# Explainable Multivariate Time Series Classification: A Deep Neural Network Which Learns to Attend to Important Variables As Well As Time Intervals

Tsung-Yu Hsieh
The Pennsylvania State University
University Park, PA, USA
tuh45@psu.edu

Suhang Wang
The Pennsylvania State University
University Park, PA, USA
szw494@psu.edu

Yiwei Sun
The Pennsylvania State University
University Park, PA, USA
yus162@psu.edu

Vasant Honavar
The Pennsylvania State University
University Park, PA, USA
vhonavar@psu.edu

## ABSTRACT

Many real-world applications, e.g., healthcare, present multi-variate time series prediction problems. In such settings, in addition to the predictive accuracy of the models, model transparency and explainability are paramount. We consider the problem of building explainable classifiers from multi-variate time series data. A key criterion to understand such predictive models involves elucidating and quantifying the contribution of time varying input variables to the classification. Hence, we introduce a novel, modular, convolution-based feature extraction and attention mechanism that simultaneously identifies the variables as well as time intervals which determine the classifier output. We present results of extensive experiments with several benchmark data sets that show that the proposed method outperforms the state-of-the-art baseline methods on multi-variate time series classification task. The results of our case studies demonstrate that the variables and time intervals identified by the proposed method make sense relative to available domain knowledge.

## CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; *Feature selection.*

## KEYWORDS

Multivariate time series; attentive convolution; explainability

**ACM Reference Format:**
Tsung-Yu Hsieh, Suhang Wang, Yiwei Sun, and Vasant Honavar. 2021. Explainable Multivariate Time Series Classification: A Deep Neural Network Which Learns to Attend to Important Variables As Well As Time Intervals. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3437963.3441815


*WSDM '21, March 8–12, 2021, Virtual Event, Israel*

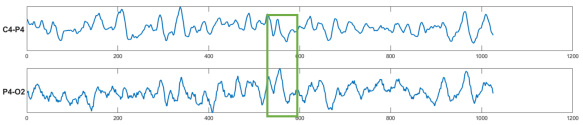
ACM ISBN 978-1-4503-8297-7/21/03…$15.00
https://doi.org/10.1145/3437963.3441815

## 1 INTRODUCTION

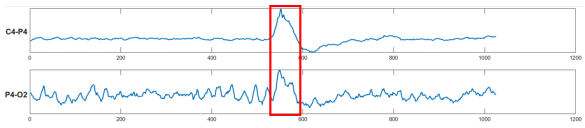
Recent advances in high throughput sensors and digital technologies for data storage and processing have resulted in the availability of complex multivariate time series (MTS) data, i.e., measurements from multiple sensors, in the simplest case, sampled at regularly spaced time points, that offer traces of complex behaviors as they unfold over time. There is much interest in effective methods for classification of MTS data [3] across a broad range of application domains including finance [58], metereology [8], graph mining [55, 60], audio representation learning [17, 54], healthcare [13, 34], human activity recognition [38, 57], among others. The impressive success of deep neural networks on a broad range of applications [31] has spurred the development of several deep neural network models for MTS classification [15]. For example, recurrent neural network and its variants LSTM and GRU are the state-of-the-art methods for modeling the complex temporal and variable relationships [10, 27].

In high-stakes applications of machine learning, the ability to explain a machine learned predictive model is a prerequisite for establishing *trust* in the model's predictions, and for gaining scientific insights that enhance our understanding of the domain [28, 39]. MTS classification models are no exception: In healthcare applications, e.g., monitoring and detection of epileptic seizures, it is important for clinicians to understand how and why an MTS classifier classifies EEG signal as indicative of onset of seizure [50]. Similarly, in human activity classification, it is important to be able to explain why an MTS classifier detects activity that may be considered suspicious or abnormal [62]. Although there has been much recent work explaining black box predictive models and their predictions [23, 28, 39], the existing methods are not directly applicable to MTS classifiers.

Developing explainable MTS data presents several unique challenges: Unlike in the case of classifiers trained on static data samples, MTS data encode *the patterns of variable progression over time*. For example, compare the brain wave signals (electroencephalogram or EEG recordings) from a healthy patient with those from a patient suffering from epileptic seizure as shown in Figure 1 [22, 50]. The two EEG recordings differ with respect to the temporal patterns




(a) Normal example




(b) Seizure example

**Figure 1: Normal and seizure brain wave signal examples.**

in the signals [37]. Because EEG measurements obtained at high temporal resolution suffer from low signal-to-noise ratio, the EEG recordings from healthy patients (see Figure 1(a)) display some of the spike-like signals that are similar to those indicative of seizure Figure 1(b). However, the temporal pattern of EEG signals over a larger time window shows clear differences between healthy and seizure activity. Thus, undue attention to local, point-wise observations, without consideration of the entire temporal pattern of activity [32] would result in failure to correctly recognize abnormal EEG recordings that are indicative of seizure. In contrast, focusing on the temporal pattern of activity over the relevant time windows as shown in Figure 1, would make it easy to distinguish the EEG recordings indicative of healthy brain activity from those that are indicative of seizure, and to explain how they differ from each other. In the case of MTS data, each variable offers different amounts of information that is relevant to the classification task. Furthermore, different variables may provide discriminative information during different time intervals. Hence, we hypothesize that MTS classifiers that can simultaneously identify not only important variables but also the time intervals during which the variables facilitate effective discrimination between different classes can not only improve the accuracy of MTS classifiers, but also enhance their explainability.

Hence, we introduce a novel, modular, convolution-based feature extraction and attention mechanism that simultaneously (i) identifies informative variables and the time intervals during which they contain informative patterns for classification; and (ii) leverages the informative variables and time intervals to perform MTS classification. Specifically, we propose Locality Aware eXplainable Convolutional ATtention network (LAXCAT), a novel MTS classifier which consists of dedicated convolution-based feature extraction network and dual attention networks. The convolution feature extraction network extracts and encodes information from a local temporal neighborhood. The dual attention networks help identify the informative variables and the time intervals in which each variable helps discriminate between classes. Working in concert, the convolution-based feature extraction network and the dual attention networks maximize predictive performance and the explainability of the MTS classifier. The major contributions of this work are as follows:

- We consider the novel problem of simultaneously selecting informative variables and time intervals with informative patterns for discrimination between the classes to optimize the accuracy and explainability of MTS classifiers;
- We describe a novel modular architecture consisting of a convolution-based feature extraction network and dual attention networks to effectively address this problem;
- We present results of extensive experiments with several benchmark data sets and show that LAXCAT outperforms the state-of-the-art baseline methods for MTS classification;
- We present results of case studies and demonstrate that the variables and time intervals identified by the proposed model are in line with the available domain knowledge.

