

Earliness-Aware Deep Convolutional Networks for Early Time Series Classification

Wenlin Wang* Changyou Chen* Wenqi Wang† Piyush Rai‡ Lawrence Carin*

November 15, 2016

Abstract

We present Earliness-Aware Deep Convolutional Networks (EA-ConvNets), an end-to-end deep learning framework, for early classification of time series data. Unlike most existing methods for early classification of time series data, that are designed to solve this problem under the assumption of the availability of a good set of pre-defined (often hand-crafted) features, our framework can jointly perform feature learning (by learning a deep hierarchy of *shapelets* capturing the salient characteristics in each time series), along with a dynamic truncation model to help our deep feature learning architecture focus on the early parts of each time series. Consequently, our framework is able to make highly reliable early predictions, outperforming various state-of-the-art methods for early time series classification, while also being competitive when compared to the state-of-the-art time series classification algorithms that work with *fully observed* time series data. To the best of our knowledge, the proposed framework is the first to perform data-driven (deep) feature learning in the context of early classification of time series data. We perform a comprehensive set of experiments, on several benchmark data sets, which demonstrate that our method yields significantly better predictions than various state-of-the-art methods designed for early time series classification. In addition to obtaining high accuracies, our experiments also show that the learned deep shapelets based features are also highly interpretable and can help gain better understanding of the underlying characteristics of time series data.

1 Introduction

Early time series classification (ETSC), the task of predicting the label of a time series as early as possible, by only looking at an initial subsequence of the entire series, is becoming increasingly important in many time-sensitive domains. Examples include medical informatics and clinical prediction, weather reporting, financial markets, network traffic monitoring, etc [12, 34, 9, 22, 14, 23]. While the problem of classifying *fully observed* time series has been studied extensively in recent years, the earliness constraints posed by ETSC, which demands reliable *early* predictions, makes it a considerably more challenging problem than standard time series classification. An ETSC system must ensure that it not only cuts down on the prediction time, but also makes *reliable* predictions based on very few initial timestamps. These two goals are often conflicting with each other.

In order to ensure reliable predictions from an ETSC system, it is desirable to have methods that can extract highly discriminative features, even from very limited observations in the time series. Unfortunately, most existing methods for this problem are usually based on ad-hoc ways of defining subsequences or hand-crafted patterns [38], mapping them to fixed sized vectors, and subsequently applying a classification method (e.g., nearest neighbors or support vector machine). One of the key limitations of such approaches is that the extracted

*Department of Electrical & Computer Engineering, Duke University, USA

†Department of Industrial Engineering, Purdue University, USA

‡Department of Computer Science & Engineering, IIT Kanpur, India

features can usually capture only label-based local patterns, and may have difficulties in generalizing to more complex time series data sets. Therefore, the ability to automatically learn highly discriminative features from (limited-length) time series data can be very important in these problems. Motivated by this, we present a deep learning framework for learning highly discriminative *and* interpretable features in the context of early time series classification.

Shapelets [38] based representations have recently emerged as an effective way to model time series data in problems such as time series classification. Shapelets are subsequences in time series with good discriminative power. Shapelets also provide good interpretability by capturing the salient characteristics in the time series data. However, most of the existing methods perform shapelet extraction using time-consuming and ad-hoc ways. e.g., by trying out all possible prefix subsequences and evaluating them based on their ability to predict the target variable of interest. Some recent work has tried learning shapelets from data [15] but the learned shapelets are usually not ideal for *early* time series classification.

In this paper, we present a deep feature learning and classification framework based on an *earliness-aware* deep convolutional neural network architecture. Our framework, henceforth referred to as Earliness-Aware Deep Convolutional Networks (EA-ConvNets), provides, to the best of our knowledge, the first deep, nonlinear feature learning based method for fast shapelet discovery, and integrates it with a nonlinear classification model to perform end-to-end deep learning for early classification of time-series data. Our deep learning framework can learn highly discriminative shapelets that are also highly interpretable, a property that is especially desirable in many ETSC problems, such as applications in medical and health informatics. As compared to the state-of-the-art methods for that can learn interpretable features for ETSC problems [36], the deep shapelet based representations learned using our framework, in a purely data-driven manner, can capture patterns in the time series at multiple levels of granularities.

Our framework is flexible to dynamic lengths of the input time series, and can make predictions at any time based on the current observations. A key step towards this goal is to train the ConvNet on *stochastic truncated* time series data. Through an extensive set of experiments, we show that EA-ConvNets outperforms several state-of-the-art models for early time series classification, and is also competitive with time series classification methods that work with full time series.

2 Convolutional Neural Nets for Time-Series

Our framework is based on a convolutional neural network (CNN) architecture for time series data. We first provide a brief review of a basic CNN for time series data [37], before describing the details of our proposed framework.

Consider a sequence representing a univariate time series $d = [d_1, d_2, \dots, d_L]$, where L is the length of the sequence. A classical CNN for such time series data works by composing alternating layers of convolution and pooling operations. The convolutional layer extracts patterns within local regions throughout the input sequence. This operation is performed by convolving a filter, $g = [g_1, g_2, \dots, g_m]$, of length m over the sequence, and computing the inner product of the filter at each location in the sequence. For each layer, let us denote the outputs of this filter applied over the entire sequence, as a feature map vector \mathbf{o} , which the i -th element written as

$$\mathbf{o}_i = \sum_{j=1}^m g(j) \cdot d(i - j + \frac{L}{2}).$$

. Intuitively, it measures the similarity between the filter to each portion of the sequence. Afterwards, a nonlinear function $f(\cdot)$ is applied element-wise to the feature map \mathbf{o} as $\mathbf{a} = f(\mathbf{o})$. The activations \mathbf{a} are then passed to a pooling layer for nonlinear dimension reduction.

