M2NN: Rare Event Inference through Multi-variate Multi-scale Attention

Manjusha Ravindranath
Arizona State University
Tempe, USA
mravind1@asu.edu

K. Selçuk Candan
Arizona State University
Tempe, USA
candan@asu.edu

Maria Luisa Sapino
University of Torino
Torino, Italy
mlsapino@di.unito.it

***Abstract*—With the increasing availability of sensory data, inferring the existence of relevant events in the observations is becoming a critical task for smart data service delivery in applications that rely on such data sources. Yet, existing solutions tend to fail when the events that are being inferred are rare, for instance when one attempts to infer *seizure events in electroencephalogram (EEG) data*. In this paper, we note that multi-variate time series often carry robust localized multi-variate temporal features that could, at least in theory, help identify these events; however, the lack of sufficient data to train for these events make it impossible for neural architectures to identify and make use of these features. To tackle this challenge, we propose an LSTM-based neural architecture, $M2NN$, with an attention mechanism that leverages robust multivariate temporal features that are extracted *a priori* and fed into the NN as a side information. In particular, multi-variate temporal features are extracted by simultaneously considering, at multiple scales, temporal characteristics of the time series along with external knowledge, including variate relationships that are known a priori. We then show that a single layer LSTM with dual-layer attention that leverages these multi-scale, multi-variate features provides significant gains in rare seizure detection on EEG data. In addition, in order to illustrate the broader applicability (and reproducibility) of $M2NN$, we also evaluate it in other publicly available rare event detection tasks, such as anomaly detection in manufacturing. We further show that the proposed $M2NN$ technique is beneficial in tackling more traditional inference problems, such as travel-time prediction, where rare accident events can cause congestions.**

***Keywords*-Brain EEG analysis, Rare event inference, Multi-variate time series, Multi-scale attention**

I. Introduction

There are many applications generating and consuming multivariate timeseries. With the increasing availability of such data, inferring the existence of relevant events in the observations is becoming a critical to effective delivery of smart data services in critical domains, including healthcare, manufacturing and logistics, and transportation.

A. Motivating Application: Post Traumatic Seizure Detection

Seizures occur as a result of various damaging events to the brain like central nervous system infections, intracranial hemorrhage, stroke, brain injury or cancer. Unfortunately, seizures are more wide-spread in population than most expects – about one percent of Americans have some form of epilepsy, and nearly four percent will develop epilepsy at some point in their lives [1]. Furthermore, the cumulative incidence of post-traumatic epilepsy (PTE) ranges widely, from 2% to over 50% depending on injury severity [2].

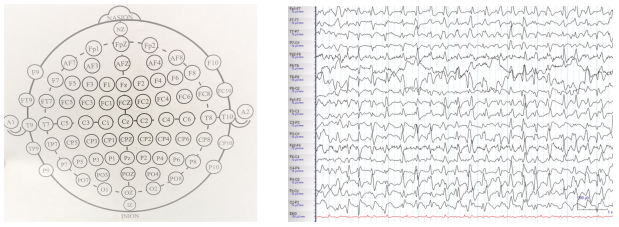
EEG can be used as a brain computer interface to read people's brain signals (Figure 1). In [3], EEG activities are transformed into sequence of topology-preserving multi-spectral images, as opposed to standard EEG analysis techniques that ignore such spatial information. A deep recurrent-convolutional network is then used to learn robust representations from the sequence of images. In [4] a convolutional neural network is used for detecting sharp waveforms called 'spikes' occurring between seizures.

A particular challenge in detecting *post-traumatic seizures*, on the other hand, is that they are *very diverse*, While seizure detection and prediction requires modeling of complex non-linear spatio-temporal dynamics in electroencephalogram (EEG) signals, most investigators consider a single late post-traumatic seizure as being sufficient for the diagnosis of post-traumatic epilepsy (PTE) [2] [5]. Since each trauma is unique, this implies that developing sufficiently rich models of seizures is a very difficult task. Recently, several machine learning based techniques, including deep neural networks, have been proposed to tackle to infer from EEG data. [6] studies the use of multi channel "envelope" EEG trends that monitor waveforms within a specific frequency range over a period of time. In [7], authors propose an ICU seizure detection algorithm using signal amplitude variation and fifth order Butterworth filter to reduce unwanted detections caused by activity in very low and high frequency ranges.

B. Contributions: $M2NN$ for Rare Event Detection

Despite the above advances, existing smart data services tend to fail when the events that are being inferred are rare, for instance when one attempts to infer very rare seizure events in highly personalized post-traumatic EEG data. In this paper, we note that multi-variate time series often carry robust localized multi-variate temporal features that could, at least in theory, help identify these events; however, the lack of sufficient data to train for these events make it impossible for neural architectures to identify and make use of these features. To tackle this challenge, we propose an LSTM-

978-1-7281-8777-8/20/$31.00 ©2020 IEEE
DOI 10.1109/SMDS49396.2020.00014



(a) EEG channels (b) temporal data

Figure 1: EEG data for seizure detection – here, the International 10-20 system [8] is used to annotate the EEG channels (C=central, T=temporal P=parietal F=frontal Fp=frontal polar O=occipital)

based neural architecture, $M2NN$, with an attention mechanism that leverages robust multivariate temporal features that are extracted *a priori* and fed into the NN as a side information. In particular, multi-variate temporal features are extracted by simultaneously considering, at multiple scales, temporal characteristics of the time series along with external knowledge, including variate relationships that are known a priori. We then show that a single layer LSTM with dual-layer attention that leverages these multi-scale, multivariate features provides significant gains in rare seizure detection tasks. We also evaluate $M2NN$ in other rare event detection tasks, such as anomaly detection in manufacturing. We further show that the proposed $M2NN$ technique is also beneficial in tackling more traditional inference problems, such as travel-time prediction, which can nevertheless be afflicted with congestions due to rare accident events.