The rest of the paper is organized as follows. Section 2 reviews related work; Section 3 introduces the problem definition; Section 4 describes our proposed solution; Section 5 describes our experiments and case studies; Section 6 concludes with a brief summary and discussion of some directions for further research.

## 2 RELATED WORK

**Multivariate Time Series Classification.** Multi-variate time series classification has received much attention in recent years. Such methods can be broadly grouped into two categories: distance-based methods [1] and feature-based methods [21]. Distance-based methods classify a given time series based on the label(s) of the time series in the training set that are most *similar* to it or closest to it where closeness is defined by some distance measure. Dynamic time warping (DTW) [5] is perhaps the most common distance measure for assessing the similarity between time series. DTW, combined with the nearest neighbors (NN) classifier is a very strong baseline method for MTS classification [3]. Feature-based methods extract a collection of informative features from the time series data and encode the time series using a feature vector. The simplest such encoding involves representing the sampled time series values by a vector of numerical feature values. Other examples of time series features include various statistics such as sample mean and variance, energy value from the Fourier transform coefficients, power spectrum bands [7], wavelets [43], *shapelets* [61], among others. Once time series data are encoded using finite dimensional feature vectors, the resulting data can be used to train a classifier using any standard supervised machine learning method [26]. The success of deep neural networks on a wide range of classification problems [31] has inspired much work on variants of deep neural networks for time series classification (see [15] for a review). However, as noted earlier, the black box nature of deep neural networks makes them difficult to understand. Deep neural network models for MTS classification are no exceptions.

**Explainable Models.** There has been much recent work on methods for explaining black box predictive models (reviewed in [28, 39], typically, by attributing the responsibility for the model's predictions to different input features. Such post hoc model explanation techniques include methods for visualizing the effect of the model inputs on its outputs [65, 67], methods for extracting simplified rules or feature interactions from black box models [20, 41], methods that score features according to their importance in prediction [2, 9, 36], gradient based methods that assess how changes in inputs impact the model predictions [9, 51, 52], and methods for approximating local decision surfaces in the neighborhood of the input sample via localized regression [6, 47, 49].

An alternative to post hoc analysis is explainability by design, which includes in particular, methods that identify an informative subset of features to build parsimonious, and hence, easier to understand models. Such methods can be further categorized into *global* methods which discover a single, instance agnostic subset of relevant variables, and *local* methods which discover instance-specific subsets of relevant features. Yoon et al. [63] proposed a principal component analysis-based recursive variable elimination approach to identify informative subset of variables on an fMRI classification task. Han et al. [25] use class separability to select optimal subset of variables in a MTS classification task. When the data set is heterogeneous, it may be hard to identify a single set of features that are relevant for classification over the entire data set [64]. Such a setting calls for local methods that can identify instance-specific features. One such local method uses an attention mechanism [4]. Choi et al. [11] proposed RETAIN, an explainable predictive model based on a two-level neural attention mechanism which identifies significant clinical variables and influential visits to the hospital in the context of electronic health records classification. RAIM [59] introduced a multi-channel attention mechanism guided by discrete clinical events to jointly analyze continuous monitoring data and discrete clinical events. Qin et al. [46] proposed a dual-stage attention-based encoder-decoder RNN to select the time series variables that drive the model predictions. Guo et al. [24] explored the structure of LSTM networks to learn variable-wise hidden states to understand the role of each variable in the prediction. A key limitation of the existing body of work on explaining black box neural network models for MTS classification is that they focus on either identifying a subset of relevant time series, or a subset of discrete time points. However, many practical applications of MTS classification, require identifying not only the relevant subset of the time series variables, but also the time intervals during which the variables help discriminate between the classes.

## 3 PROBLEM DEFINITION

Let $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}\}$ be a multivariate time series sequence, where $\mathbf{x}^{(t)} \in \mathbb{R}^P$ denotes the $P$ dimensional observation at time point $t$. $\mathbf{x}_i^{(t)} \in \mathbb{R}$ means the value of the $i$-th variable sampled at time point $t$. We use $\mathcal{X} = \{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_N, y_N)\}$ to denote a set of $N$ input sequences along with their true labels, where $\mathbf{X}_i$ is the $i$-th multivariate time series sequence and $y_i$ is its corresponding label. Based on the context, $t$ can be used to index either a time point or a time interval. In multivariate time series classification (MTSC), the goal is to predict the label $y$ of a MTS data $\mathbf{X}$. For example, given sequences of EEG recordings of a subject from multiple channels corresponding to different locations on the brain surface, the task is to predict whether it denotes healthy or seizure activity. As noted earlier each multivariate time series, not all the features equally inform the classification. In addition, for the important variables, only few key time intervals are typically important for discrimination between the different classes. Hence the problem of explainable MTSC is formally defined as follows

*Given an MTS training data set $\mathcal{X} = \{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_N, y_N)\}$, learn a function $f$ that can simultaneously predict the label of a MTS data, and identify the informative variables and the time intervals over which their values inform the class label.*

## 4 THE PROPOSED FRAMEWORK - LAXCAT

We proceed to describe Locality Aware eXplainable Convolutional ATtention network (LAXCAT). Figure 2 provides an overview of the LAXCAT architecture. LAXCAT consists of three components: (i) a convolutional module that extracts time-interval based features from the input multivariate time series sequence; (ii) variable attention module, which assigns weights to variables according to their importance in classification; and (iii) temporal attention module, which identifies the time intervals over which the variables identified by the variable attention module inform the classifier output. The LAXCAT architecture is designed to learn a representation of the MTS data that not only suffices for accurate prediction of the class label for each MTS data instance, but also helps explain the assigned class label in terms of the variables *and* the time intervals over which the values they assume inform the classification. We now proceed to describe each module of LAXCAT in detail.

### 4.1 Feature Extraction via Convolutional Layer

The first step is to extract useful features from the input time series. The key idea of the feature extraction module is to incorporate temporal pattern of values assumed by a time series variable as opposed to focusing only on point-wise observations. Given a multivariate time series input sequence $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}\}$, with $\mathbf{x}^{(t)} \in \mathbb{R}^P$, where $T$ is the length of the sequence and $P$ is the number of covariates, we adopt convolutional layer to automatically extract features from the time series. Specifically, a 1-$d$ convolutional layer with kernel size $1 \times L$ is applied on each input variable where $L$ is the length of the time interval of interest. The kernel window slides through the temporal domain with overlap. The convolutional weight is shared along the temporal domain and each input variable has its own dedicated feature extraction convolutional layer. In our model, we adopt a convolutional layer with $J$ filters so that a $J$-dimensional feature vector is extracted for each variable from each time interval. The convolutional layer encodes multivariate input sequence by:

$$\left\{\mathbf{c}_{i,t}\right\} = CNN_i\left(\mathbf{x}_i\right), \quad i = 1, \ldots, P \tag{1}$$

where $\mathbf{c}_{i,t} \in \mathbb{R}^J$ is the feature vector for $\mathbf{x}_i$ extracted from the $t$-th time interval of interest, $t = 1, \ldots, l$. Number of intervals, $l$, depends on the convolution kernel length $L$ and convolution stride size.

