Renyi entropy. Also, the feature vector has Shannon entropy and Tsallis entropy. Furthermore, the absolute and relative band powers calculated as features in B-FET are δ, θ, α, β, and γ. These features are commonly considered relevant by clinicians in the context of epilepsy [45]. The predicted seizure location is the window with the highest score returned by the random forest.

C. Evaluation Method

In this work, our goal is to find the location of a seizure within a long EEG recording. We define the evaluation error as the distance between the detected seizure location to the nearest $\mathbf{w}_i$, which is a seizure signal in the ground truth. Thus, if the detected point is inside the interval of seizures, the error will be zero. We only consider sessions containing at least one seizure. In a real-world case, we assume that the patient is able to indicate that they experienced a seizure within one hour (e.g., via interaction with a monitoring device). In general, being able to localize seizures in long time periods is useful since after a seizure attack, patients may be disoriented or unconscious.

Following standard practice, we partition our data into train, validation, and test sets. The training set is used to fit model parameters, the validation set is used for hyperparameter tuning and model selection, and the test set is used to obtain a final estimate of the out-of-sample error for the model minimizing the validation error. Our hyperparameter tuning methodology is described in the Appendix. We use the following methods for partitioning the data:

1) Leave-One-Out Cross Validation (LOOCV): We here partition the data into train, test, and validation sets using the principle of "leave-one-out" cross-validation. In LOOCV, one cycles through each patient in the dataset, holding out their data as a test set. The remaining patients are used for training and validation.

2) New Unseen Test Set: To evaluate the performance of our method on a completely different dataset from which it was trained, we perform another set of experiments in which we train and validate on one dataset, but test on the other (e.g., train on CHB-MIT, test on Epilepsiae). This setting is potentially more challenging since the test set is derived from a different clinical setting. However, it is more reflective of the actual performance our model would achieve if it were applied beyond the immediate clinical setting in which it was trained. To the best of our knowledge, we are the first work to perform this type of evaluation in the context of evaluating seizure detection procedures.

D. Implementation Details

We here describe key details of M2D2 implementation. A more detailed description can be found in the Appendix.

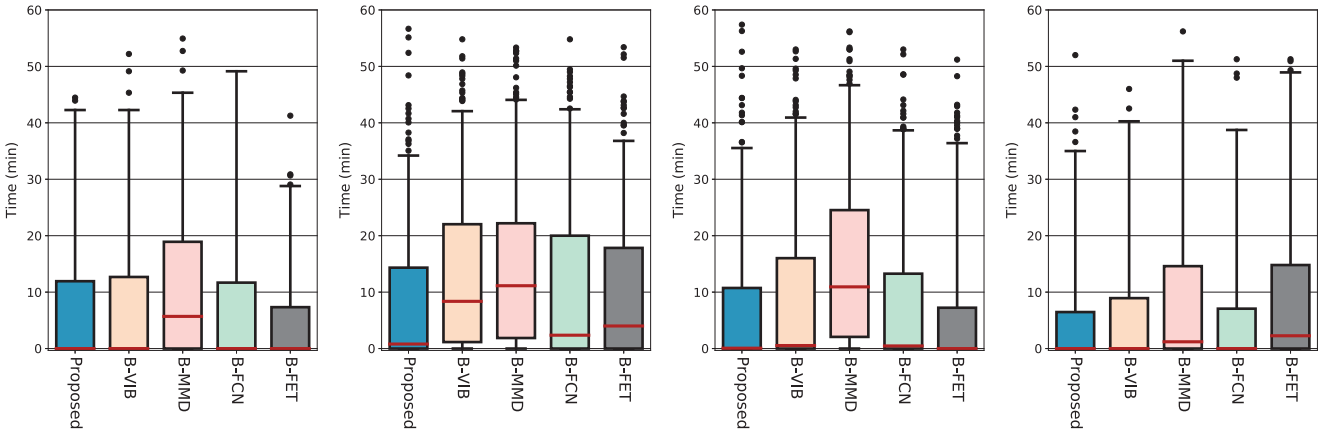
1) Hyper-Parameters Tuning: In the LOOCV evaluation, for every patient, we train a separate model. The EEG recordings associated with the held out patient form the test set while the other 22 patients are in the training and validation set. In the unseen new dataset evaluation, the test set is from the Epilepsiae (CHB-MIT) dataset; thus, we choose all the hyper-parameters, based on the validation set in the CHB-MIT (Epilepsiae) dataset.

The latent space length d is the most important hyper-parameter, which is chosen based on grid search. The possible values are 2, 4, 8, 16, and 32. We select the value leading to the lowest cross-validation error. Regarding the kernel selection, when computing the MMD, we choose k using cross-validation between polynomial kernels: $k(x,z) = (1 + \langle x, z\rangle)^n,\ n \in \{1,2,3,4,5\}$, and radial basis functions (RBF) kernels: $k(x,z) = \exp(-\gamma||x-z||^2),\ \gamma \in \{0.01, 0.1, 1, 10, 100\}$. However, we show in Section VI-E that our method is generic to any particular choice of kernel. We train the models for 100 epochs or until the validation loss fails to decrease for ten consecutive epochs.