The pooling layer aggregates the information within a set of small regions of \mathbf{a} , say $\{\mathbf{r}_k\}_{k=1}^K$ with \mathbf{r}_k being some element of \mathbf{a} , and produces a *pooled feature map* $\mathbf{s} = [\mathbf{s}_1, \dots, \mathbf{s}_K]$ as the output. Let us denote the aggregation function to be $\text{POOL}()$. Then for each feature map, \mathbf{o} , we have

$$\mathbf{s}_k = \text{POOL}(f(\{\mathbf{o}_i\}_{i \in \mathbf{r}_k}))$$

where \mathbf{r}_k is pooling region k in the feature map \mathbf{o} .

3 Earliness-Aware ConvNets

In this section, we formally define the early time series classification (ETSC) problem, and describe our framework, Earliness-Aware Deep Convolutional Networks (EA-ConvNets), in detail.

3.1 Problem Definition

We will denote a univariate time series (a single sequence) as $\mathbf{T} = \{t_1, t_2, \dots, t_L\}$, where t_i is a scalar measurement* of this time series at timestamp i , L is the total number of timestamps for this time series. Following this notation, we denote a collection of time series as $\mathbf{D} = \{(\mathbf{T}_i, y_i)\}_{i=1}^N$, where N is the number of time series in the dataset, $\mathbf{T}_i = \{t_i^1, t_i^2, \dots, t_i^L\}$ represents the i th time series and y_i is the associated label. The element t_i^j indicates the j th timestamp value in the i th time series. For simplicity of exposition, we will assume all the time series in \mathbf{D} to have identical timestamps, though our framework does not require this.

The task in time series classification (TSC) is to build a classifier to predict the class label y_* of a *new* time series \mathbf{T}_* , given past training data \mathbf{D} . The time series in the training set as well as in the test data are assumed to be *fully observed*.

In contrast, ETSC requires making prediction when the time series may only be *partially* observed (i.e., upto a certain timestamp $l < L$) in the training and/or the test data. In these settings, it is desirable to predict class labels as early as possible, while retaining an acceptable level of accuracy.

3.2 Earliness-Aware ConvNets

We present a deep feature learning framework for early time series classification by building a multi-scale convolutional neural network (CNN) architecture, which is coupled with a time series truncation model that allows focusing on early parts of each time series. The multi-scale CNN enables extracting features at multiple scales from the time series data. In the context of time series data, the extracted features correspond to *shapelets* [38] which are subsequences with high discriminative power. Unlike most other existing methods that extract shapelets from time series data using hand-crafted rules (which is often ad-hoc, tedious, and time-consuming), a distinguishing aspect of our framework is that the shapelets are learned in a purely data-driven manner. Although some recent methods have looked at ways of *learning* shapelets that are optimized for a given data set [15], unlike our approach which learns a deep representation of the shapelets, these methods can only learn a shallow representation, which consequently are less expressive and have less discriminative power.

We call our propose framework Earliness-Aware Deep Convolutional Networks (EA-ConvNets). The EA-ConvNets takes as input a raw time series and outputs the class label of that time series. The overall architecture of EA-ConvNets is shown in Figure 1, which consists of three stages: Early-Awareness, deep convolutional feature (shapelet) extractor, and the final classifier, each of which is described below.

*Extending to the vector-value setting is straightforward in our framework.

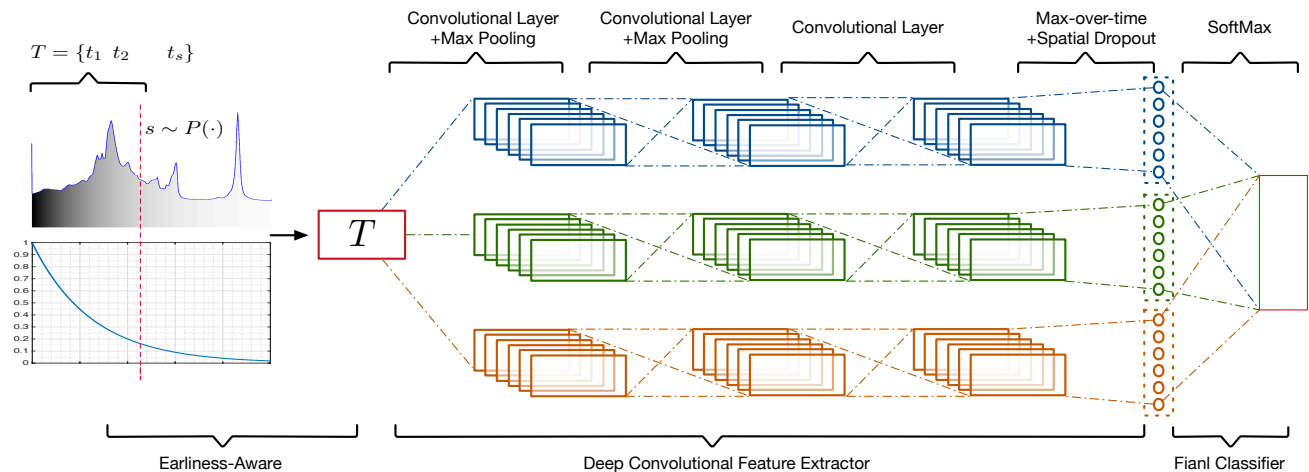


Figure 1: Overall achitecture of EA-ConvNets. In the earliness-aware part, the color of shade indicats the density of timestamps to be observed during training, and the figure below shows the exact probability distribution.