The convolution-based feature extraction yields features that incorporate the temporal pattern of values assumed by the input variables within a local context (determined by the convolution window). The attention mechanism applied to such features measures the importance of the targeted time interval, as opposed to specific time points. Thus, the convolutional layers can learn to adapt to the dynamics of each input time series variable while ensuring that the attention scores are attached to the corresponding input variables. The multiple filters attend to different aspects of the signal and jointly extract a rich feature vector that encodes the relevant information from the time series in the time interval of interest. Note that for each variable, the convolution computation on each time interval can be carried out in parallel, as opposed to the sequential processing in canonical RNN models. Furthermore, the number of effective time points is significantly reduced by considering intervals as opposed to discrete time points. This also reduces the computational complexity for downstream attention

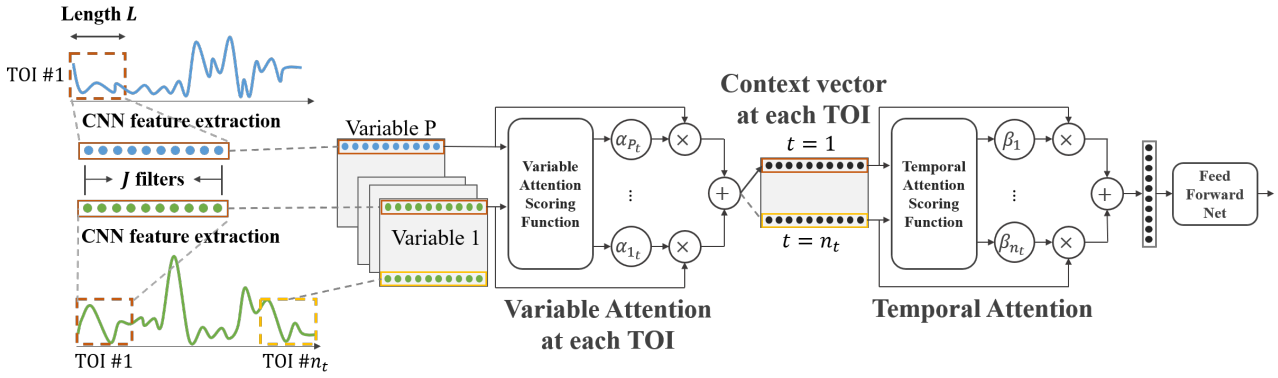



**Figure 2: Proposed LAXCAT model framework. The framework is comprised of three major components: CNN feature extraction module and two attention modules. The CNN layer extracts informative features within each time interval of interest (TOI). The two attention modules work together to identify informative variables and key TOIs.**

mechanism. While we limit ourselves to the simple convolution structure described above, the LAXCAT architecture can accommodate more sophisticated e.g., dilated [42] convolution structures for more flexible feature extraction from MTS data.

The feature extraction module accepts an input time series $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ and produces a sequence of feature matrices $\{\mathbf{C}_1, \ldots, \mathbf{C}_l\}$, where $\mathbf{C}_t \in \mathbb{R}^{P \times J}$. Each row in $\mathbf{C}_t$ stores the feature vector specific to each variable within time interval $t$ in the input sequence, i.e., $\mathbf{C}_t = [\mathbf{c}_{1,t}, \ldots, \mathbf{c}_{P,t}]^T$. The variable attention module (see below) considers the feature matrices at each time interval so as to obtain a local context embedding vector $\mathbf{h}_t, t = 1, \ldots, l$, for each interval. The temporal attention module construct the summary embedding $\mathbf{z}$, which is used to encode the MTS data for classification. In the model, temporal attention measures the contribution of each time interval to the embedding whereas variable attention controls the extent to which each variable is important within each interval. We proceed to discuss the detail of the two attention modules.

## 4.2 Variable Attention Module

The variable attention module evaluates variable attention and constructs local context embedding. Specifically, the local context embedding is an aggregation of the feature vectors weighted by their relative importance measures within the specific time interval. The context vector $\mathbf{h}_t \in \mathbb{R}^J$ for the $t$-th time interval is obtained by

$$\mathbf{h}_t = \sum_{i=1}^{P} \alpha_{i_t} \mathbf{c}_{i,t} \tag{2}$$

where $\alpha_{i_t}$ is the attention score for $\mathbf{c}_{i,t}$ and is equivalently the attention score dedicated to variable $\mathbf{x}_i$ in the $t$-th time interval.

To precisely evaluate the importance of each variable, we use a feed forward network to learn the attention score vectors $\mathbf{a}_t = [\alpha_{1_t}, \ldots, \alpha_{P_t}], t = 1, \ldots, l$. The network can be characterized by

$$\mathbf{a}_t = softmax(\mathbf{s}_t), \ \mathbf{s}_t = \sigma_2\left(\sigma_1\left(W_1^{(V)}\mathbf{C}_t + B_1^{(V)}\right)W_2^{(V)} + B_2^{(V)}\right) \tag{3}$$

where $W_1^{(V)} \in \mathbb{R}^{1\times P}, W_2^{(V)} \in \mathbb{R}^{J\times P}, B_1^{(V)} \in \mathbb{R}^{1\times J}, B_2^{(V)} \in \mathbb{R}^{1\times P}$ are the model parameters, and $\sigma_1(\cdot), \sigma_2(\cdot)$ are non-linear activation functions such as tanh, ReLU among others. In the feature extraction stage in Sec. 4.1, each time series input variable is processed independently and the correlations among the variables have not been considered. The input data to the variable attention network is the feature matrix $\mathbf{C}_t$ which contains all feature vectors in time interval $t$. The attention network considers the multivariate correlation and distributes attention weights to each variable so as to maximize the predictive performance. Note that the local context embeddings in different intervals can be constructed independently of each other (and hence processed in parallel). In addition, the parameters of the variable attention module are shared among all intervals to ensure parsimony with respect to the model parameters.