2) MMD Simplification: Given m samples from $\mathcal{P}$ (the candidate ictal region), and n samples from $\mathcal{Q}$, the exact computation of the MMD is $O((m+n)^2)$. However, in our case, $\mathcal{P}$ is very small compared to $\mathcal{Z}$ (the entire recording), and so the second term in (1) changes very little as $\mathcal{P}$ is varied. Accordingly, to simplify implementation, we set $\mathcal{Q} = \mathcal{Z}$ in which case the second term of (1) is a constant. This approximate MMD reduces the computation significantly to $O(m(m+n))$ (recall that m is small) and consequently reduces training and inference time.

To understand the implications of using the simplified form of MMD, we perform the following experiment. After training the proposed model, we freeze the weights of the encoder. Then, we extract the latent representation of all input signals using the "LOOCV" method in CHB-MIT. Next, we calculate both the exact and simplified MMD for different window sizes, and compute the correlation between the exact and simplified values. The average of correlation coefficient for different window length m is obtained as 0.95 ± 0.01. On the other hand, the amount of computation saved using the simplified MMD is between 98.1% to 99.4% for different m values. Consequently, simplified values are tightly correlated with the exact ones while dramatically reducing computational overhead. In Section VI-F, we discuss, visualize, and compare exact and simplified MMD in more detail.

3) Pre-Processing and Implementation: The EEG signal is pre-processed using a Butterworth 50 Hz low-pass filter and



(a) CHB-MIT, LOOCV (b) CHB-MIT → Epilepsiae (c) Epilepsiae, LOOCV (d) Epilepsiae → CHB-MIT

Fig. 2. Distribution of errors in temporal localization under different evaluations when models are trained on one dataset and tested on the same dataset (LOOCV) or on a new unseen test set (→).

by standardizing each 4-second window to have zero mean and unit-variance input signals.

We train and test our models on a platform with an 8-core Intel i7-9700 K CPU and a single NVIDIA RTX 2080 GPU with 2944 CUDA cores.

V. Results

A. Temporal Localization Error Distribution

Fig. 2 summarizes the results. The boxplots show the distribution of the errors in seizure temporal localization. The red line indicates the median error in each box, and the upper and lower limits of the colored box indicate the 75^{th} and 25^{th} percentile, respectively. The upper and lower whiskers indicate the 95^{th} and 5^{th} percentiles, respectively, and the black dots indicate outliers. To keep plots readable, we do not show outliers with over one hour of error. We emphasize that this is just for display purposes, and reported numeric results are inclusive of all data.

As shown in Figs. 2(b) and 2(d), we find that our approach yields superior performance when evaluated on a different dataset than was used for training. In other words, when training on CHB-MIT and evaluating on Epilepsiae (shown in Fig. 2(b)) or vice-versa (shown in Fig. 2(d)), our approach is able to localize seizures with lower error than any of the baselines. These results are consistent with our goal of developing techniques for temporally localizing seizures that offer better generalization in new data settings.

In Figs. 2(a) and 2(b), the same dataset is considered for training and testing. As shown, our approach delivers competitive performance in the leave-one-out evaluation. For instance, in these figures, our proposed method has a median of zero, meaning that the imputed temporal location of a seizure falls within an actual seizure in over half of the cases.

By comparing the results of our proposed model with the baselines, we see that B-MMD has a wider distribution with a substantially higher median error. While at first glance the B-MMD and M2D2 methods appear similar, they are trained quite differently. In M2D2 the encoder is trained end-to-end in a supervised fashion and, thus, can fine-tune the extracted features for the MMD computation. By contrast, in B-MMD, the encoder cannot be fine-tuned, and thus the features are extracted by optimizing an unsupervised cost function. Moreover, the B-MMD is only able to consider a fixed length candidate region for seizures, whereas our approach can consider multiple possible window lengths. This underscores the value of using a neural network to learn good signal features in our approach. On the other hand, we can see that B-FET offers the best performance when evaluated using leave-one-out. However, as can be seen in Fig. 2(d), this model has the largest median error when evaluated on CHB-MIT as an unseen test set. This emphasizes the need to develop models which can maintain performance when applied outside of the data environment from which they were trained.

B. Quantitative Results

The proposed M2D2 model and the decoders in B-VIB, B-FCN, and B-FET return a value $\hat{y}_i \in [0, 1]$ for every input $\mathbf{w}_i$ corresponding to the probability that $\mathbf{w}_i$ is a seizure. We define a threshold τ with the condition that if $\hat{y}_i < \tau$, then the signal in i is detected as a non-seizure. Also, if $\hat{y}_i >= \tau$, a seizure point is detected by the model. This definition is used in the temporal localization task, and new metrics are defined inspired by [46]. For a one-hour EEG recording that contains seizures, if all the outputs $\hat{y}_i$ for the whole signal are less than τ, then we categorize the signal as a False Negative (FN). True Positive (TP) is defined as the points i^* where $\hat{y_{i^*}} > \tau$ and i^* is in the ground truth. Likewise, False Positive (FP) points are the points i^* where $\hat{y_{i^*}} > \tau$ but they are not in the ground truth.