3.2.1 Early-Awareness

The Early-Awareness component lets EA-ConvNets to focus on the early stage of each time series. To accomplish Earliness-Aware, we draw inspiration from the idea of *nested dropout* [27], which was originally proposed to stochastically drop coherent nested sets of hidden units in a neural network by drawing a decay point from some decay distribution. We leverage a similar strategy to stochastically choose a truncation point for each time series, which gives rise to a new truncated version of the original time series (with the same label). To select the truncation point, a prior distribution $P(\cdot)$ on the timestamps (length) of the time series is employed. For each time series \mathbf{T}_i , we first sample an index $s \sim P(\cdot)$, generate the *stochastically* truncated time series as $\mathbf{T}_{i \downarrow s} = \{t_i^1, \dots, t_i^s\}$, which then forms the input for the subsequent convolutional neural networks based deep shaplet detector module. For choosing the truncation point, as also done in [27], we use a *geometric distribution* because of its properties of exponential decay and memorylessness. The geometric distribution is parametrized as

$$P(s; \rho) = \rho^{s-1}(1 - \rho) .$$

An illustration of applying a geometric distribution based truncation of the original full-length time series is shown in the left sub-figure in Fig 1. Specifically, we refer to the process of drawing a truncation point s from $P(\cdot)$, and truncating the time series at s as *s-truncation*; the *s-truncation* of a time series T is denoted as $T_{\downarrow s}$.

3.2.2 Deep Convolutional Feature (Shapelet) Extractor

The deep feature extraction stage of our framework is based on a deep convolutional neural network architecture, which takes as input the truncated time series obtained from the Earliness-Aware stage, extracts a hierarchy of shapelet based features for each time series, and passes these on to the final classifier.

In order to extract the deep, shapelet based features, we apply independent 1D convolutions over the (truncated) time series. Since the time series data may exhibit latent features at multiple scales, a single filter of fixed length may not be able to capture this. Therefore, to capture the *multi-scale* characteristics of the time series, we apply three channels of ConvNets in parallel. Each channel of the ConvNets has a different filter size, which is fixed within each ConvNet.

In our experiments, we found that choosing the filter sizes from $\{3\%, 5\%, 10\%\}$ of L performs well. The specific architecture is shown in Fig 1, with different filters shown in different colors. Applying each filter

on the (truncated) time series gives as output a sequence of real values, each of which can be viewed as the similarity of a local subsequence of the time series with the applied filters. A larger filter size means a wider *receptive field*, capturing a larger scale of information.

After the convolutional operation, down-sampling is applied to reduce the output dimensions. Two types of down-sampling strategies are commonly used: average-pooling and max-pooling. In our experiments, we found that max-pooling produces better results and also leads to faster training. As a result, we adopt the max-pooling without overlapping. In addition to reducing the output dimension, down-sampling also has other benefits, such as capturing scale-invariant features and reducing model complexity to avoid overfitting.

We would also like to note that our EA-ConvNets framework is adaptive to dynamic timestamps. Denoting the output feature map of the last convolutional layer to be $O(\mathbf{T}) = \{O_1(\mathbf{T}), O_2(\mathbf{T}), \dots, O_M(\mathbf{T})\}$, where M is the number of filters in the layer. Note that the feature map $O_i(\mathbf{T})$ itself is vector-valued (representing another time-series in a deep feature space). From the convolutional property, in the output feature map vector, the one with the largest value contains the most significant information. Collecting all these maximum values from the feature maps $O(\mathbf{T})$ constitutes the final layer shapelet feature representation, $F(\mathbf{T})$, for the time-series \mathbf{T} . This is essentially the *max-over-time* method described in [39], with the i th element of $F(\mathbf{T})$ defined as

$$F_i(\mathbf{T}) = \max_j O_i(\mathbf{T})[j]$$

Our feature extractor can be viewed as a hierarchical feature extractor, aiming for extracting high level abstract features, which correspond to shapelets learned in a purely data-driven fashion. To maintain the independence of each filter and avoid overfitting, we employ SpatialDropout[†] (originally proposed in [31]) on $F(\mathbf{T})$. Our experiments demonstrate the effectiveness of the regularization achieved by SpatialDropout.

3.2.3 The Final Classifier

For the final classifier, we concatenate all the shapelet based features from the outputs of the three ConvNets (one for each filter), and feed it into the final output layer, followed by a softmax transformation. The output of EA-ConvNets is the predictive distribution of each possible label for the input time series. The objective function is to minimize the cross-entropy loss defined as

$$\mathcal{L}(\mathbf{W}) = - \sum_{i=1}^N \log p_{y_i}(\mathbf{T}_i)$$

where \mathbf{W} represents all the model parameters (*e.g.*, the weights of the deep network), $p_{y_i}(\mathbf{T}_i)$ is the probability of \mathbf{T}_i having the true label y_i , which is easily calculated by forwarding the input \mathbf{T}_i through the network (line 6 in Algorithm 1). Since all the operations in the objective function are differentiable, all the model parameters can be optimized jointly through backpropagation.

The full algorithm is shown in Algorithm 1. In the algorithm, Forward(\mathbf{B}_*) in line 8 denotes the standard feedforward pass of a deep neural network, which takes a minibatch as input, and returns the loss on the corresponding minibatch. Backprop(L, y) in line 9 is the standard backpropagation pass, which takes the loss and labels of a minibatch as input, calculates the gradients of model parameters, and then update the parameters via a stochastic optimization algorithm, *e.g.*, stochastic gradient descent. For the cross-validation in line 10, we propose a method to emphasize early-stage classification accuracy, with details provided in Section 5.2 in the experiments.

[†]Applying DropOut on the convolutional layers is also possible, but it typically increases the computation burden without significantly improving the results [17], we thus did not consider this setting.