The preceding process yields local context embeddings for each of the time intervals by considering the relative variable importance. The result is a context matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_l]^T \in \mathbb{R}^{l\times J}$ consisting of the context vector at each interval. In the next subsection, we describe how to compose the summary embedding of the MTS instance using the temporal attention mechanism.

## 4.3 Temporal Attention Module

The goal of temporal attention module is to identify key segments of signals which contain information that can discriminate between classes. The summary vector $\mathbf{z}$ is composed by aggregating the context embedding vectors weighted by their relative temporal contribution as follows:

$$\mathbf{z} = \sum_{t=1}^{l} \beta_t \mathbf{h}_t \tag{4}$$

where $\beta_t$ is the temporal attention score for the context vector $\mathbf{h}_t$ and it quantifies the contribution of the information carried in interval $t$. Similarly, the temporal attention module is instantiated by a feed-forward network. The vector of temporal attention scores $\mathbf{b} = [\beta_1, \ldots, \beta_l]$ is learned using the following procedure:

$$\mathbf{b} = softmax(\mathbf{u}), \ \mathbf{u} = \sigma_2\left(\sigma_1\left(W_1^{(T)}\mathbf{H} + B_1^{(T)}\right)W_2^{(T)} + B_2^{(T)}\right) \tag{5}$$

where $W_1^{(T)} \in \mathbb{R}^{1\times l}, W_2^{(T)} \in \mathbb{R}^{J\times l}, B_1^{(T)} \in \mathbb{R}^{1\times J}$, and $B_2^{(T)} \in \mathbb{R}^{1\times l}$ are model parameters, and $\sigma_1(\cdot), \sigma_2(\cdot)$ are non-linear activation functions. The input to the temporal attention module is the entire context matrix $\mathbf{H}$. The module takes into account the correlations among time intervals and the predictive performance of each interval to distribute attention scores.

This concludes the three modules of LAXCAT architecture. To summarize, the convolutional feature extraction module extracts a

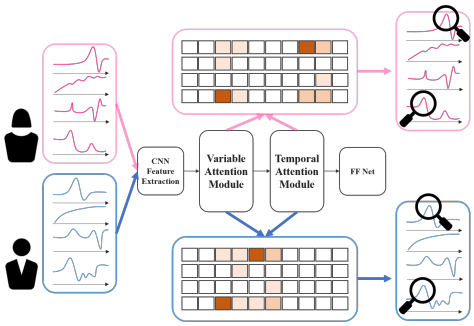



**Figure 3: Explainability of the LAXCAT model.**

rich set of features from time series data. The variable and temporal attention modules, construct an embedding of the MTS data to be classified by attending to the relevant variables and time intervals.

## 4.4 Learning LAXCAT

Given the encoding $\mathbf{z}$ which captures the important variables and time intervals of the input multivariate time series sequence, we can predict the class label of the sequence as follows:

$$y = f(\mathbf{z}; \mathbf{W}) \tag{6}$$

where $\mathbf{W}$ is the weights of $f(\cdot)$, a fully connected network[1].

Given a set of training instances $\{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_N, y_N)\}$, the parameters $\Theta$ of the variable and temporal attention modules and the classifier network can be jointly learned by optimizing the following objective function:

$$\min_{\Theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\mathbf{X}_i, y_i; \Theta) + \alpha \|\Theta\|_F^2 \tag{7}$$

where $\|\Theta\|_F^2$ is the Frobenius norm on the weights to alleviate overfitting and $\alpha$ is a scalar that control the effect of the regularization term. In this study, $\mathcal{L}$ is chosen to be the cross entropy loss function. The resulting objective function is smooth and differentiable allowing the objective function to be minimized using standard gradient back propagation update of the model parameters. We used the Adam optimizer [29] to train the model and the hyperparameters are set to their default values.

**Explainability of LAXCAT**. LAXCAT is designed to accurately classify MTS data and also facilitate instance-level explanation of the predicted class label. Given an input sequence, we can extract two attention measures from the model, namely temporal attention scores and variable attention scores. As shown in Figure 3, a summary of the importance of each variable in each interval is given by the product of the two attention scores,

$$JointAtt_{i,t} = \alpha_{i_t} \times \beta_t \tag{8}$$

for $i = 1, \ldots, P$, $t = 1, \ldots, l$. These results can then be compared against domain knowledge or used to guide further experiments.

[1] Although here we focus on classifying MTS data, the LAXCAT framework can be readily applied to forecasting and other related tasks by choosing an appropriate $f(\cdot)$ .

**Table 1: A summary of the datasets.**

| Dataset | # Var. ($P$) | # Time Points | # Classes | # Samples |
|---|---|---|---|---|
| PM2.5 w/ | 8 | 24 | 6 | 1013 |
| PM2.5 w/o | 7 | 24 | 6 | 1013 |
| Seizure | 23 | 1025 | 2 | 272 |
| Motor Task | 15 | 641 | 2 | 405 |

# 5 EXPERIMENTS AND RESULTS

We proceed to describe our experiments aimed at evaluating the performance LAXCAT in terms of the accuracy of MTS classification as well as the explainability of the resulting classifications.

## 5.1 Datasets

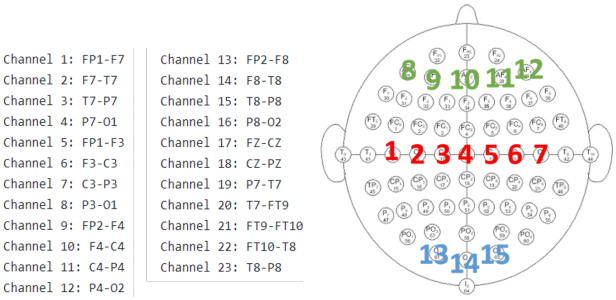
We used three publicly available real-world multivariate time series data sets and a summary of the data sets is provided in Table 1.