The metrics precision, recall, and F1-score are defined as follows for all the models. The threshold τ is chosen for every

TABLE I
PRECISION, RECALL, AND F1-SCORE UNDER DIFFERENT EVALUATIONS. FOR THE LOOCV EVALUATIONS, THE RESULTS ARE REPRESENTED AS THE AVERAGE OF EACH METRIC FOR EVERY PATIENTS

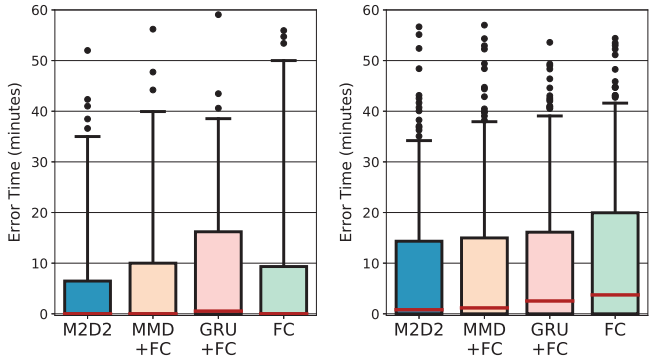
Method	Recall(%)	Precision (%)	F1-score (%)
CHB-MIT, LOOCV			
Proposed	65.6	62.4	60.3
B-VIB	70.7	58.7	61.2
B-FCN	72.2	55.0	58.4
B-FET	79.9	65.4	66.2
CHB-MIT → Epilepsiae			
Proposed	63.9	78.2	70.4
B-VIB	52.6	14.6	22.8
B-FCN	46.6	51.9	49.2
B-FET	3.0	99.9	5.8
Epilepsiae, LOOCV			
Proposed	63.7	63.5	61.4
B-VIB	63.5	54.1	55.1
B-FCN	75.4	47.9	55
B-FET	56.3	74.8	60.1
Epilepsiae → CHB-MIT			
Proposed	92.7	64.4	76.0
B-VIB	99.9	6.1	11.5
B-FCN	99.9	7.6	14.1
B-FET	99.7	8.7	16.0

single model to optimize the F1-score in the validation set.

$$\text{Precision} = \frac{TP}{TP+FP}, \ \text{Recall} = \frac{TP}{TP+FN},$$
$$\text{F1-score} = \frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}}$$

The results are represented in Table I. As seen in this table, in the new unseen test set evaluation, the proposed model has an F1-score significantly better than the baseline methods. To understand the reason for this gap between the proposed results with the baseline methods, one can note that M2D2 is trained to predict the seizure probability of each point using the comparison between two distributions. On the other hand, the baseline methods predict seizures only based on the input window $\mathbf{w}_i$ regardless of the rest of the signal. Therefore, the M2D2 model has a limited number of points with high probability in the output, while the baseline methods can have high probable seizure points as many as all the inputs. As a consequence, the FP points increase in the baselines, and the precision metric decrease. Note that the baseline models still provide the maximum probability for a point close to the ground truth; thus, they perform well in the temporal localization discussed in Section V-A. However, using the quantitative results provided in this section, we show the better performance of M2D2 if the model is applied beyond the immediate clinical setting in which it was trained.

The B-MMD baseline is excluded from the experiment because its MMD output is not a probability limited between zero and one, and the definition of τ is not as same as the other baselines.



(a) Epilepsiae → CHB-MIT (b) CHB-MIT → Epilepsiae

Fig. 3. Ablation study for the M2D2 decoder and comparison with models without MMD block, GRU layer or both. As seen, the least error is when the complete M2D2 is used.

C. Temporal Localization With Acceptable Errors

In Table II, we show the number of EEG sessions in which at least one seizure is correctly localized in time. A seizure is correctly localized with an acceptable temporal error of Δ if, the time distance between the predicted seizure point and the nearest ground truth (GT) seizure is less than Δ. Therefore, each model chooses a single point i^*, and if there is a point g in GT where $(|i^* - g| \leq \Delta)$, it is a hit, otherwise, it is a miss. The total number of hits are represented in the Table 1 as Top-1 results. Similarly, for the top-3 result in this table, the model chooses three different points $I^* = \{i_1^*, i_2^*, i_3^*\}$, and if there is a point g in GT where for any i^*, we obtain $(|i^* - g| \leq \Delta)$, it is a hit.

All the projected numbers in this table are obtained by running every experiment three times and reporting their median number. As can be seen, these results provide corroboration that our method is able to localize seizures with higher precision than the baselines when evaluated on the unseen test set.

VI. DISCUSSION OF RESULTS

We here offer additional discussion of results as well as some additional analysis of our approach that aims to provide insight into why it exhibits better generalization than the baselines.

A. Ablation Study

The M2D2 decoder contains MMD, GRU, and fully-connected (FC) layers. To evaluate the contribution of these layers, we performed the ablation study and trained four different models with identical encoders and different decoders. The decoders are (1) a complete M2D2 module, (2) an MMD block followed by an FC layer, (3) a GRU followed by an FC layer, and (4) a single FC layer. The results of the unseen evaluation are shown in Fig. 3.

We observe that removing any component of the M2D2 decoder causes an error increase. Therefore, the results underscore that each element is necessary for the model to perform as desired. In particular, the decoder with GRU and FC layers is an

TABLE II
NUMBER OF CORRECTLY LOCALIZED SEIZURES UNDER DIFFERENT EVALUATIONS FOR VARIOUS DURATION OF THE TARGET WINDOW. THE NUMBERS ARE OUT OF THE TOTAL NUMBER OF SEIZURE SESSIONS IN THE TEST SET, I.E., 262 SESSIONS IN EPILEPSIAE AND 129 SESSIONS IN CHB-MIT. THE PROPOSED METHOD OUTPERFORMS THE BASELINES IN THE TEMPORAL LOCALIZATION FOR THE NEW UNSEEN TEST SET EVALUATION