II. Related Works

A. EEG and other Brain Data

Focal EEG onset can be predicted using convolutional networks as shown in [9]. The primary important finding for diagnosis of epilepsy is the 'spike' and 'wave' pattern. [10] utilize a subject independent convolutional recurrent attention model (CRAM) that utilizes a convolutional neural network to encode the high-level representation of EEG signals and a recurrent attention mechanism to explore the temporal dynamics of the EEG signals as well as to focus on the most discriminative temporal periods.

Aside from detection or prediction of individual seizure events from EEG signals as discussed in the Introduction, machine learning techniques have also been useful in image analysis [11] [12] in homologous brain regions on resting-state functional Magnetic Resonance Imaging(fMRI). Post-traumatic epilepsy is studied in [5] wherein an injury to brain generates seizures after weeks, months, or years. A k-NN classifier is used in [13] to predict epileptic seizures.

B. Zero-Shot Learning

Zero-shot Learning through cross modal transfer has been studied in [14] to recognize objects in images when training data are sparse; the necessary knowledge about the unseen categories comes only from unsupervised large text. Extreme data imbalance is circumvented in [15] by feature generating networks like generative adversarial networks. Zero-shot learning and knowledge transfer in music classification and speech is studied in [16] and [17], respectively. A classifier learns to recognize new image classes given only few examples from each in few-shot learning studied in [18].

C. Time Series Analysis and Rare Event Detection

Knowledge discovery in time series data has been studied from since the early 1990s. Berndt [19] proposed the dynamic time warping (DTW) technique to detect patterns in data streams or time series. DTW considers all possible warping paths that can transform one timeseries to another and picks the path that has the lowest cost. [20] proposed Discrete Wavelet Transform (DWT) and Discrete Fourier Transform (DFT) for studying temporal patterns. Lowe proposed the Scale Invariant Feature Transform (SIFT) [21] feature extraction technique in 2004.

More recently, deep learning [22] has been shown to perform well in rare inference tasks compared to other conventional methods in signal processing and other applications [23] [24]. One of the deep learning networks is the recurrent neural network and a recurrent neural network having long short term memory is usually referred to as an LSTM network [25]. LSTM has been shown to be more effective than the conventional feed-forward neural networks and recurrent neural networks in terms of sequence prediction. The LSTM network have the ability to selectively remember important information for a longer period of time. One difficulty with neural network based inference is the large number of model parameters that need to be learned from data. This is especially problematic for sparse and noisy data sets where it is difficult to learn these model parameters for accurate inference. Recent research has shown that *attention mechanisms*, that help the neural network to focus on different aspects of the data at different stages of inference, has the potential to alleviate this difficulty to some degree. The challenge with such attention mechanisms, however, is that the attention model itself needs to be constructed carefully to ensure that the model focuses on the most relevant parameters, without mistakenly ignoring parameters critical for the inference task.

III. $M2NN$: LSTM with Dual, Multi-variate, Multi-scale Attention

As described in the introduction, in this section we propose an LSTM-based neural architecture, $M2NN$, with an attention mechanism that leverages robust multivariate temporal features that are extracted *a priori* and fed into the NN as a side information. In particular, $M2NN$ leverages available metadata to extract robust localized multi-variate temporal features that help the neural architecture to focus

	Meaning
$\mathcal{V}$	Set of variates
m	Number of variates
$\mathcal{M}$	Set of metadata showing variate relationships
T	Temporal length of multi-variate time series
$\mathcal{Y}$	Data matrix describing the multi-variate time series
$\vec{q}$	Query vector in the attention model
$\vec{k}$	Key vector in the attention model
$\vec{v}$	Value vector in the attention model
$\vec{h}$	Number of attention heads
$\mathcal{S}$	Feature scales created by the RMT algorithm
$\mathcal{F}$	RMT feature set identified in the input data
l	Length of the RMT feature descriptor vector
n_t	Number of selected RMT features covering time instance t
r	Target rank for feature dimensionality reduction (also the reduced feature descriptor length)
ρ	Number of trees in the random forest (for RF-based dimensionality reduction)
k	Number of variates (for k-means based variate reduction)

Table I: Key notations

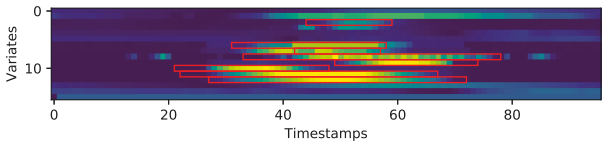


Figure 2: Localized temporal events and their scopes (red boxes)

on aspects of the multi-variate data that are potentially relevant for rare event inference. We therefore first present the underlying multivariate time series model (Table I presents the key notations used throughout this paper)

A. Meta-Data Enriched Multi-Variate Time Series Model

In this paper, we consider a metadata-enriched, multi-variate timeseries model. In particular, a multi-variate time series is defined as a triple $\mathbf{Y} = (\mathcal{V}, \mathcal{Y}, \mathcal{M})$, where

- $\mathcal{V} = \{v_1, \ldots, v_m\}$ is a set of m variates;
- $\mathcal{Y}$ is an $T \times m$ data matrix where T is the temporal length of multi-variate time series; and
- $\mathcal{M}$ is an application specific metadata graph that describes how the various variates in $\mathcal{V}$ are related to each other.