- **PM2.5 data set** [35] contains hourly PM2.5 value and the associated meteorological measurements in Beijing. Given the measurements on one day, the task is to predict the PM2.5 level on the next day at 8 am, during the peak commute. The PM2.5 value is categorized into six levels according to the United States Environmental Protection Agency standard, i.e., good, moderate, unhealthy for sensitive, unhealthy, very unhealthy, and hazardous. We arranged this data set into two versions, the first one contains PM2.5 recordings as one of the covariates, called **PM2.5 w/**, and the second one excludes PM2.5 recordings, denoted as **PM2.5 w/o**. Aside from PM2.5 values, the meteorological variables include dew point, temperature, pressure, wind direction, wind speed, hours of snow and hours of rain. We keep the measurements for weekdays and exclude the measurements for weekends yielding a data set of 1013 MTS instances in total.
- **Seizure data set** [22, 50] consists of electroencephalogram (EEG) recordings from pediatric subjects with intractable seizures collected at the Children's Hospital Boston. EEG signals at 23 positions, as shown in Figure 4(a), according to the international 10-20 system, were recorded at 256 samples per second with 16-bit resolution. Each instance is a four second recording containing either seizure attack period or non-seizure period.
- **Movement data set** [22, 48] consists of EEG recordings of subjects opening and closing left or right fist. EEG signals were recorded at 160 samples per second and 15 electrode locations were used in this study, covering the central-parietal, frontal and occipital regions as shown in Figure 4(b). Each instance contains 4 second recordings and the subjects were at rest state during the first two seconds and performed the fist movement during the latter two seconds. The task is to distinguish between left and right fist movement based on the 15-channel EEG recordings.

## 5.2 Baseline Methods and Evaluation Setup

We compare the classification performance of LAXCAT with representative and state-of-the-art baselines:

- **$k$NN-DTW** [19, 40] is the dynamic time warping (DTW) distance measure combined with $k$-nearest neighbor ($k$NN) classifier. DTW provides a similarity score between two time series by warping the time axes of the sequences to achieve alignment. The classification phase is carried out by $k$NN classifier.




(a) Seizure data (b) Movement data

**Figure 4: Variable and EEG location correspondence.**

- **LR** is the logistic regression classifier. For multivariate time series input, we concatenate all variables and the input to the LR model is a multivariate vector.
- **LSTM** [27] is the long short-term memory recurrent neural network. An LSTM network with one hidden layer is adopted to learn an encoding from multivariate time series data and the classification phase is carried out by a feed forward neural network.
- **DARNN** [46] is a dual attention RNN model. It uses an encoder-decoder structure where the encoder is applied to learn attentions and the decoder is adopted for prediction task.
- **IMV-LSTM** [24] is the interpretable multivariate LSTM model. It explores the structure of LSTM networks to learn variable-wise hidden states. With hidden states, a mixture attention mechanism is exploited to model the generative process of the target.

We implemented the proposed model and deep learning baseline methods with PyTorch. We used the Adam optimizer [29] to train the networks with default parameter settings and the mini-batch size is 40. The number of filters in the feature extraction step is chosen from $\{8, 16, 32\}$. The kernel size $L$ is selected from $\{2, 3, 5\}$, $\{16, 32, 64\}$, $\{16, 32, 64\}$ in PM2.5, Seizure, Motor Task respectively, and the stride size is set to 50% of kernel size. For the number of hidden nodes in the classifier feed forward network, we conduct grid search over $\{8, 16, 32\}$. The coefficient for the regularization term is chosen from $\{0.001, 0.01, 0.1\}$. In the case of the $k$NN-DTW method, $k$ is set to 1, yielding an one nearest neighbor classifier. In the case of the LSTM baseline, the number of hidden nodes is selected from $\{8, 16, 32, 64\}$. For IMV-LSTM, we implemented IMV-Tensor as it was reported to perform better [24]. The parameter selection for the baseline methods DARNN and IMV-LSTM follows the guidelines provided in the respective papers. We train the models using 70% of the samples, and 15% of the samples are for validation. The remaining 15% are used as test set. We repeat the experiment five times and report the average performance.

## 5.3 Performance of LAXCAT

**Classification Accuracy**. We compared LAXCAT with the baseline methods on MTS classification using the different benchmark data sets described above and report the results in Table 2. The results of our experiments show that deep learning-based methods outperform the other two simple baseline methods, 1NN-DTW and LR. LR mostly outperforms 1NN-DTW with the only exception on Seizure data set. Among the deep learning based methods, those equipped with an attention mechanism achieve better classification accuracy than the canonical LSTM model. Among the attention based deep neural network models, LAXCAT outperforms the other two attention based deep neural network baselines. We further note that, in the case of the two PM2.5 data sets, perhaps not surprisingly, all models consistently make better future PM2.5 value predictions when past PM2.5 value recordings are included as an input .

**Time Complexity.** We also compare the computational complexity of deep learning based methods in terms of run-time per training iteration and run-time per testing iteration. As reported in Table 3, the LSTM baseline does not include any attention mechanism to track variable and temporal importance and hence takes the least amount of run-time in each training and test iteration across all the data sets. Among the attention based deep neural network models, LAXCAT has the shortest run-time. The difference in execution time between LAXCAT and the two baselines is quite substantial in the case of Seizure and Motor Task data sets, due to the lengths of the time series in question: Each sequence in Seizure data set contains 1025 sampling points while sequences in Motor Task contain 641 sampling points. DARNN and IMV-LSTM evaluate time point-based attention, which places a greater computational burden compared to the time interval-based attention in LAXCAT . We further note in contrast to LSTM based methods which are inherently sequential, many aspects of LAXCAT are parallelizable.

## 5.4 Explaining the LAXCAT Predictions

We proceed to describe several case studies designed to evaluate the effectiveness of LAXCAT in producing useful explanations of its classification. For qualitative analysis, we report the meaningful variables and time intervals identified by the attention mechanism and compare them with domain findings in related literature. To quantitatively assess the effectiveness of the allocation of attention, we define the attention allocation measure (AAM)

$$AAM = \frac{\text{Amount of attention allocated correctly}}{\text{Total amount of attention}} \times 100\% \quad (9)$$

This measure is only applied to the cases where a solid understanding of important variable and time interval is present.

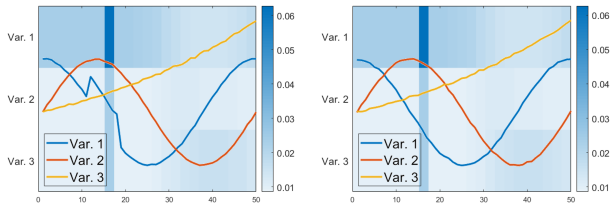
*5.4.1 Case Study I: Synthetic Data.* To thoroughly examine the attention mechanism, we constructed a synthetic data set that reflects concrete prior knowledge regarding the key variables and the time intervals that determine class labels. This synthetic data set consists of 3 time series variables, i.e. $\mathbf{x}_1^{(t)} = \cos(2\pi t) + \epsilon$, $\mathbf{x}_2^{(t)} = \sin(2\pi t) + \epsilon$, and $\mathbf{x}_3^{(t)} = \exp(t) + \epsilon$, where $\epsilon$ is Gaussian noise and $t$ takes value from a vector of 50 linearly equally spaced points between 0 and 1. To generate two classes, we randomly select half of the instances and manipulate the first variable $\mathbf{x}_1^{(t)}$ by adding a square wave signal to the raw sequence. The square wave is controlled by three random variables, the starting point, the length, and the magnitude of the square wave. We treat instances with square wave as positive and that without square wave as negative. For the synthetic data, we define correct attention allocation as the attention assigned to variable 1 within the interval of the square wave. The AAM scores on the synthetic data are reported in Table 4. From the table, we observe that LAXCAT outperforms DARNN and IMV-LSTM, suggesting that LAXCAT can better identify important variables and the relevant time intervals. We give