Acceptable Temporal Error Δ		CHB-MIT LOOCV (/129)					CHB-MIT → Epilepsiae (/262)					Epilepsiae LOOCV (/262)					Epilepsiae → CHB-MIT (/129)				
		proposed	B-VIB [21]	B-FCN [40]	B-FET [13]	B-MMD [34]	proposed	B-VIB [21]	B-FCN [40]	B-FET [13]	B-MMD [34]	proposed	B-VIB [21]	B-FCN [40]	B-FET [13]	B-MMD [34]	proposed	B-VIB [21]	B-FCN [40]	B-FET [13]	B-MMD [34]
2 Sec.	Top-1	**79**	67	71	72	39	**115**	41	93	85	32	138	116	121	**147**	68	72	69	**76**	54	56
	Top-3	83	70	77	**84**	44	**133**	73	126	118	42	145	142	160	**170**	88	82	82	**87**	74	61
18 Sec.	Top-1	**83**	68	71	76	46	**130**	54	98	93	41	144	123	124	**152**	74	**82**	71	78	56	61
	Top-3	88	75	77	**89**	59	**168**	105	148	132	79	169	167	172	**185**	113	**92**	85	91	78	66
30 Sec.	Top-1	**85**	68	72	78	51	**132**	60	106	98	50	153	131	132	**160**	76	**82**	72	79	57	62
	Top-3	**92**	76	79	**92**	68	**172**	115	164	146	102	184	174	183	**195**	137	**93**	86	92	82	69
60 Sec.	Top-1	**85**	70	74	81	54	**147**	63	114	106	57	158	131	141	**168**	80	**84**	75	79	59	64
	Top-3	93	78	82	**95**	73	**179**	126	178	160	138	194	178	193	**202**	159	**95**	88	92	85	73
150 Sec.	Top-1	**90**	76	77	85	59	**155**	84	132	117	77	165	143	153	**177**	100	**90**	79	84	65	70
	Top-3	**104**	90	91	102	86	**202**	165	201	178	171	204	200	211	**217**	188	99	94	**101**	92	85
300 Sec.	Top-1	**93**	81	84	90	64	**163**	99	142	135	93	176	156	167	**189**	117	**95**	86	91	73	75
	Top-3	107	95	104	**111**	97	**230**	209	224	215	213	211	231	234	**240**	220	**115**	109	108	105	102

RNN where it is impossible to define a candidate seizure region. The GRU layer only compares the current feature vectors with a non-linear combination of the remainder of the signal.

B. Class Separability Measures

We hypothesize that the MMD layer in our approach may lead to better separation between the ictal (seizure) and non-ictal representations ($\mathbf{z}$). Since the decoder only has access to these $\mathbf{z}$-space features, discriminating the two classes will be easier if the MMD between the $\mathbf{z}$-corresponding to each class is large. We compare the separability of the $\mathbf{z}$ produced by our method and the baselines using the J-score. Intuitively, the J-score compares the distances between samples "within" a class, and "between" samples in different classes. If the within-class distance is small relative to the between-class distance, then the J-score is large, indicating better separability. The J-score is computed from the within and between-class scatter matrices [47] as follows:

$$\mathbf{S}_W = \sum_{i=1}^{n^+} (\mathbf{z}_i^+ - \mathbf{m}^+)(\mathbf{z}_i^+ - \mathbf{m}^+)^T + \sum_{i=1}^{n^-} (\mathbf{z}_i^- - \mathbf{m}^-)(\mathbf{z}_i^- - \mathbf{m}^-)^T$$

$$\mathbf{S}_B = n^+(\mathbf{m}^+ - \mathbf{m}) + n^-(\mathbf{m}^- - \mathbf{m})$$

where n^+ and n^- denote the number of samples in the seizure and non-seizure classes, respectively. Similarly, $\mathbf{z}_i^+$ and $\mathbf{z}_i^-$ denote the i-th sample in the seizure class and the i-th sample in the non-seizure class, respectively. $\mathbf{m}^+$ and $\mathbf{m}^-$ denote the mean vectors of the samples in seizure and non-seizure classes. Finally, $\mathbf{m}$ denotes the mean vector of all samples. The class separability measure is defined as $J = {}^{\text{trace}(\mathbf{S}_B)}/_{\text{trace}(\mathbf{S}_W)}$, where a small within-class scatter and large between-class scatter cause a large separability.

TABLE III
CLASS SEPARABILITY OF CODEWORDS $\mathbf{z}$ FOR THE PROPOSED METHOD AND THE BASELINES WITH THE SAME ENCODER STRUCTURE

Evaluation	proposed	B-VIB [21]	B-MMD [34]
CHB-MIT → Epilepsiae	0.29±0.1	0.08±0.03	0.01±0.002
Epilepsiae → CHB-MIT	0.27±0.11	0.02±0.02	0.01±0.001

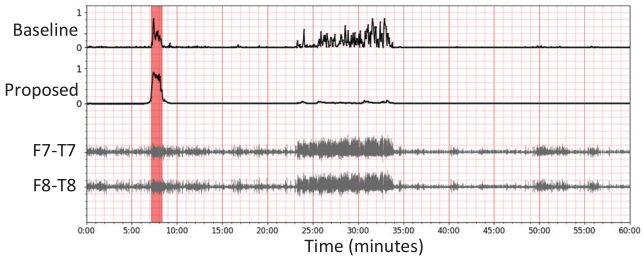


Fig. 4. Two-channel EEG signal for one hour recording and the output of B-VIB and the proposed models.