Below we describe how the data matrix and metadata are constructed for the EEG data.

B. EEG Time Series and Meta-data Graph

Brain seizures are (thankfully) rare – even in patients with post-traumatic seizure. For the data set we use in our studies, time steps with seizure labels is at most 7% of the total, with seizure-positive labels forming $< 1\%$ in many of the cases (the labels are provided by expert physicians; see Section IV-B1 for dataset details). Therefore, as we discussed in the introduction, our goal in this paper is to tackle the rare event inference challenge.

In the case of the EEG data, the multi-variate time series consists of the recorded signals from each of the sensors

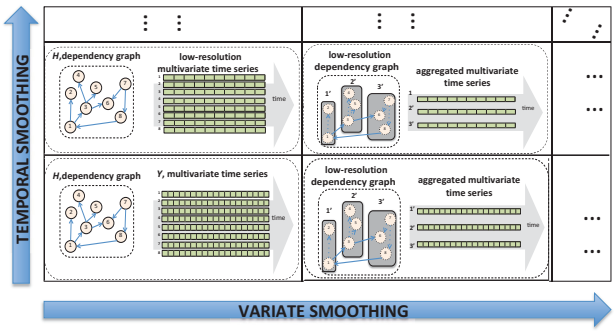


Figure 3: Scale-space generation through multi-variate smoothing

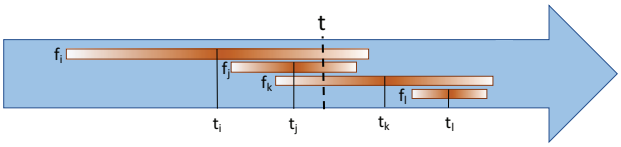


Figure 4: Three features, f_i, f_j, f_k, and f_l centered around time instances, t_i, t_j, t_k, and t_l respectively – note that the scopes of the features are defined by the Gaussian smoothing parameters (σ_i, σ_j, σ_k and σ_l corresponding to each feature); the time instance t is within the scopes of the first three of these four features, but since t is closest to t_j its contribution is highest relative to the feature f_j

shown in Figure 1 and is taken into consideration for analysis. Note that, depending on the system and configuration target being used, there can be 15 to 26 sensors used for different patients. The raw EEG data is segmented into eight second windows and power spectral density of each time window is computed by performing Fast Fourier transform on each of the individual signal segments. The result is a multi-variate EEG time series with a total of upto 520 variates. This series is accompanied with a metadata graph that describe the frequency context as described in Section IV-B1.

C. Robust Multi-Variate Temporal Features

Our key argument in this paper is that multi-variate time series carry robust localized multi-variate temporal features that could help identify critical events; however, the lack of sufficient data to train for these events make it impossible for neural architectures to identify and make use of these features. We therefore, propose that these features are identified through a process external to the NN architecture and then used as a side information to train the neural network.

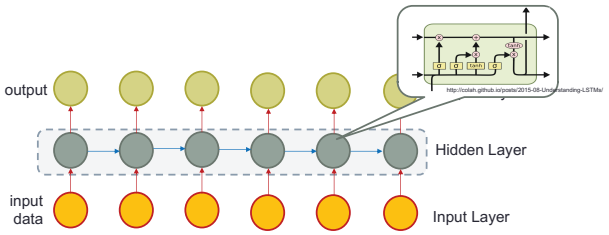
1) Key Event Detection: In this paper, we rely on the metadata supported robust multi-variate temporal (RMT) feature extraction algorithm proposed in [26]. Intuitively, a RMT feature is a fragment of a multi-variate time series that is maximally different from its immediate neighborhood,

both in time and across variate relationships specified by the metadata (Figure 2).

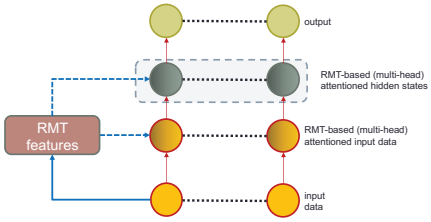
The following process is used to identify RMT features: *(Step 1): Scale-space construction:* Multi-variate temporal features of interest can be of different lengths and may cover different number of variates. In order to be able to locate such features of different sizes, the RMT features are extracted from a scale-space constructed for the given multi-variate time series through iterative smoothing (Figure 3). As shown in [26], the (Gaussian) smoothing process is guided by a metadata graph, which captures the relationship of the variates – and the scale space is obtained by smoothing both the time-series and the metadata graph. This creates different resolution versions of the input data and, thus, helps identify features with different amount of details in time and in terms of the number of variates involved. We denote the set of scales, each corresponding to a different temporal; feature size, created by this process with $\mathcal{S}$. *(Step 2): Identifying feature candidates:* Next, the process identifies candidate features of interest across multiple scales of the given multi-variate time series by searching over multiple scales and variates of the given series. Each candidate RMT feature has a *temporal-scope* (a beginning and an end in time) and a *variate-scope* (a set of variates involved in the feature). These candidate features of interest are those with the largest variations with respect to their neighbors in time, variates, and scale. *(Step 3): Eliminating poor candidates:* At the following step, those candidate features that are poorly localized (and hence are inappropriate to use as key events) are eliminated.

2) RMT Features: The above process leads to a set, $\mathcal{F}$, of RMT features, where each feature, $f_i \in \mathcal{F}$, extracted from $\mathbf{Y}$, is a pair of the form, $f_i = \langle pos_i, \vec{d_i} \rangle$: Here, $pos_i = \langle v_i, t_i, s_i \rangle$ is a VTS triple denoting the position of the feature in the scale-space of the multi-variate time series, where v_i is the index of the variate at which the feature is *centered*, t_i is the time instant around which the duration of the feature is *centered*, and $s_i \in \mathcal{S}$ is the temporal/variate smoothing scale in which the feature is identified, and $\vec{d_i}$ is a descriptor vector, representing a gradient histogram describing the temporal structure (in terms of the distribution of local gradients) corresponding to the identified key event.