**Table 2: Classification results (Accuracy±std) of different algorithms on the four data sets**

| Dataset | 1NN-DTW | LR | LSTM | DARNN | IMV-LSTM | Proposed |
|---|---|---|---|---|---|---|
| PM2.5 w/ | 36.05 ± 3.24 | 38.29 ± 0.86 | 40.40 ± 1.89 | 41.19 ± 2.89 | 48.16 ± 3.30 | **50.66 ± 4.58** |
| PM2.5 w/o | 30.39 ± 3.85 | 35.92 ± 2.11 | 37.06 ± 3.04 | 38.98 ± 5.43 | 39.34 ± 4.40 | **45.53 ± 6.20** |
| Seizure | 53.66 ± 7.72 | 52.20 ± 5.88 | 70.24 ± 2.67 | 71.31 ± 2.78 | 72.19 ± 2.78 | **76.59 ± 4.08** |
| Movement | 53.44 ± 5.13 | 71.80 ± 8.31 | 75.32 ± 3.26 | 83.28 ± 1.37 | 84.09 ± 1.12 | **87.21 ± 1.37** |

**Table 3: Run-time (per iteration in seconds) comparison.**

| Dataset | PM2.5w/ | Seizure | Movement |
|---|---|---|---|
| LSTM (Train) | 0.5 | 4 | 2.8 |
| DARNN (Train) | 4.8 | 430 | 218 |
| IMV-LSTM (Train) | 3.3 | 430 | 150 |
| Ours (Train) | 1.4 | 4.5 | 3.5 |
| LSTM (Test) | 0.001 | 0.01 | 0.01 |
| DARNN (Test) | 0.08 | 24 | 13 |
| IMV-LSTM (Test) | 0.06 | 20 | 4.3 |
| Ours (Test) | 0.02 | 0.03 | 0.03 |




(a) Positive Sample (b) Negative Sample

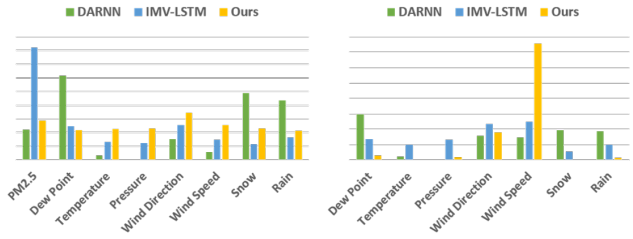
**Figure 5: Positive and negative synthetic examples are drawn in solid lines and the heat maps of the attention allocation are depicted in the background. The attention for variable 1 is located in the top row, variable 2 in the center row, and variable 3 in the bottom row. (Best viewed in color)**

**Table 4: AAM score on Synthetic data and Motor Task.**

| Dataset | DARNN | IMV-LSTM | Ours |
|---|---|---|---|
| Synthetic | 5.42% | 7.91% | 10.93% |
| Motor Task | 19.91% | 22.08% | 24.17% |

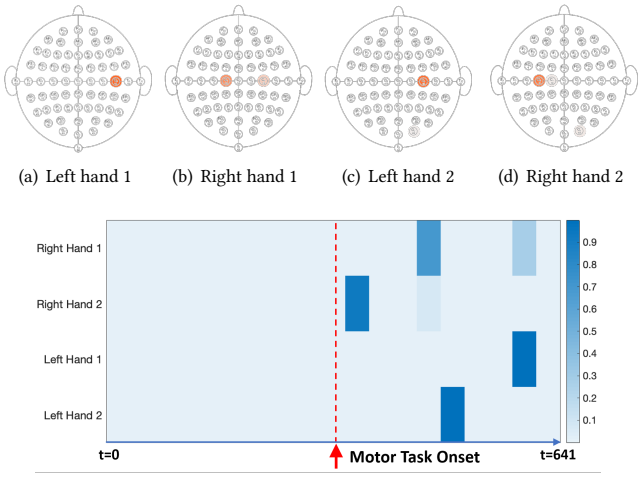
an illustration of positive instance and negative instance with the attention allocation by LAXCAT in Figure 5. We observe that the proposed model distributes most of its attention to variable 1, and specifically, in the interval that covers the location of the square wave in both positive and negative class instances.

*5.4.2 Case Study II: PM2.5.* PM2.5 value is the concentration of particles with a diameter of 2.5 micrometers or less suspended in air. Studies have found a close link between exposure to fine particles and premature death from heart and lung disease [18]. We report the attention learning results in Figure 6. Variable-wise speaking, as shown in Figure 6, when past PM2.5 recordings are available for future prediction, IMV-LSTM ranks PM2.5 as the most important variable, and the LAXCAT method ranks it as the second most important variable. DARNN consistently selects dew point, snow, and rain as important predictive variables. When PM2.5 value is not available, wind speed, wind direction and pressure are high ranked by IMV-LSTM, which is consistent with that reported in [24]. LAXCAT attends to wind direction and wind speed besides




(a) PM2.5 w/ (b) PM2.5 w/o

**Figure 6: Average attention allocation on the PM25 datasets.**



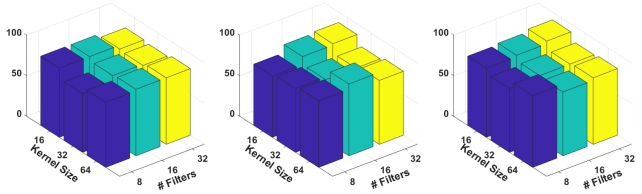

(a) Left hand 1 (b) Right hand 1 (c) Left hand 2 (d) Right hand 2

(e) Temporal attention

**Figure 7: Important channels and time intervals identified by LAXCAT on the Movement data set. The darker the color, the more attention allocated at the location.**

PM2.5 value. According to [45], wind direction and speed are critical factors that affect the amount of pollutant transport and dispersion between Beijing and surrounding areas.