Table III compares J-scores of our approach and the baselines for the unseen evaluation methods. As we see in the table, the J-score for the proposed method is much larger than the baseline methods indicating that the derived representations of seizure and non-seizure points are better separated than in the baseline methods. Note that the B-FET and B-FCN baselines do not have an analogous derived representation and so are not included here.

C. Artifact Study

Fig. 4 shows a one-hour session of EEG signal extracted from the CHB-MIT dataset with two different channels, F7-T7 and F8-T8. The annotation of this signal indicates that the seizure occurs from time 7':12" until 8':21". This seizure time is shown in Fig. 4 with a red rectangle span on the background. Other

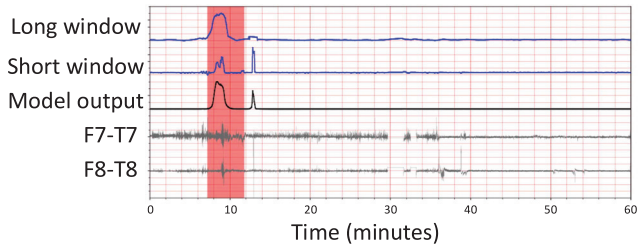


Fig. 5. The effect of a spike on the output of the MMD with short and long candidate region of seizure.

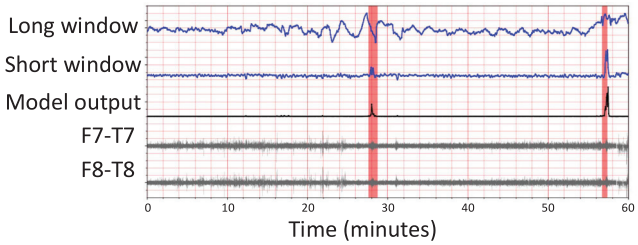


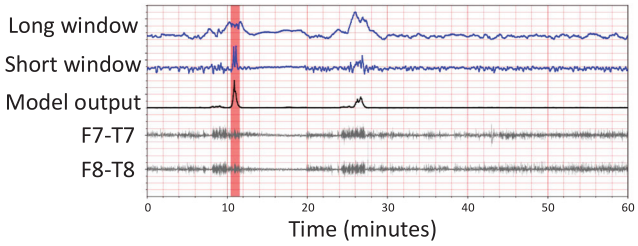
Fig. 6. The effect of multiple seizures on the MMD with short and long candidate region of seizure.

parts of the signal are normal EEG, but as we can see in the figure, from 23‘:18“ until 33‘:50“, there is an artifact. As shown in the figure, the proposed model can make the artifact ineffective while the baseline model can predict the artifact as a seizure window. This is because the MMD layer compares the distribution of codewords inside and outside a short window. The seizure lengths (1:09 for this case) are usually shorter than these artifacts (the artifact duration is 10:32). Therefore, our proposed model can realize the similarity of samples from inside and outside the candidate window (with the nominal duration of seizures for each patient) and then detect them as non-seizure segments.

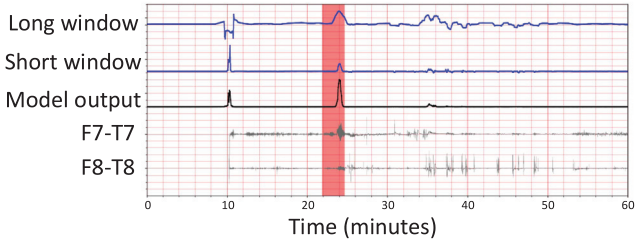
D. Window Length in M2D2

As mentioned in Section III, seizures vary fairly widely in length, and in M2D2, we use a variety of window lengths m to cover different seizure lengths. In this section, we describe some usual problems in EEG signals, and we show and discuss how various window lengths in M2D2 help to address the problems.

1) Spikes and Sharp Waves: In Section VI-C, we discussed the artifacts in EEG signals and how M2D2 improves robustness to them. “Spikes” and “Sharp waves” are other abnormal waveforms which may appear in EEG signals. A spike is a sharp-pointed peak clearly distinguished from the background and typically lasts between 20 to 70 milliseconds. If the duration is between 70 to 200 milliseconds, the wave becomes a sharp wave [45]. Since we assume the segment length as 4 seconds, and the segments are longer than the duration of spikes and sharp waves, they are usually addressed by the convolution layers in the encoder. However, in some cases, we can see the effect of spikes in the output. Fig. 5 shows a one-hour EEG signal and the output prediction of the proposed method. The seizure is shown with a red rectangle span in the background. A spike occurs one minute after the seizure and lasts 50 ms, and it notable perturbs the corresponding $\mathbf{z}_t$. As shown in this figure, for short window lengths (e.g. 20 seconds), the spike meaningfully changes the output of the MMD layer. On the other hand, for the long window length, which is 68 seconds, the spike is nearly eliminated from the output of the MMD layer. Note that the model output in the figure is the output of the model after the GRU and fully connected layers.



(a) The shorter m works more accurately in a short seizure.



(b) The longer m works more accurately in a long seizure

Fig. 7. The effect of m on temporal localization in recordings with different seizure length.