Note that the above approach to identify RMT features has several advantages: First of all, the identified salient features are robust against noise and common transformations, such as temporal shifts or dropped/missing variates. Scale invariance enables the extracted salient features to be robust against variations in speed and enables multi-resolution analysis. Moreover the temporal and relationship scales at which a multi-variate feature is located give an indication about the scope (both in terms of duration and the number of variates involved) of the multi-variate feature. The value of s_i is the temporal/variate scope of the key event



(a) Outline of an LSTM network



(b) Outline of M2NN

Figure 5: $M2NN$ with dual RMT-based regional attention

corresponding to the RMT feature. In particular, since we use Gaussian smoothing to obtain the scale-space, each scale s_i has a corresponding smoothing parameter, σ_i, and the temporal scope of the feature is $6\sigma_i$ since $3\sigma_i$ from the center point t_i, in both directions, would cover approximately 99.73% of the contributions to the smoothing (Figure 4)

D. LSTM with Dual RMT-based Regional Attention

As shown in Figure 5, the proposed $M2NN$ model extends the conventional single-layer LSTM architecture, with dual RMT-based regional attention layers. In particular, a multi headed attention unit has been used in the model inspired by the transformer from [27] to operate on input data $\mathcal{Y}$, along with the RMT features extracted from this $\mathcal{Y}$.

Intuitively, the multi headed attention maps a query and set of key-value pairs to an output. The query vector $\vec{q}$ represents the inference question, the key $\vec{k}$ represents the available context information, and a value vector $\vec{v}$ specified the values on which the attention is applied.

The attention matrix is constructed through the dot product of all keys and queries, normalized via softmax, to create a mapping of elements in the key sequence corresponding to the data needed for each query. After taking softmax, the normalized attention matrix is applied on the value vector.

More specifically, given query, key, and value vectors, $\vec{q}$, $\vec{k}$, and $\vec{v}$, respectively, we have

$$Attention(\vec{q}, \vec{k}, \vec{v}) = softmax(\frac{\vec{q}\vec{k}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}})\vec{v}, \qquad (1)$$

where d_k is the length of the key vector $\vec{k}$.

1) First RMT-based Attention Layer: Intuitively, the first attention layer of $M2NN$ helps the LSTM model to focus on different parts of the data, as a function of the RMT

features corresponding to each time step. Therefore, in the first layer, the query vector, $\vec{q}$ is the RMT descriptors extracted from the input data. The key, $\vec{k}$, and value, $\vec{v}$, vectors both are set to be the input multi-variate time series.

As we see in Figure 4, a given time instance, t can be within the scopes of multiple RMT features. For example, in the figure, the time instance t is covered by three RMT features. Nevertheless, as we also see in the figure, the distance of t to the centers of these features may be different, therefore its contribution to these features may vary. To account for this, for each feature f_* that covers t in its scope, we compute a contribution value

$$contrib(t, f_*) = e^{-\frac{1}{2}(\frac{t_* - t}{\sigma_*})^2}$$

which captures the Gaussian nature of the smoothing process applied to obtain the features. Note that, since the $contrib(t, f_*)$ takes a value between 0 and 1, it can be treated also as a probability of contribution. Therefore, to identify a set of features, $\mathcal{F}_t$, that correspond to time instance t, we randomly select n_t RMT features based on the individual contribution probabilities of the features covering t. Let us denote the length of the RMT feature descriptor vector with l (in our experiments $l = 128$). In $M2NN$, for each time instance t, we stack the n_t many RMT feature descriptors corresponding to features in $\mathcal{F}_t$, constructing a data structure (a matrix, M_t) of size $n_t \times l$. This matrix M_t is then fed into $M2NN$ to support attention at time t.

2) Second RMT-based Attention Layer: The first attention layer of helps $M2NN$ to focus on different latent semantics, as a function of the RMT features. Therefore, the query vector, $\vec{q}$, is the output of the LSTM model combined with the attention weights from first layer; whereas the key, $\vec{k}$, and value vectors, $\vec{k}$, are the LSTM output sequences, each with its own descriptive vector. For the second attention layer, we are using a multihead attention unit as it allows the model to jointly attend to information from the different latent subspaces. Each attention head is of the form

$$head_{\mathrm{i}} = Attention(\vec{q}W_{Q,i}, \vec{k}W_{K,i}, \vec{v}W_{V,i}), \qquad (2)$$

where $W_{Q,i}$,$W_{K,i}$, and $W_{V,i}$ are the weights corresponding to the query, key, and value vectors, respectively. Given these an h-headed model is trained by considering

$$MultiHead(\vec{q}, \vec{k}, \vec{v}) = [head_1; \ldots; head_{\mathrm{h}}]W_O, \qquad (3)$$

where W_O captures the weights for the overall output. In our experiments, the number of heads is set to eight as in the paper [27] in the second layer.

Note that, as we see in Figure 5, at the final step, the output of the dual attention layer goes through a final activation step to complete the inference process: *sigmoid* activation is used for binary ("no-event", "event") classification, whereas for regression tasks, we have used mean squared error metric.