Algorithm 1 EA-ConvNets in pseudo-code.

```
1: Input:  $\mathbf{D} = \{\mathbf{T}_i, y_i\}_{i=1}^N$ ; parameter of  $P(\cdot)$  to be  $\rho$ 
2: Output: weights  $\mathbf{W}$  of EA-ConvNets
3: Initialize weights  $\mathbf{W}$  of EA-ConvNets with uniform distribution range from  $[-0.5, 0.5]$ 
4: for iteration do
5:   Sample mini-batch  $\mathbf{B} = \{\mathbf{T}_j, y_j\}_{j=1}^n$  from  $\mathbf{D}$ 
6:   Sample a set of truncation variables  $\mathbf{s} = \{s_1, \dots, s_n\}$  for the current mini-batch, where  $s_j \sim P(\rho)$ 
7:   Get the  $s$ -truncation mini-batch  $\mathbf{B}_* = \{T_{\downarrow s_j}, y_j\}_{j=1}^n$ 
8:    $L = \text{Forward}(\mathbf{B}_*)$ 
9:    $\text{Backprop}(L, \{y_j\}_{j=1}^n)$ 
10:  Evaluate on a validation set with EA-ConvNets
11:  if early-stop criteria satisfied then
12:    Break;
13:  end if
14: end for
15: Return  $\mathbf{W}$ 
```

4 Related Work

Here we review the related work in two directions relevant to this work: algorithms for early time series classification (ETSC) and deep learning methods for time series classification (TSC). ETSC was arguably first proposed in [10] and developed along several independent lines, where classifiers make predictions from partial observations of temporal data.

Some of the early work on ETSC is based on ensemble learning methods. In particular, [24, 32] use boosting to facilitate a chronological restriction of features to achieve early recognition. More recently, [1] proposed an ensemble of classifiers learned on subsequences of different lengths, and the prediction is made by the earliest individual classifier. Among other recent work, [2] learns the ensemble based early classifier by minimizing an empirical risk function and the response time required to achieve the minimum risk, which is formulized as a quadratic programming problem and solved by an iterative constraint generation algorithm.

Another prominent line of work is based on instance-based learning methods. In particular, [35] proposed ETCS (Early Classification on Time Series), which introduced the notion of Minimum Prediction Length (MPL), the earliest timestamp for each time series in the training data to find the correct nearest neighbor, and searches all MPLs in a hierarchical fashion. At test time, for a given time series, we predict its label as that of its nearest neighbor with MPL.

Inspired by the idea of shaplets [38], which are subsequences of time series that can be highly discriminative for identifying a class label, [36][13][16] extend instance-based methods by discovering a set of discriminative shapelets early in time. These methods predict a label for a test instance when a matching time series is found in the shaplets library. Further, [20] generalizes these methods to the multivariate setting with both numerical and categorical features. Another recent interesting approach is MD-MDPP+TD [19], which models multivariate time series as a Multivariate Marked Point-Process (Multi-MPP) and leverages the sequential cues as the temporal patterns to make early classification.

Another class of methods for ETCS are based on learning a classifier and a detector simultaneously. The former aims to make real-time classifications while the latter is used to decide which result can be trusted and the label will be assigned subsequently. For example, [14] combines hidden markov models (HMM) and support vector machines (SVM) for early classification and set up a threshold on the prediction probability to ensure reliability. [26] proposes a local quadratic discriminant analysis (QDA) based approach. Its reliability is guaranteed by the

threshold on the similarity of prediction on truncated and complete time series. [8] introduces a formal criterion for the detector to express the trade-off between earliness and accuracy.

Among other recent work, [33] utilizes the structure characteristic of the long short term memory (LSTM) to do early recognition during testing time. More recently, [25] proposes a probability based classifier, which learns timestamps in which the prediction accuracy for each class surpass a user defined threshold. At test time, a decision will be made no early than the learned timestamps and reliability is controlled by the differences between the top two probabilities.

For TSC, shaplets based methods have received a significant attention recently [15][29][18]. Since, shaplets can be viewed as a special case of features learned by ConvNets [7], ConvNets to TSC is a promising direction. Recently, [40] explored multi-channel CNN to deal with multivariate time series. Features are extracted through a parallel channel and then combined via an ensemble. In another recent work, [7] proposed a generalized CNN capable of extracting multi-scale characteristics in time series data and achieves state-of-the-art performance on multiple benchmark datasets. In contrast, ConvNets based models for early classification of time series have not been explored before and the EA-ConvNets framework proposed in our work is a first attempt in this direction.

5 Experiments

We evaluate the effectiveness of EA-ConvNets on 12 publicly available benchmark data sets. In the following, we give a brief description of the data sets and the experimental settings, and then conduct the following experiments

- Evaluating our model for the task of early time series classification by comparing with several state-of-the-art methods for this problem.
- Comparing EA-ConvNets with state-of-the-art algorithms for *fully observed* time series data, and show that our model is also competitive with these algorithms in terms of classification accuracies.
- Visualizing the learned features (shapelets) in each of the layers of the deep architecture.
- Assessing the sensitivity of EA-ConvNets to the hyper-parameters.
- Comparing EA-Convnets with the setting of training multiple ConvNets, where each ConvNet is trained only on the data with a particular truncated level.

5.1 Datasets

We select 12 benchmark data sets from the UCR time series archive [5], with time series having a variety of lengths. Our data sets include: ADIAC, FISH, GUN POINT, ITALYPOWERDEMAND, SYNTHETIC CONTROL, TRACE, CRICKET-X, CRICKET-Y, CRICKET-Z, TWO PATTERNS, NON-INVASIVE FETAL ECG THORAX1(NONINVTHORAX1) and NON-INVASIVE FETAL ECG THORAX2(NONINVTHORAX2). Some statistics of the data sets are listed in Table 1. Our experiments use the default training and test splits provided by UCR.

5.2 Experimental Setting

For consistency, we conduct all the experiments on all the data sets using the same architecture as shown in Fig 1. The number of filters is set to $\{48, 48, 96\}$ for each layer in each channel of the networks, and is the same

Table 1: Summary of datasets used in the evaluation.