2) Multiple Seizures in One Session: The frequency of seizures varies from patient to patient. Some patients suffer from more frequent seizures and may have multiple seizures in an hour–the typical length of recordings in our data. Fig. 6 shows a session with two different seizures. The first seizure happened from 27‘:44“ until 28‘:45“, and the second one from 56‘:54“ until 57‘:26“. As we see in the figure, the output of the MMD layer using a short window length (20 seconds) is largest in the second seizure, i.e., the shorter one, which lasts 32 seconds. The first seizure is better detected by the longer window length, underscoring that different window lengths are appropriate for different seizures and that using a static window length as in [9] is not optimal.

3) Different Seizure Length: In general, short seizures are detected by shorter window lengths and longer seizures by longer window lengths. Therefore, in the dataset, there are some cases in which the short and long windows detect different localization. Fig. 7 shows two different sessions with seizures of 64 seconds and 164 seconds, respectively. The MMD layer

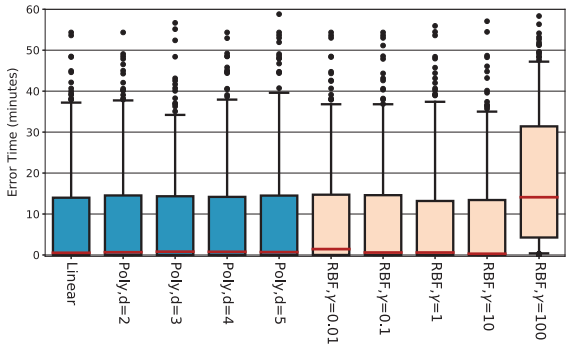


Fig. 8. The robustness of M2D2 to the chosen kernels.

works differently for these seizures, and interestingly, the output is correct for both cases, which shows the rest of the layers in M2D2, i.e., GRU and fully-connected layers, work properly to choose the best value of window length.

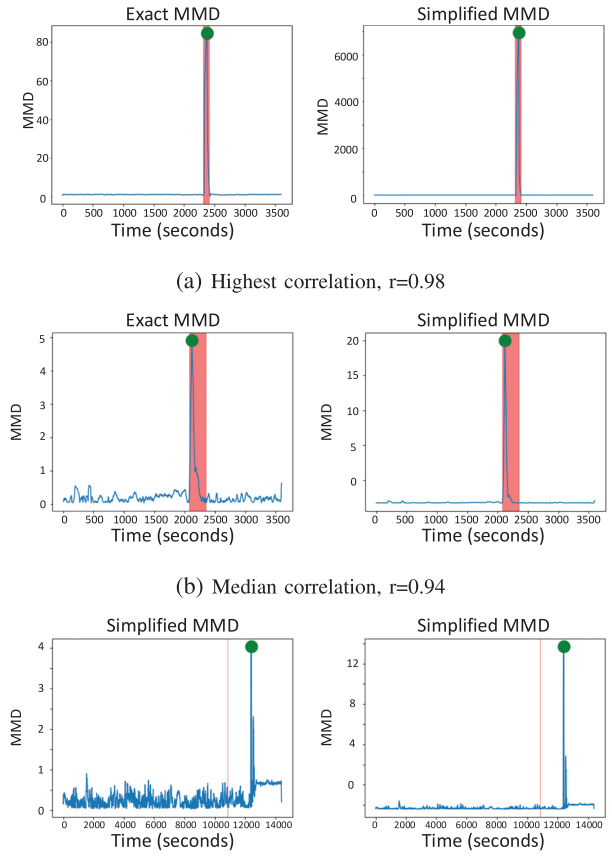
E. Kernel Robustness

In this section, we show the robustness of M2C2 in choosing the kernel function. In this experiment, we train several models with an identical pre-trained encoder with different M2D2 decoders, which vary in the kernel functions. The kernels are chosen between linear, polynomial kernel with orders of 2, 3, 4, and 5, and RBF kernels with γ in range of 0.01, 0.1, 1, 10, and 100. The error time in the unseen test set evaluation of CHB-MIT $\rightarrow$ Epilepsiae is shown in Fig. 8. As we see in this figure, all of the chosen kernels except the RBF kernel with $\gamma = 100$ temporally localize the seizures almost in the same way.

F. Simplified MMD Visualization

As mentioned in Section IV-D2, we use a simplified MMD because the exact MMD is compute-intensive. To discuss the differences between the "simplified MMD" and "Exact MMD," we performed a new experiment. We visualize the MMD for every EEG recording in CHB-MIT containing seizure. We used the leave-one-out cross-validation, and thus, the models are trained and tested on CHB-MIT. The following figures show how much the simplified and exact MMD are different. In Fig. 9(a), we choose the EEG recording, which has the highest correlation between the simplified and exact MMD. The seizure time is shown in a red rectangle span in the background. As we can see in this figure, the trend of the MMD is similar; however, the MMD values shown in the y-axis are different by two orders of magnitude. The absolute values of the MMD has no effect on our work because of the following reason. In this paper, the goal is temporal localization of the seizures in the EEG recording, i.e., to find t in which δ_t has the highest value. Therefore, the absolute value of δ_t is not essential.

Fig. 9(b) corresponds to the EEG recording with a correlation that has the median value among all correlations. The correlation value is between the simplified and exact MMD. Finally, Fig. 9(c) shows the MMDs for an EEG recording with



(a) Highest correlation, r=0.98

(b) Median correlation, r=0.94

(c) Lowest correlation, r=0.84

Fig. 9. The MMD for EEG recordings with the **highest** (a), **median** (b), and **lowest** (c) correlation between the simplified and exact MMD.

the lowest correlation value. The figure shows that the model cannot detect the seizure location because the maximum of the MMD is not inside the red rectangle. However, the simplified MMD and exact MMD have almost the same trends in their values.