E. Noise Reduction in RMT Features used for Attention

The process for RMT feature extraction described in Section III-D1 leads to a descriptor vector for each RMT feature – the descriptor size[1] must be selected in a way that reflects the temporal characteristics of the time series; if a series contains many similar features, it might be more advantageous to use large descriptors that can better discriminate: these large descriptors would not only include information that describe the features, but would also describe the temporal contexts in which they are located.

As described in the previous section, in $M2NN$, we stack multiple RMT feature descriptors corresponding to each time step, before feeding these into the attention mechanism, leading to a data structure (matrix) M_t for each time instant t. While this structure can be fed as is to the attention mechanism, we note that due to its size and noise inherent in the feature extraction process, this may be not be a very effective strategy. We instead consider noise elimination and dimensionality reduction of the RMT feature descriptors before they are fed into the attention process. This is done for the entire series/data channel once, before stacking operation. In particular, in this paper, we consider principal component analysis (PCA), random forest (RF), and nonnegative matrix factorization (NMF) based latent semantic extraction techniques on the RMT feature descriptors.

1) PCA-based Reduction: In PCA based approach, the input is an $a \times b$ matrix M_t, where $a = n_t$ is the number RMT feature descriptors corresponding to time t and $b = l$ the length of the RMT feature descriptor vector. We first obtain the corresponding $a \times a$ covariance matrix C_t, which is then decomposed into $C_t = U_t \Sigma_t U_t^T$, where the $a \times c$ matrix U_t records the c eigenvectors and diagonal matrix Σ_t records the corresponding eigenvalues of the matrix C_t. Given a target rank $r \leq c$, we then decompose C_t as $\hat{C}_t = U_t^{'} \Sigma_t{}' U_t^{'T}$ where the $a \times r$ matrix $U_t^{'}$ records the r eigenvectors and diagonal matrix $\Sigma_t{}'$ records the corresponding eigenvalues of the matrix $\hat{C}_t$ with target rank r. The matrix U_t' is used as input instead of matrix M_t.

2) NMF-based Reduction : In NMF based approach, we follow a similar strategy. Given the $a \times b$ matrix M_t and a target decomposition rank r, we seek a low rank non-negative matrix decomposition $\hat{M}_t \simeq H_t V_t$ where H_t is of size $a \times r$ and V_t is of size $r \times b$. The matrix H_t is used as input instead of matrix M_t.

3) RF-based Reduction: The RF feature importance value is computed based on the impact of the feature descriptors in the overall prediction using mean decrease in impurity or GINI importance. Given the $a \times b$ matrix M_t and the labels corresponding to the RMT features, a random forest of ρ trees is built for each data channel. RF feature importance

[1]In Section III-D1, the descriptor vector length is 128.

$M2NN$ Hyperparameters(using Keras)	Value
Batch size	60
Epochs for classification	$\leq$17
Epochs for regression	15
Learning rate(Adam optimizer)	0.001
Hidden nodes of LSTM for EEG	100
Hidden nodes of LSTM for process mining	80
Hidden nodes of LSTM for traffic	15
Number of attention heads (h)	8
Number of trees (ρ) in RF	10

RMT Hyperparameters	Value
Smallest scope	$\sim$ 60 time units
Largest scope	$\sim$ 420 time units
Number of scales ($\|\mathcal{S}\|$)	12
Descriptor length (l)	128
Reduced descriptor length (r)	10

Table II: Default hyperparameters

metric is used to get the indices for the top-r features. This process is repeated for all data channels.

F. Variate Reduction in the Input Data

We also consider an additional variate reduction strategy to complement the learning process: we apply k-means clustering to the input data to reduce the number of variates from m to k. The clustering is applied on the variates in the combined time series data of all the data channels. After the clusters are obtained under the Euclidean distance model, the resulting k cluster centroids are used to construct the data matrix passed to the first layer of $M2NN$ (note that the RMT features used for attention are extracted directly from the original data matrix before the variate reduction).

IV. Experiments

In this section, we present experiment results to evaluate the effectiveness of $M2NN$ (a single layer LSTM with dual-layer regional attention that leverages these multi-scale, multi-variate features) in identifying rare events in multi-variate time series. Since our motivating smart data service is seizure detection, the first data set we use is EEG data, with rare seizure events labeled by physicians. We also evaluate $M2NN$ in other rare event detection and prediction tasks, including anomaly detection in manufacturing and travel-time prediction, which can be afflicted with congestions due to rare accident events.[2] Unless specified otherwise, the experiments are conducted using the default hyperparameter values in Table II. Linux machines(Ubuntu 18.0) with GPU 16GB RAM were used for experiments.

[2]Since the healthcare data is HIPAA protected, we make the data and code for process mining and traffic available at *https://shorturl.at/btBHN* .

A. Data Preparation

Time series in all data sets are split into three (*train*; *validation*; and *test*) regions. In order to ensure that each region has similar distribution of positive and negative labels, the time series are chunked and these chunks are shuffled in a way that preserves the rate of positive labels in each of the three regions.

B. Data Sets

Since the EEG data set cannot be released due to HIPAA protections, in order illustrate the broader applicability and reproducibility of $M2NN$, we also evaluate it in two other publicly available rare event detection tasks, anomaly detection in manufacturing and travel-time prediction under congestion.

1) EEG Seizure Dataset: The first set of experiments were performed on the EEG dataset provided by Phoenix Children's Hospital. The dataset records EEG time series and seizure events, marked by physicians, for 6 patients. Three of these six patients had intermittent seizures, whereas the other three had one or more cluster of seizure events, each.