Dataset	$\#train$	$\#test$	$ Y $	d
ADIAC	390	391	37	176
FISH	175	175	7	463
GUN POINT	50	150	2	150
ITALYPOWERDEMAND	67	1029	2	24
SYNTHETIC CONTROL	300	300	6	60
TRACE	100	100	4	275
CRICKET-X	390	390	12	300
CRICKET-Y	390	390	12	300
CRICKET-Z	390	390	12	300
TWO PATTERNS	1000	4000	4	128
NONINVTHORAX1	1800	1965	42	750
NONINVTHORAX2	1800	1965	42	750

across the three channel of the network. As a result, the final feature representation has a size $96 \times 3 = 288$. ReLU is adopted as the nonlinear activation function after each hidden layer. EA-ConvNets is implemented using Torch7 [6] and trained on NVIDIA GTX TITAN graphics cards with 6GB RAM. We optimize the model with RMSprop [30] using minibatches of size 50. We use 5-fold cross-validation for hyperparameter tuning.

Specifically, the hyperparameters of EA-ConvNets include the value of geometric distribution hyperparameter ρ , pooling factor chosen from among $\{2, 3, 5\}$, SpatialDropout rate chosen from among $\{0.4, 0.5, 0.6\}$ and L_2 weight decay chosen from $\{0, 10^{-6}, 10^{-5}, 10^{-4}\}$. We apply early stopping to avoid overfitting. We use “AUC” on the validation set as the criteria to determine when to early stop. The “AUC” (not the usual area under the ROC curve) indicates the area under a curve, plotting the trend of average truncation length vs accuracy. Early stopping is triggered when the maximum “AUC” on validation sets maintains for a number of epochs.

Also note that, unlike several existing time series classification methods, we use raw time series as input and do not use any other data augmentation, except for the stochastic truncation. The automatic feature learning capability of our deep model allows it to learn the best features for the task.

5.3 Baselines

We conduct two types of experiment: (1) early time series classification (ETSC), and (2) fully observed time series classification (TSC). For the former, we compare EA-ConvNets against the following state-of-the-art baselines:

- Early classification on time series (ECTS)[35]
- Early Distinctive Shaplet Classification (EDSC) [36]
- Reliable early classification method (RelClass) [26]
- Reliable early classification framework for time series based on class discriminativeness and reliability of predictions (ECDIRE) [25]
- Deep convolutional neural network classifier with the same network architecture as EA-ConvNets with *full observation* as the training set (EA-ConvNets-Full).

Note that our baselines also include methods that can learn *shapelets* from data (EDSC), just like EA-ConvNets (which however can learn a deep feature representation of the shapelets).

The goal of the latter experiment is to show that EA-ConvNets is able to do early classification without compromising much on the classification accuracies as compared to methods that can work with fully observed time series. For this experiment, we compare EA-ConvNets against the following state-of-the-art baselines for fully-observed time series:

- 1-Nearest Neighbor (1NN) classifier under *full observation* with Euclidean distance[11]
- 1NN with Dynamic Time Warping (DTW) [4]
- Bag-of-SFA-Symbols (BOSS) [28]
- Elastic Ensemble (PROP) [21]
- Learning Shapelets Models(LTS) [15]
- COTE [3], a weighted vote of over 35 classifiers
- Multi-scale convolutional neural networks (MCNN)[7]

5.4 Early Time Series Classification

In this section, we compare EA-ConvNets with several state-of-the-art baseline methods for early time series classification. To assess the robustness of all the methods on short time series data, for each data set, we also evaluate each method on varying lengths of the input time series (Fig 2).

For ECTS [35], we set the parameter *minimum support* from 0 to 1 with the dense interval to be 0.05. For EDSC [36], we adopt the version based on Chebyshev Inequality and set the bound for Chebyshev condition to be 2.5, 3 and 3.5, which has been reported to be efficient by the authors. For RelClass [26], we choose the values for the reliability threshold from among $\{10^{-30}, 10^{-25}, 10^{-20}, 10^{-15}, 10^{-10}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.25, 0.50, 0.75, 0.90, 0.99\}$ and for each individual test, the reliability threshold is calculated with *Gaussian Naive Bayes box* method. In contrast, for EA-ConvNets, a range of classification thresholds is set for the output of the SoftMax layer in EA-ConvNet. For a particular threshold, if the highest confidence score (the output of the SoftMax layer) surpasses the value, we accept the prediction for the current time series (thus can determine if the time series is classified correctly, and in the end calculate the average length under this threshold); otherwise we proceed until next available timestamps. The overall results are plotted in Fig 2, which show that.

1. For all fractions of the time series given as input to each method, EA-ConvNets outperforms the other baselines on all the data sets by a significantly large margin, except for NONINVTHORAX1 and NONINVTHORAX2 data sets, on which the ECDIRE baseline is comparable to EA-ConvNets. This demonstrates the robustness and generality of EA-ConvNets.
2. Even when the full time series is given, EA-ConvNets is competitive with EA-ConvNets-Full (which does not perform stochastic truncation and thus does not have the early-awareness). This indicates that the earliness aspect of EA-ConvNets does not compromise on the classification accuracies.
3. It is also worthwhile to note that the variance of EA-ConvNets is much smaller than that of EA-ConvNets-Full. Therefore, earliness-aware can also be viewed as a robust regularizer (akin to a “dropout” mechanism).

4. EA-ConvNets is particularly effective at *extremely early classification*. This is evidenced especially from the high classification accuracies even when only 20%-40% of the time series is given as the input to each of the baselines and EA-ConvNets.