VII. Limitations and Future Work

We see two notable limitations of our work. First, our work is in furtherance of developing models for seizure localization that can be deployed on lightweight wearable devices. While we address one important limitation of prior work by developing an empirical approach that offers substantially improved generalization, our model is still heavy-weight (in terms of latency and energy-use) compared to the types of approaches that can be deployed in real world devices. While we experiment with modified versions of the MMD computation that can improve total computation, and hence energy efficiency and latency, an important component of future work will be to develop light-weight realizations of our techniques which can be deployed on practical, wearable devices. In addition, we have evaluated our model on two large publicly available EEG datasets with long relevant EEG recordings; however, this is

likely not reflective of the full diversity of patients with epilepsy, due to the limited amount of data available in the context of epilepsy.

VIII. Conclusion

In this work, we have considered the problem of automatically localizing epileptic seizures from EEG recordings. Existing deep learning-based methods for this problem typically need to be fine-tuned to be applied beyond the immediate data environment in which they were trained. However, this process requires acquiring new labeled training data which is costly to obtain. In this work, we have taken a step towards resolving this issue by introducing the M2D2 neural network architecture for automatic temporal localization of epileptic brain activities in long EEG recordings. Our approach groups together a set of low-dimensional codewords corresponding to a candidate seizure region and introduces a novel decoder architecture which computes a set of features based on the maximum-mean-discrepancy between each candidate region and the remainder of the signal. These features are used by a recurrent decoder to impute the location of a seizure. Using an extensive empirical evaluation, we have shown that this approach leads to substantially better generalization than prior approaches when tested in a completely new data environment without any fine-tuning. From a methodological perspective, our work has introduced a new technique for detecting phenomena of interest in time-series. From a practical perspective, our work has improved existing techniques by reducing the need for fine-tuning and specialization of models for seizure detection to new data environments.

References

[1] F. Tang, A. Hartz, and B. Bauer, "Drug-resistant epilepsy: Multiple hypotheses, few answers," *Front. Neurol.*, vol. 8, 2017, Art. no. 301.

[2] X. Zhang, L. Yao, M. Dong, Z. Liu, Y. Zhang, and Y. Li, "Adversarial representation learning for robust patient-independent epileptic seizure detection," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 2852–2859, Oct. 2020.

[3] L. Orosco, A. G. Correa, P. Diez, and E. Laciar, "Patient non-specific algorithm for seizures detection in scalp EEG," *Comput. Biol. Med.*, vol. 71, pp. 128–134, 2016.

[4] M. Zabihi, S. Kiranyaz, V. Jäntti, T. Lipping, and M. Gabbouj, "Patient-specific seizure detection using nonlinear dynamics and nullclines," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 543–555, Feb. 2020.

[5] P. Detti, G. Z. M. de Lara, R. Bruni, M. Pranzo, F. Sarnari, and G. Vatti, "A patient-specific approach for short-term epileptic seizures prediction through the analysis of EEG synchronization," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 6, pp. 1494–1504, Jun. 2019.

[6] A. Burrello, S. Benatti, K. Schindler, L. Benini, and A. Rahimi, "An ensemble of hyperdimensional classifiers: Hardware-friendly short-latency seizure detection with automatic iEEG electrode selection," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 4, pp. 935–946, Apr. 2021.

[7] S. Baghersalimi, T. Teijeiro, D. Atienza, and A. Aminifar, "Personalized real-time federated learning for epileptic seizure detection," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 2, pp. 898–909, Feb. 2022.

[8] S. Jenssen, E. J. Gracely, and M. R. Sperling, "How long do most seizures last? a systematic comparison of seizures recorded in the epilepsy monitoring unit," *Epilepsia*, vol. 47, no. 9, pp. 1499–1503, 2006.

[9] D. Pascual, A. Aminifar, and D. Atienza, "A self-learning methodology for epileptic seizure detection with minimally-supervised edge labeling," in *Proc. Des., Automat. Test Europe Conf. Exhib.*, 2019, pp. 764–769.

[10] J. W. Britton et al., *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*. Chicago, IL, USA: American Epilepsy Society, 2016.

[11] J. A. Urigüen and B. Garcia-Zapirain, "EEG artifact removal—state-of-the-art and guidelines," *J. Neural Eng.*, vol. 12, no. 3, 2015, Art. no. 0 31001.

[12] X. Jiang, G.-B. Bian, and Z. Tian, "Removal of artifacts from EEG signals: A review," *Sensors*, vol. 19, no. 5, 2019, Art. no. 987.

[13] D. Sopic, A. Aminifar, and D. Atienza, "e-glass: A wearable system for real-time detection of epileptic seizures," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2018, pp. 1–5.

[14] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, 2012.

[15] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Berlin, Germany: Springer, 2011.

[16] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Allerton Conf. Commun., Control Comput.*, vol. 49, 2001.

[17] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *J. Mach. Learn. Res.*, vol. 19, no. 1, pp. 1947–1980, 2018.

[18] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Int. Conf. Learn. Representations*, 2017.

[19] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, "Nonlinear information bottleneck," *Entropy*, vol. 21, 2019, Art. no. 1181, doi: 10.3390/e21121181.

[20] A. Achille and S. Soatto, "Information dropout: Learning optimal representations through noisy computation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2897–2905, Dec. 2018.