Overall the seizure events were very rare, with positive labels being $\sim 7\%$ at the best case and $< 1\%$ in the worst case. The data set contains EEG recordings of 8 second windows upto 106,000 windows. The raw EEG data were recorded from 26 channels with a sampling rate of 256 Hz, using both referential and bipolar montage. While the sensor readings are used directly in referential montage, in bipolar montage the signals are differenced according to a spatial connectivity graph and the differenced data are used instead of the original readings. The EEG time series are segmented into eight second windows and, for each window, the corresponding power spectral density, with 20 frequency bands, is computed using Fast Fourier transform. This leads to a time series with $(26 \times 20) = 520$ variates and $(216 \times 10^6 \div (256 \times 8)) = 105944$ time steps. In these experiments, the metadata graph (represented as a matrix) is used to capture the relationships among neighboring frequencies. In other words, for each sensor channel, a 20×20 matrix is created where if two frequencies are neighbors, the corresponding pair has 1 in the metadata matrix and the matrix contains 0 otherwise. The time series were chunked into sequences of length 500 for training the LSTM. The data set is then partitioned into a training set, validation set, and test set, with 60%, 20%, and 20% of the original data each, respectively. As described earlier, the chunks were shuffled in such a way that each of these three set have similar ratios of events.

2) Process Mining Dataset: As a second rare event detection task, we considered anomaly (paper break) detection in a pulp-and-paper manufacturing data set [28]. Paper manufacturing is a continuous rolling process – when a break happens, the entire process has to be stopped, the reel has to be taken out, and the production is restarted only after

the problem is fixed. The cost of this process is very high and, therefore, a detection of a paper break event is very critical. Sensors are placed in different parts of the machine and record both the status of raw materials (e.g. amount of pulp fiber and various chemicals) and process variables (e.g. blade type, couch vacuum, and rotor speed). In particular, the available data set contains 61 variates, each with 18398 time steps – each time unit corresponds to 2 minutes. Out of these only 124 entries are marked with a breakage event.

Since we do not have *a priori* information about the relationships among the different process variables, the 61×61 metadata matrix is created to represent a clique (i.e., all entries are 1 – indicating a potential relationship).

For this dataset, we consider anomaly detection, formulated as binary classification problem. The time series were chunked into sequences of length 500 for training and the chunks were shuffled in such a way that each of these three set have similar ratios of events. Since the anomalies are exceptionally rare, to prevent this bias to negatively impact accuracy, we further leverage a sampler unit that performs undersampling of the negative labeled chunks (i.e., chunks that do not contain any anomaly). The undersampling rate of negative labels is chosen in such a way to achieve balanced positive and negative labels. As before, the data set is partitioned into a training set, validation set, and test set, with 60%, 20%, and 20% of the original data each.

3) Traffic Prediction Dataset: The two data sets described above were used for classification tasks. Our third data set focuses on a regression task for traffic prediction in road networks, to predict the travel time for a given departure time.

These experiments were performed on the dataset provided by Highways England [29]. This database provides average travel/journey time for 15-min time periods since April 2009 on all motorways and 'A' roads managed by the Highways Agency, known as the Strategic Road Network, in England. Along with average travel time within 15-min time intervals, speed and traffic flow information on motorways and 'A' roads is also given. Each day contains 96 time intervals – which means that there are 96 distinct potential departure times. Travel times in the dataset are obtained from real vehicle observations using GPS. In this section, we consider 31 day period for the month of March 2011, the length of the time series being $31 \times 96 = 2976$ time steps. The time series were split into sequences of length 4 for training. The inputs to the LSTM model are the previous travel times and departure times. There is an embedding layer used for representing time for regression problem, this is to learn traffic congestion similarities between previous timestamps and the query timestamp. In the experiments $p = 4$, we look at the previous 4 travel times (one hour history) to predict the next travel time.

We have selected the highway 'AL1165A' for the travel time study as it has congestions on the road. The majority of journey times recorded in the highway are between 100 and 200, whereas there are some significantly slower journeys, indicating a rare event slowing the travel time. There are a few outlier travel times as well greater than 400.

There are 5 variates in traffic dataset: travel time, day type, total traffic flow, average speed, and quality index. As in the case of the process mining application, since we do not have *a priori* information about the relationships among the different process variables, the 5×5 metadata matrix is created to represent a clique (i.e., all entries are 1 – indicating a potential relationship).The training set is 70%, valid set 20%, test set 10% of the input data.

C. Accuracy Metrics

For the rare event detection tasks using the EEG and manufacturing data, we assess the accuracy of different models using the F1-score metric (i.e., harmonic means of recall and precision):

$$F1Score = \frac{(2 \times Recall \times Precision)}{(Recall + Precision)} \tag{4}$$

Here, $Recall$ is the ratio of the time steps with positive labels that have been identified by the model and $Precision$ is the ratio of the time steps marked as positive by the model that is, in fact, marked also positive by domain experts.

For the traffic prediction task (which requires a regression model, rather than a classification model), we report root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Each experiment has been executed 5 times and we report averages.

D. Competitors

As competitors we consider the following techniques: **CNN (1D).** We train a multi layer CNN with batch normalization as a competitor. **LSTM, Bidirectional LSTM [30].** We train both uni-directional and bi-directional LSTM's as competitors. LSTM is a specific form of RNN [31] [32] and it works better at remembering long time steps of signal data.

E. Results

Before we present the detailed results for the EEG, process mining, and traffic data sets, we first investigate the impact of various noise removal/dimensionality reduction techniques – results are presented in Table III. As we see, PCA based dimensionality reduction of RMT features used for attention, along with k-means clustering based variate reduction provide the best F1-score; consequently, unless specified otherwise, we use PCA and variate reduction.