*5.4.3 Case Study III: Movement Data.* On the Movement data set, we use EEG recordings to distinguish whether the subject is moving the left or right hand. A subset of attention results are reported in Figure 7. Extensive research has shown that the motor cortex is involved in the planning, control, and execution of voluntary movements [12, 44]. The motor cortex, located in the rear portion of the frontal lobe, is closest to the locations of variables 1 to 7 in our empirical analysis. In Figures 7(a)- 7(d), the heatmaps of accumulated attention in the entire time period from 2 subjects are depicted. We observe that most attention is distributed to the channels around the central region, namely the $C_1$, $C_3$, $C_4$, and $C_6$ channels. On the two left hand movement examples, i.e. Figure 7(a), 7(c), the proposed model allocates attention to the right brain. On the contrary, LAXCAT assigns most attention to the left brain during right hand




(a) 8 hidden nodes (b) 16 hidden nodes (c) 32 hidden nodes

**Figure 8: Parameter analysis on the Movement data set.**

**Table 5: Ablation study on variable (var.) attention and temporal time-interval (temp.) attention.**

| | PM2.5w/ | PM2.5w/o |
|---|---|---|
| LAXCAT | 50.66 | 45.53 |
| LAXCAT - var. attention | 40.35 | 42.50 |
| LAXCAT - temp. attention | 41.67 | 41.71 |
| LAXCAT - both attentions | 40.26 | 38.68 |

movements as shown in Figure 7(b), 7(d). This observation is in line with the current theory of contralateral brain function, which states that the brain controls the opposite side of the body. For temporal attention, as reported in Figure 7(e), LAXCAT puts most attention to the time intervals later to the motor task onset. For qualitative analysis, we define correct attention assignment as the attention distributed to channels $C_5, C_3, C_1, C_z, C_2, C_4, C_6$ in the time period after motor task onset. As reported in Table 4, LAXCAT assigns about 24% of attention to the target zone as compared to around 20% for DARNN and 22% for IMV-LSTM.

## 5.5 Ablation Study

We also conducted an ablation study to examine the relative contributions of variable attention and temporal attention in LAXCAT. Specifically, we remove the variable attention module and obtain local context embeddings by averaging the feature vectors in each time interval (the second row in Table 5). Similarly, we remove the temporal attention module and obtain summary embedding vector by averaging over the context embedding vectors (the third row in Table 5). Lastly, we remove both attention modules (the last row of Table 5). We conclude that both variable and temporal attention modules contribute to improved classification accuracy of LAXCAT.

## 5.6 Parameter Sensitivity Analysis

We investigated how the kernel size (interval length), number of filters, and number of hidden nodes in the classifier neural network affect classification accuracy. Due to space limitation, we only report the results on the Movement data set, shown in Figure 8. The results show no clear pattern as to how the numbers of filters and hidden nodes affect the predictive performance. As for kernel size, 16 and 64 consistently yield better results than 32. When we set the kernel size to 1 (corresponding to time point based temporal attention) while fixing number of filters and hidden nodes to 8, the classification accuracy falls to around 75%, which further underscores the benefits of interval-based temporal attention.

# 6 SUMMARY

We considered the problem of MTS classification, in settings where besides achieving high accuracy, it is important to identify both the key variables that drive the classification, and the time intervals during which their values provide information that helps discriminate between the classes. We introduced LAXCAT, a novel, modular architecture for explainable MTS classification. LAXCAT consists of a convolution-based feature extraction along with a variable based and a temporal interval based attention mechanism. LAXCAT is trained to optimize classification accuracy while simultaneously selecting variables and time intervals over which the pattern of values they assume drive the classifier output. We present results of extensive experiments with several benchmark data sets and show that the proposed method outperforms the state-of-the-art baseline methods for explainable MTS classification. The case studies demonstrate that the variables and time intervals identified by LAXCAT are in line with available domain knowledge. Some directions for ongoing and future research include generalizations of the LAXCAT framework to the settings with transfer learning [56, 68], multi-modal [14] or multi-view [53, 66], sparsely and irregularly observed [30, 33, 34], multi-scale [16], MTS data.

# ACKNOWLEDGMENTS

This work was funded in part by the NIH NCATS grant UL1 TR002014 and by NSF grants 2041759, 1636795, 1909702, and 1955851, the Edward Frymoyer Endowed Professorship at Pennsylvania State University and the Sudha Murty Distinguished Visiting Chair in Neurocomputing and Data Science funded by the Pratiksha Trust at the Indian Institute of Science (both held by Vasant Honavar).

# REFERENCES

[1] Amaia Abanda, Usue Mori, and Jose A Lozano. 2019. A review on distance based time series classification. *Data Mining and Knowledge Discovery* 33, 2 (2019), 378–412.

[2] Marco Ancona, Cengiz Oztireli, and Markus Gross. 2019. Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Value Approximation. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 272–281. http://proceedings.mlr.press/v97/ancona19a.html

[3] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. 2017. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery* 31, 3 (2017), 606–660.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

[5] Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series.. In *KDD workshop*, Vol. 10. Seattle, WA, 359–370.

[6] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 648–657.

[7] Peter Bloomfield. 2004. *Fourier analysis of time series: an introduction*. John Wiley & Sons.

[8] Prithwish Chakraborty, Manish Marwah, Martin Arlitt, and Naren Ramakrishnan. 2012. Fine-grained photovoltaic output prediction using a bayesian ensemble. In *AAAI*.

[9] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In *ICML*. 883–892.

[10] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv:1409.1259* (2014).