[21] A. H. Thomas, A. Aminifar, and D. Atienza, "Noise-resilient and interpretable epileptic seizure detection," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2020, pp. 1–5.

[22] B. Litt and J. Echauz, "Prediction of epileptic seizures," *Lancet Neurol.*, vol. 1, no. 1, pp. 22–30, 2002.

[23] P. Celka and P. Colditz, "A computer-aided detection of EEG seizures in infants: A singular-spectrum approach and performance comparison," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 5, pp. 455–462, May 2002.

[24] N. Kannathal, M. L. Choo, U. R. Acharya, and P. Sadasivan, "Entropies for detection of epilepsy in EEG," *Comput. Methods Programs Biomed.*, vol. 80, no. 3, pp. 187–194, 2005.

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

[26] S. B. Wilson, M. L. Scheuer, R. G. Emerson, and A. J. Gabor, "Seizure detection: Evaluation of the Reveal algorithm," *Clin. Neurophysiol.*, vol. 115, no. 10, pp. 2280–2291, 2004.

[27] A. Emami, N. Kunii, T. Matsuo, T. Shinozaki, K. Kawai, and H. Takahashi, "Seizure detection by convolutional neural network-based analysis of scalp electroencephalography plot images," *NeuroImage: Clin.*, vol. 22, 2019, Art. no. 101684.

[28] M. Zhou et al., "Epileptic seizure detection based on EEG signals and CNN," *Front. Neuroinform.*, vol. 12, 2018, Art. no. 95.

[29] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[30] G. Liu, L. Tian, and W. Zhou, "Patient-independent seizure detection based on channel-perturbation convolutional neural network and bidirectional long short-term memory," *Int. J. Neural Syst.*, vol. 32, no. 06, 2022, Art. no. 2150051.

[31] S. Tang et al., "Self-supervised graph neural networks for improved electroencephalographic seizure analysis," in *Int. Conf. Learn. Representations*, 2021.

[32] S. Baghersalimi, A. Amirshahi, F. Forooghifar, T. Teijeiro, A. Aminifar, and D. Atienza, "Many-to-one knowledge distillation of real-time epileptic seizure detection for low-power wearable Internet of Things systems," 2022, *arXiv:2208.00885*.

[33] M. Sinn, A. Ghodsi, and K. Keller, "Detecting change-points in time series by maximum mean discrepancy of ordinal pattern distributions," in *Proc. 28th Conf. Uncertainty Artif. Intell.*, 2012.

[34] B. Hamzi, T. N. AlOtaiby, S. AlShebeili, and A. AlAnqary, "Kernel methods and the maximum mean discrepancy for seizure detection," in *Proc. 1st Int. Conf. Comput. Appl. Inf. Secur.*, 2018, pp. 1–6, doi: 10.1109/CAIS.2018.8441977.

[35] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural network," in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 28, no. 3, 2013, pp. 1310–1318.
[36] C. Hoppe, M. Feldmann, B. Blachut, R. Surges, C. E. Elger, and C. Helmstaedter, "Novel techniques for automated seizure registration: Patients' wants and needs," *Epilepsy Behav.*, vol. 52, pp. 1–7, 2015.
[37] G. H. Klem et al., "The ten-twenty electrode system of the international federation," *Electroencephalography Clin. Neurophysiol.*, vol. 52, no. 3, pp. 3–6, 1999.
[38] J. Klatt et al., "The EPILEPSIAE database: An extensive electroencephalography database of epilepsy patients," *Epilepsia*, vol. 53, no. 9, pp. 1669–1676, 2012.
[39] D. Pascual, A. Amirshahi, A. Aminifar, D. Atienza, P. Ryvlin, and R. Wattenhofer, "EpilepsyGAN: Synthetic epileptic brain activities with privacy preservation," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 8, pp. 2435–2446, Aug. 2021.
[40] C. Gómez, P. Arbeláez, M. Navarrete, C. Alvarado-Rojas, M. Le Van Quyen, and M. Valderrama, "Automatic seizure detection based on imaged-EEG signals through fully convolutional networks," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, 2020.
[41] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Massachusetts Institute of Technology, Harvard University, Cambridge, MA, USA, 2009.
[42] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. 2nd Int. Conf. Learn. Representations*, Banff, AB, Canada, Apr. 14–16, 2014.
[43] A. Anier, T. Lipping, V. Jäntti, P. Puumala, and A.-M. Huotari, "Entropy of the EEG in transition to burst suppression in deep anesthesia: Surrogate analysis," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, 2010, pp. 2790–2793.
[44] N. Mammone, J. Duun-Henriksen, T. W. Kjaer, and F. C. Morabito, "Differentiating interictal and ictal states in childhood absence epilepsy through permutation Rényi entropy," *Entropy*, vol. 17, no. 7, pp. 4627–4643, 2015.
[45] N. Kane et al., "A revised glossary of terms most commonly used by clinical electroencephalographers and updated proposal for the report format of the EEG findings. revision 2017," *Clin. Neurophysiol. Pract.*, vol. 2, 2017, Art. no. 170.
[46] R. Padilla, W. L. Passos, T. L. Dias, S. L. Netto, and E. A. Da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," *Electronics*, vol. 10, no. 3, 2021, Art. no. 279.
[47] L. Wang and K. Chan, "Learning kernel parameters by using class separability measure," in *Proc. 6th Kernel Mach. Workshop, Conjunction Neural Inf. Process. Syst.*, 2002, pp. 1–8.