1) Seizure Detection in EEG Data: Before we compare $M2NN$ against other competitors, in Table IV we evaluate the impact of the proposed dual attention technique on the accuracy of the LSTM. As we see in Table IV, basic LSTM with no attention has high recall for the EEG data set, but

Model	Mean Recall	Mean Precision	Mean F1 Score
$M2NN$ with Principal Component Analysis	0.83	0.56	0.65
$M2NN$ with Non-negative Matrix Factorization	0.72	0.45	0.54
$M2NN$ with Random Forest	0.83	0.32	0.43
$M2NN$ with PCA and variate clustering ($k = 200$)	**0.96**	**0.94**	**0.95**

Table III: The impact of dimensionality and variate reduction techniques (EEG data set; the higher, the better)

Model	Mean Recall	Mean Precision	Mean F1 Score
LSTM with no attention	0.40	0.04	0.06
LSTM with single layer of RMT attention to input w/o variate clustering	0.80	0.18	0.28
LSTM with single layer of RMT attention to output w/o variate clustering	0.74	0.48	0.56
$M2NN$ w/o variate clustering	0.83	0.56	0.65
$M2NN$ with variate clustering ($k = 200$)	**0.96**	**0.94**	**0.95**

Table IV: Comparison of different attention architectures for anomaly detection in EEG data (7% rare events) – PCA reduction applied by default on RMT attention (the higher, the better)

has a rather low precision. Unfortunately, adding RMT-based attention at the input or the output alone is not effective. As observed the accuracy jumps significantly when we use $M2NN$ with dual RMT-based attention. This is because we have two types of patients: patients who have intermittent seizures (with patterns) and patients who have one or more seizure clusters in the EEG data. For the first type of patients, paying attention to the LSTM output tends to be effective; for the second type of patients, on the other hand, it is more effective to pay attention to the input data. Therefore dual attention (i.e., attention to the input as well as the output) serves well both types of patients. The results also show that the accuracy further jumps when we complement $M2NN$ with variate clustering, with F1-score reaching 0.95 on average.

In Table V, we compare the proposed $M2NN$ technique against the various competitors. As we see in this table, basic non-attentioned LSTM and bidirectional LSTM fails in this rare-event detection problem. The 1D CNN is able to learn a model that has high recall, but with low precision. The proposed LSTM based $M2NN$ technique with variate clustering, however, is able to significantly boost both recall and precision in rare event detection.

2) Anomaly Detection in Process Mining Data: Table VI compares the accuracies of various competitors in the paper breakage detection problem. As we see in this paper, all models have difficulty in addressing this rare event inference problem, having very low precision (requiring three digit precision). The table also shows that, against its competitors, $M2NN$ is able to significantly boost precision. The highest F1-score is achieved using $M2NN$ with variate clustering.

Model	Mean Recall	Mean Precision	Mean F1 Score
Bi-LSTM w/o variate clustering w/o attention	0.05	0.03	0.04
Bi-LSTM with variate clustering(k = 200) w/o attention	0.19	0.16	0.18
CNN(1D) w/o variate clustering w/o attention	**1.00**	0.12	0.21
CNN(1D) with variate clustering(k = 200) w/o attention	0.93	0.14	0.24
Bi-LSTM w/o variate clustering with dual RMT attention	0.90	0.53	0.63
CNN(1D) w/o variate clustering with dual RMT attention	0.72	0.54	0.61
$M2NN$ w/o variate clustering	0.83	0.56	0.65
$M2NN$ with variate clustering (k = 200)	0.96	**0.94**	**0.95**

Table V: Comparison of different models for EEG data (7% rare events) – PCA reduction applied by default on RMT attention (the higher, the better)

Model	Mean Recall	Mean Precision	Mean F1 Score
Bi-LSTM w/o variate clustering w/o attention	0.45	0.006	0.012
CNN(1D) w/o variate clustering w/o attention	0.55	0.006	0.012
LSTM with no attention	0.66	0.010	0.020
$M2NN$ w/o variate clustering	0.52	**0.018**	0.035
$M2NN$ with variate clustering ($k = 40$)	0.44	**0.024**	**0.050**

Table VI: Comparison of different models for anomaly detection in the process mining data (0.66 % extreme rare event) – PCA reduction applied by default on RMT attention (the higher, the better)

3) Travel Time Prediction in Traffic Data: Finally, in Table VII, we present the results for the travel time prediction problem in the traffic data set. As we discussed earlier, for this experiments, we present RMSE results – i.e., the lower the values, the better the results. As we see in the table, we obtain the lowest error values using the proposed $M2NN$ model, with dual RMT-based regional attention. In the table, we also see that the number of training epochs needed to obtain these accuracies is also lower when using RMT based attention, indicating that the side information provided by the RMT features are rich in information and support more effective learning.

V. Conclusions

Smart data solutions for post traumatic seizure detection and prediction tasks are hampered by the rareness of such events. Arguing that multi-variate EEG time series carry robust localized multi-variate temporal features that could help identify these rare seizure events, we proposed an LSTM-based $M2NN$ architecture which leverages robust multivariate temporal features that are extracted *a priori*. Experiments on EEG data (along with additional experiments on manufacturing and travel data sets) show that the proposed $M2NN$ model is highly effective in improving model accuracy for rare event detection and prediction tasks.

Model	RMSE	MAE	MAPE
Bi-LSTM with no attention(74 epochs)	9.00	6.38	4.67
LSTM with no attention(74 epochs)	9.06	6.38	4.67
LSTM with single layer attention to input (20 epochs)	9.00	6.52	4.79
LSTM with single layer attention to output(20 epochs)	9.02	6.52	4.79
$M2NN$ (15 epochs)	**8.63**	**6.24**	**4.63**

Table VII: Comparison of different architectures for the traffic data set – PCA reduction applied by default on RMT attention (the lower, the better).