Table 2: Test Errors: Comparison with (Full) Time-Series Classification Methods

Dataset	INN(ED)	DTW	BOSS	PROP	LTS	COTE	MCNN	EA-ConvNets-Full	EA-ConvNets
ADIAc	0.389	0.391	0.220	0.353	0.437	0.233	0.231	0.221 ± 0.016	0.252 ± 0.008
FISH	0.217	0.160	0.011	0.034	0.066	0.029	0.051	0.013 ± 0.005	0.037 ± 0.003
GUN POINT	0.087	0.087	0	0.007	0	0.007	0	0 ± 0	0.005 ± 0.003
ITALYPOWERDEMAND	0.045	0.045	0.053	0.039	0.030	0.036	0.030	0.023 ± 0.001	0.038 ± 0.005
SYNTHETIC CONTROL	0.120	0.017	0.030	0.010	0.007	0	0.003	0.004 ± 0.002	0.019 ± 0.003
TRACE	0.240	0.01	0	0.010	0	0.010	0	0 ± 0	0 ± 0
CRICKET-X	0.426	0.236	0.259	0.203	0.209	0.154	0.182	0.161 ± 0.003	0.183 ± 0.010
CRICKET-Y	0.356	0.197	0.208	0.156	0.249	0.167	0.154	0.135 ± 0.006	0.150 ± 0.006
CRICKET-Z	0.379	0.180	0.246	0.156	0.201	0.128	0.142	0.136 ± 0.004	0.148 ± 0.007
TWO PATTERNS	0.093	0.002	0.016	0.067	0.003	0	0.002	0 ± 0	0.003 ± 0.001
NONINVTHORAX1	0.171	0.185	0.161	0.178	0.131	0.093	0.064	0.091 ± 0.002	0.093 ± 0.004
NONINVTHORAX2	0.120	0.129	0.101	0.112	0.089	0.073	0.060	0.078 ± 0.001	0.083 ± 0.004

5.5 Comparison with (Full) Time Series Classification Methods

We also compare EA-ConvNets with several state-of-the-art baselines for full time series classification (TSC). The results are shown in Table 2. None of the previous work on has compared ETSC classifiers with TSC classifiers under full observations, due to the fact that ETSC trades off accuracies for early classification with full classification[‡]. To the best of our knowledge, we are the first to conduct such experiments, to investigate the extent of the gap between ETSC and TSC algorithms. As can be seen in Table 2, EA-ConvNets with full observations outperforms most TSC baselines on multiple datasets, and is mostly competitive with state-of-the-art models. For some datasets, though EA-ConvNets does not outperform the state-of-the-art baselines, the gaps are marginal. We would like to stress that EA-ConvNets is not designed for full time-series classification tasks, though a comparable performance could be achieved as compared to state-of-the-art TSC baselines. Therefore, EA-ConvNets is practically a reliable model for anytime time-series classification.

5.6 Qualitative Results: Deep Shapelet Discovery

An appealing aspect of our framework is its ability to learn a deep feature representation of the time series data in form of a hierarchy of shapelets (highly discriminative subsequences) automatically from data. In addition to being highly discriminative features, shapelets can also be useful in understanding/visualizing the characteristics of the underlying time series and can be useful in other decision-making tasks (e.g., in medical diagnosis). To show the interpretability of the shapelets learned by EA-ConvNets, we visualize the learned shapelets at each stage of the deep architecture, using data set TRACE as an example.

We visualize the learned filters across all the three channels of the network. To visualize the feature maps of a given time series at each stage, we only plot those in the first channel for ease of visualization. Feature maps in the other two channels show similar patterns as the first channel, and are now shown here. Fig 3 shows the results.

[‡]As a result, ETSC usually performs worse than TSC under full observations.