[11] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for

healthcare using reverse time attention mechanism. In *NeurIPS*. 3504–3512.
[12] BA Conway, DM Halliday, SF Farmer, U Shahani, P Maas, AI Weir, and JR Rosenberg. 1995. Synchronization between motor cortex and spinal motoneuronal pool during the performance of a maintained motor task in man. *The Journal of physiology* 489, 3 (1995), 917–924.
[13] Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 853–862.
[14] Vijay Ekambaram, Kushagra Manglik, Sumanta Mukherjee, Surya Shravan Kumar Sajja, Satyam Dwivedi, and Vikas Raykar. 2020. Attention based Multi-Modal New Product Sales Time-series Forecasting. In *KDD*. 3110–3118.
[15] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery* 33, 4 (2019), 917–963.
[16] Garrett M Fitzmaurice, Nan M Laird, and James H Ware. 2012. *Applied longitudinal analysis*. Vol. 998. John Wiley & Sons.
[17] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. 2019. Unsupervised scalable representation learning for multivariate time series. In *NeurIPS*. 4650–4661.
[18] Meredith Franklin, Petros Koutrakis, and Joel Schwartz. 2008. The role of particle composition on the association between PM2. 5 and mortality. *Epidemiology (Cambridge, Mass.)* 19, 5 (2008), 680.
[19] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.
[20] Nicholas Frosst and Geoffrey Hinton. 2017. Distilling a neural network into a soft decision tree. *arXiv:1711.09784* (2017).
[21] Ben D Fulcher and Nick S Jones. 2014. Highly comparative feature-based time-series classification. *IEEE TKDE* 26, 12 (2014), 3026–3037.
[22] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* 101, 23 (2000), e215–e220.
[23] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–42.
[24] Tian Guo, Tao Lin, and Nino Antulov-Fantulin. 2019. Exploring interpretable LSTM neural networks over multi-variable data. In *ICML*. 2494–2504.
[25] Min Han and Xiaoxin Liu. 2013. Feature selection techniques with class separability for multivariate time series. *Neurocomputing* 110 (2013), 29–34.
[26] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
[27] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[28] Aria Khademi and Vasant Honavar. 2020. A Causal Lens for Peeking into Black Box Predictive Models: Predictive Model Interpretation via Causal Attribution. *arXiv:2008.00357* (2020).
[29] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014).
[30] Thanh Le and Vasant Honavar. 2020. Dynamical Gaussian Process Latent Variable Model for Representation Learning from Longitudinal Data. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*. 183–188.
[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.
[32] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *NeurIPS*. 5244–5254.
[33] Junjie Liang, Yanting Wu, Dongkuan Xu, and Vasant Honavar. 2021. Longitudinal Deep Kernel Gaussian Process Regression. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. In press.
[34] Junjie Liang, Dongkuan Xu, Yiwei Sun, and Vasant G Honavar. 2020. LMLFM: Longitudinal Multi-Level Factorization Machine. In *AAAI*.
[35] Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang, and Song Xi Chen. 2015. Assessing Beijing's PM2. 5 pollution: severity, weather impact, APEC and winter heating. *Proc. R. Soc. A: Mathematical, Physical and Engineering Sciences* 471, 2182 (2015), 20150257.
[36] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *NeurIPS*. 4765–4774.
[37] Philippe Major and Elizabeth A Thiele. 2007. Seizures in Children: Laboratory. *Pediatrics in review* 28, 11 (2007), 405.
[38] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *CVPR*. IEEE, 4674–4683.
[39] Shane T Mueller, Robert R Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv:1902.01876* (2019).
[40] Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84.
[41] W James Murdoch, Peter J Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. *arXiv:1801.05453* (2018).
[42] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv:1609.03499* (2016).
[43] Donald B Percival and Andrew T Walden. 2000. *Wavelet methods for time series analysis*. Vol. 4. Cambridge university press.
[44] Tue Hvass Petersen, Maria Willerslev-Olsen, Bernard A Conway, and Jens Bo Nielsen. 2012. The motor cortex drives the muscles during walking in human subjects. *The Journal of physiology* 590, 10 (2012), 2443–2452.
[45] Wei-wei Pu, Xiu-juan Zhao, Xiao-ling Zhang, and Zhi-qiang Ma. 2011. Effect of meteorological factors on PM2. 5 during July to September of Beijing. *Procedia Earth and Planetary Science* 2 (2011), 272–277.
[46] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison W Cottrell. 2017. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. In *IJCAI*.
[47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *KDD*. ACM, 1135–1144.
[48] Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. 2004. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE TBME* 51, 6 (2004), 1034–1043.
[49] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE ICCV*. 618–626.
[50] Ali Hossam Shoeb. 2009. *Application of machine learning to epileptic seizure onset detection and treatment*. Ph.D. Dissertation. Massachusetts Institute of Technology.
[51] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *ICML*. JMLR. org, 3145–3153.
[52] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv:1605.01713* (2016).
[53] Yiwei Sun, Ngot Bui, Tsung-Yu Hsieh, and Vasant Honavar. 2018. Multi-view network embedding via graph factorization clustering and co-regularized multi-view agreement. In *ICDM Workshop*. IEEE, 1006–1013.
[54] Yiwei Sun and Shabnam Ghaffarzadegan. 2020. An Ontology-Aware Framework for Audio Event Classification. In *ICASSP*. IEEE, 321–325.
[55] Yiwei Sun, Suhang Wang, Tsung-Yu Hsieh, Xianfeng Tang, and Vasant Honavar. 2019. MEGAN: a generative adversarial network for multi-view network embedding. In *IJCAI*. AAAI Press, 3527–3533.
[56] Xianfeng Tang, Yandong Li, Yiwei Sun, Huaxiu Yao, Prasenjit Mitra, and Suhang Wang. 2020. Transferring Robustness for Graph Neural Network Against Poisoning Attacks. In *WSDM*. 600–608.
[57] Xianfeng Tang, Huaxiu Yao, Yiwei Sun, Charu C Aggarwal, Prasenjit Mitra, and Suhang Wang. 2020. Joint Modeling of Local and Global Temporal Dynamics for Multivariate Time Series Forecasting with Missing Values.. In *AAAI*. 5956–5963.
[58] Yue Wu, José Miguel Hernández Lobato, and Zoubin Ghahramani. 2013. Dynamic covariance models for multivariate financial time series. In *ICML*. III–558.
[59] Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. 2018. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *KDD*. 2565–2573.
[60] Xiang Xuan and Kevin Murphy. 2007. Modeling changing dependency structure in multivariate time series. In *ICML*. 1055–1062.
[61] Lexiang Ye and Eamonn Keogh. 2009. Time series shapelets: a new primitive for data mining. In *KDD*. 947–956.
[62] Jie Yin, Qiang Yang, and Jeffrey Junfeng Pan. 2008. Sensor-based abnormal human-activity detection. *IEEE TKDE* 20, 8 (2008), 1082–1090.
[63] Hyunjin Yoon and Cyrus Shahabi. 2006. Feature subset selection on multivariate time series with extremely large spatial features. In *ICDM Workshop)*. IEEE, 337–342.
[64] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. INVASE: Instance-wise variable selection using neural networks. In *ICLR*.
[65] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv:1506.06579* (2015).
[66] Ye Yuan, Guangxu Xun, Fenglong Ma, Yaqing Wang, Nan Du, Kebin Jia, Lu Su, and Aidong Zhang. 2018. Muvan: A multi-view attention network for multivariate temporal data. In *ICDM*. IEEE, 717–726.
[67] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *ECCV*. Springer, 818–833.
[68] Xi Sheryl Zhang, Fengyi Tang, Hiroko H Dodge, Jiayu Zhou, and Fei Wang. 2019. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *KDD*. 2487–2495.