ACKNOWLEDGMENTS

This work has been supported by NSF grants #1633381, #1909555, #1629888, #2026860, #1827757, DOD grant W81XWH-19-1-0514, and a DOE CYDRES grant. Experiments for the paper were conducted using NSF testbed: "Chameleon: A Large-Scale Re-configurable Experimental Environment for Cloud Research". We thank Phoenix Children's Hospital and Drs. Brian Appavu and Stephen Foldes for the (anonymized) EEG trauma data and their invaluable expertise and feedback.

REFERENCES

[1] "Epilepsy foundation, michigan." [Online]. Available: https://epilepsymichigan.org/page.php?id=358

[2] K. Ding, P. K. Gupta, and R. Diaz-Arrastia, "Epilepsy after traumatic brain injury," in *Translational research in traumatic brain injury*. CRC Press/Taylor and Francis Group, 2016.

[3] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from eeg with deep recurrent-convolutional neural networks," *arXiv preprint arXiv:1511.06448*, 2015.

[4] A. R. Johansen, J. Jin, T. Maszczyk, J. Dauwels, S. S. Cash, and M. B. Westover, "Epileptiform spike detection via convolutional neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 754–758.

[5] P. Perucca, G. Smith, C. Santana-Gomez, A. Bragin, and R. Staba, "Electrophysiological biomarkers of epileptogenicity after traumatic brain injury," *Neurobiology of disease*, vol. 123, pp. 69–74, 2019.

[6] N. S. Abend, D. Dlugos, and S. Herman, "Neonatal seizure detection using multichannel display of envelope trend," *Epilepsia*, vol. 49, no. 2, pp. 349–352, 2008.

[7] J. C. Sackellares, D.-S. Shiau, J. J. Halford, S. M. LaRoche, and K. M. Kelly, "Quantitative eeg analysis for automated detection of nonconvulsive seizures in intensive care units," *Epilepsy & Behavior*, vol. 22, pp. S69–S73, 2011.

[8] G. H. Klem, H. O. Lüders, H. Jasper, C. Elger *et al.*, "The ten-twenty electrode system of the international federation," *Electroencephalogr Clin Neurophysiol*, vol. 52, no. 3, pp. 3–6, 1999.

[9] H. Khan, L. Marcuse, M. Fields, K. Swann, and B. Yener, "Focal onset seizure prediction using convolutional networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 2109–2118, 2017.

[10] D. Zhang, L. Yao, K. Chen, and J. Monaghan, "A convolutional recurrent attention model for subject-independent eeg signal analysis," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 715–719, 2019.

[11] B. Abbasi and D. M. Goldenholz, "Machine learning applications in epilepsy," *Epilepsia*, 2019.

[12] R. Garner, M. La Rocca, G. Barisano, A. W. Toga, D. Duncan, and P. Vespa, "A machine learning model to predict seizure susceptibility from resting-state fmri connectivity," in *2019 Spring Simulation Conference (SpringSim)*. IEEE, 2019, pp. 1–11.

[13] M. Hasan, M. Ahamed, M. Ahmad, M. Rashid *et al.*, "Prediction of epileptic seizure by analysing time series eeg signal using-nn classifier," *Applied bionics and biomechanics*, vol. 2017, 2017.

[14] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Advances in neural information processing systems*, 2013, pp. 935–943.

[15] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5542–5551.

[16] J. Choi, J. Lee, J. Park, and J. Nam, "Zero-shot learning and knowledge transfer in music classification and tagging," *arXiv preprint arXiv:1906.08615*, 2019.

[17] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Advances in neural information processing systems*, 2009, pp. 1410–1418.

[18] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[19] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[20] F. Mörchen, "Time series feature extraction for data mining using dwt and dft," 2003.

[21] G. Lowe, "Sift-the scale invariant feature transform," *Int. J*, vol. 2, pp. 91–110, 2004.

[22] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," *Journal of neural engineering*, vol. 16, no. 3, p. 031001, 2019.

[23] S. Motamedi-Fakhr, M. Moshrefi-Torbati, M. Hill, C. M. Hill, and P. R. White, "Signal processing techniques applied to human sleep eeg signals—a review," *Biomedical Signal Processing and Control*, vol. 10, pp. 21–33, 2014.

[24] G. Vrbancic and V. Podgorelec, "Automatic classification of motor impairment neural disorders from eeg signals using deep convolutional neural networks," *Elektronika ir Elektrotechnika*, vol. 24, no. 4, pp. 3–7, 2018.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] S. Liu, S. R. Poccia, K. S. Candan, M. L. Sapino, and X. Wang, "Robust multi-variate temporal features of multivariate time series," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, p. 7, 2018.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[28] C. Ranjan, M. Reddy, M. Mustonen, K. Paynabar, and K. Pourak, "Dataset: rare event classification in multivariate time series," *arXiv preprint arXiv:1809.10717*, 2018.

[29] H. England, "Highways agency network journey time and traffic flow data," 2018.

[30] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[31] M. Jordan, "Serial order: a parallel distributed processing approach. technical report, june 1985-march 1986," California Univ., San Diego, La Jolla (USA). Inst. for Cognitive Science, Tech. Rep., 1986.

[32] A. Cleeremans, D. Servan-Schreiber, and J. L. McClelland, "Finite state automata and simple recurrent networks," *Neural computation*, vol. 1, no. 3, pp. 372–381, 1989.