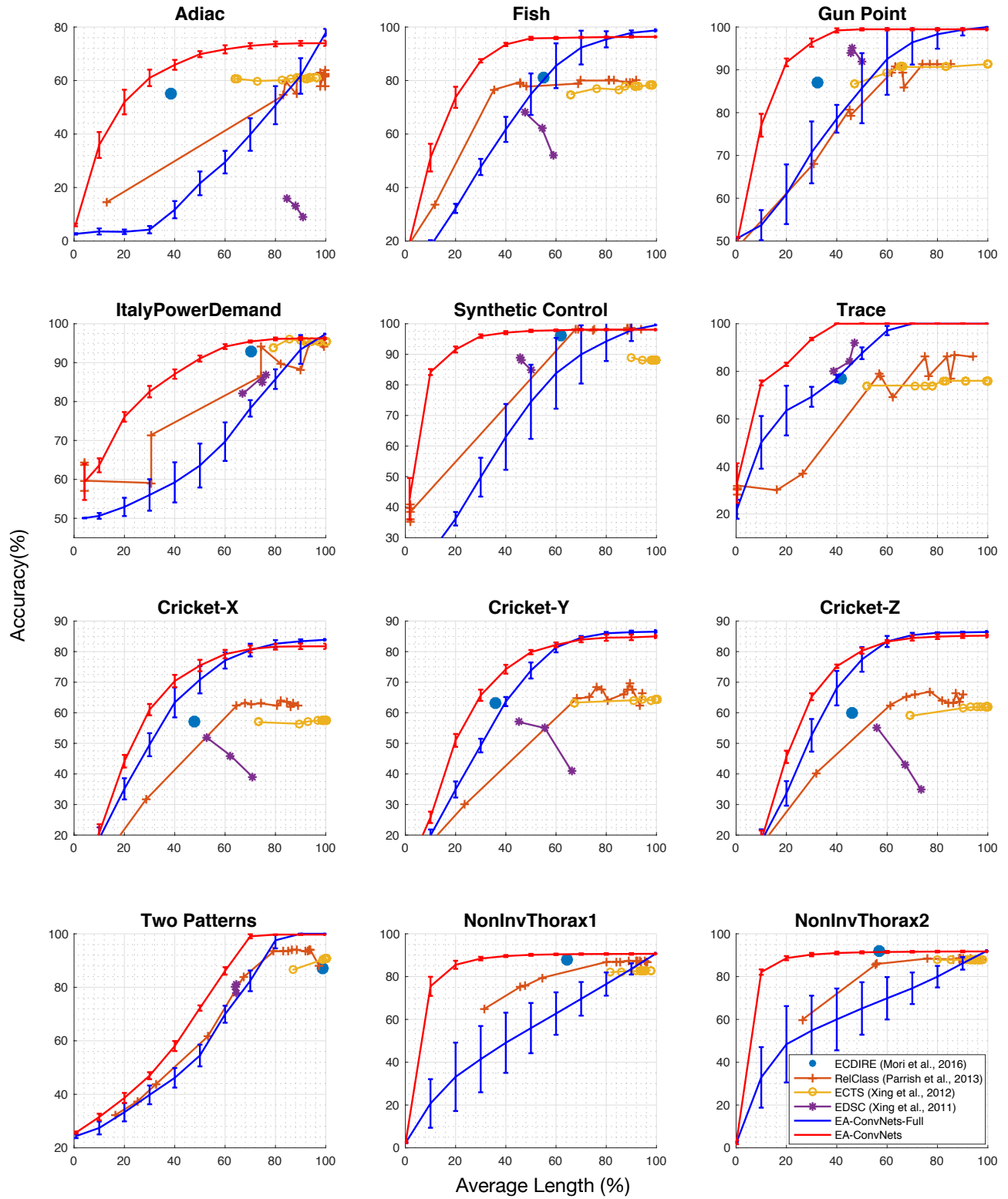


Figure 2: Test Accuracy with varying length of timestamps. The average length has been normalized to be the ratio to the full observation.

It is remarkable to note that our EA-ConvNets can automatically learn different layers of features (shapelets) simultaneously, as shown in the first layer of the networks. These features can be either low or high frequency patterns, or a mixture, depending on the data. By comparing the output feature maps of each layer, we notice that the deeper the layer goes, the more abstract features it can learn, which is typically a group of sparse and spiked features. It is also interesting to note that, for each feature map (either a sparse vector or a vector with spike in Figure 3) in the final layer, the spike (corresponding to a spike in the final feature representation) usually appears at the beginning of the feature map. This indicates that the highly discriminative features are extracted in the early stages of a time series, which provides a qualitative evidence of the effectiveness of EA-ConvNets. Finally, as shown in the figure, the final feature representation for different classes distribute discriminatively. Therefore, EA-ConvNets is a powerful model to capture robust, highly interpretable, deep features (which correspond to a *deep* hierarchy of shapelets) to make early, reliable predictions.

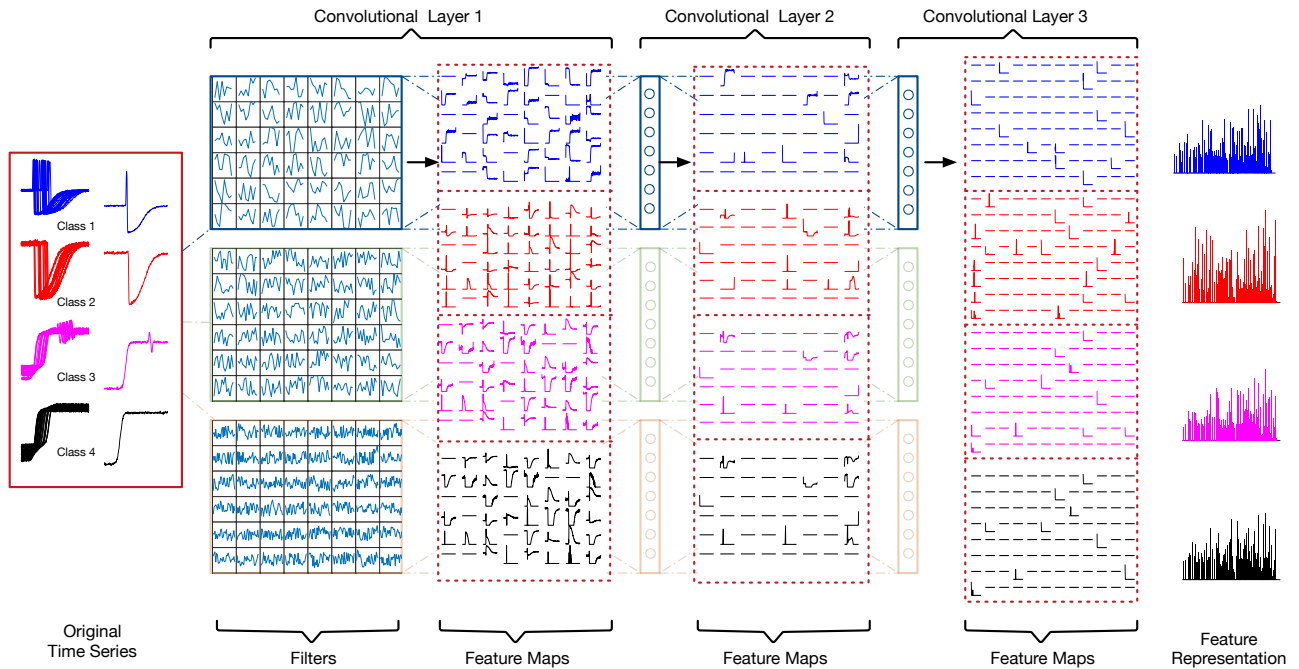


Figure 3: Visualization of dataset TRACE. The plot in the left red rectangle is the visualization of the entire datasets and one sample from each class, which is fed to the EA-ConvNet for follow up visualization. Different colors indicate different categories. The plot inside the dotted red rectangle is the output feature maps after each corresponding layers. The patterns shown in the first layer is the learned filters of EA-ConvNets. The histogram on the right is the feature representation after max-over-time.

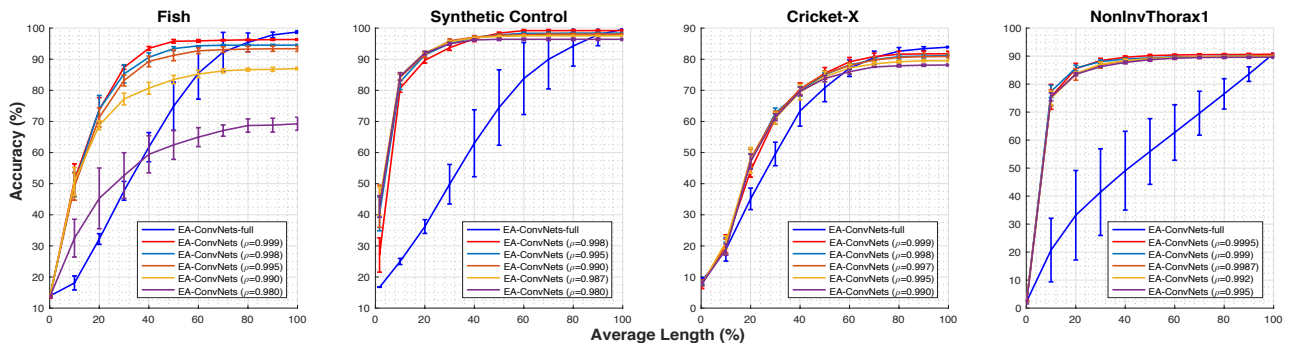


Figure 4: Results by varying the hyperparameter ρ

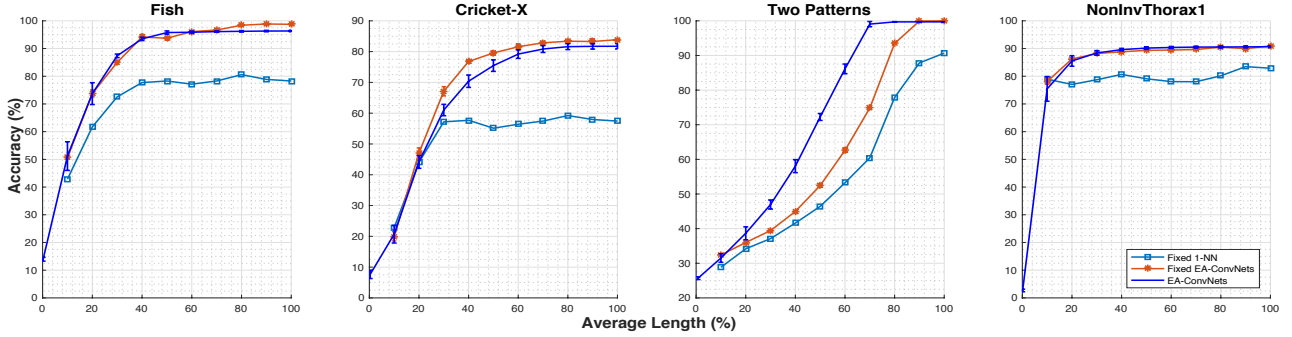


Figure 5: Comparison of EA-ConvNets with ensemble of models constructed using a variety of truncated levels

5.7 Sensitivity to hyperparameters

EA-ConvNets includes a numbers of hyperparameters. Among them, the pooling factor, SpatialDropout rate and L_2 weight decay, affect the convergence of the training algorithm. The hyperparameter, ρ , in the geometric distribution for data augmentation, not only affects the convergence of the training algorithm, but also controls the degree of earliness-aware of our model. Therefore, in this section, we mainly analyze the sensitivity of our model to ρ . We use FISH, SYNTHETIC CONTROL, CRICKET-X and NONINVTHORAX1 data sets to show the effect of this hyperparameter. The results of EA-ConvNets on these datasets for different values of ρ are shown in Fig 4. As we can see, the effect of ρ varies across different data sets. On FISH, the performance is sensitive to ρ . With an increase in ρ , EA-ConvNets results in higher accuracies across the board for various fractions of the entire time series. The story is slightly different on the SYNTHETIC CONTROL and CRICKET-X data sets, where with an increase on ρ , the curves converge to a higher accuracies under full observations but the accuracies are much lower when the fraction of observed time-series is small. This is because the larger ρ is, the higher the probability is that EA-ConvNets will observe most of the later time steps of the time series. As a result, the model would move its awareness towards the later parts of the time series. There is a trade-off between the accuracy and the earliness-aware when ρ increases. On the other hand, on NONINVTHORAX1, for a wide range of values of ρ , the accuracies are roughly the same. This can be possibly because the highly discriminative features are present in the early stage for particular data set, and are captured effectively by EA-ConvNets. To summarize, although EA-ConvNets relies on a good choice of this hyperparameter, for a sufficiently broad range of value of ρ , our model yields reasonable results. Nevertheless, we recommend using cross validation to select an appropriate ρ for different data sets.

5.8 Comparison with Ensemble of Multiple Models

To further evaluate the reliability of EA-ConvNets, we compare it with a combination (ensemble) of multiple models, each trained by truncating the original time series to a fixed length. Specifically, we compare EA-ConvNets with a group of deep learning models sharing the same architecture as EA-ConvNets but having a fixed truncation length (Fixed EA-ConvNets) and multiple k NN classifiers with fixed truncation length (Fixed 1-NN). For each truncation length, we train independent models and use them at test time to predict the label of a test time series having the same truncation length. We conduct experiments on four data sets: FISH, CRICKET-X, TWO PATTERNS and NONINVTHORAX1. Fig 5 shows the results.

It is interesting to see that on data set TWO PATTERNS, our EA-ConvNest outperforms both Fixed EA-ConvNets and Fixed 1-NN by a significant margin. The reasons may be attribute to: (1) the dynamic adaptation of EA-ConvNets. EA-ConvNets performs prediction adaptively, based on a threshold on prediction confidence; while for the ensemble of models, each model is restricted to predict based on a particular truncation length,

irrespective to specific data sets; and (2) the ability to deal with noise in the time series with EA-ConvNets. When the time series data contains noise, there is no guarantee for higher prediction accuracy with longer time series length. Our EA-ConvNets provides a proper way to deal with the noise by making early predictions with a reasonable prediction confidence. This is especially reflected in the results on NONINVTHORAX1, where EA-ConvNets performs consistently similar to Fixed EA-ConvNets, with the accuracies for both saturating after roughly 20% of the observed time-series length. Note although EA-ConvNets is a little weaker than the corresponding multiple models on the CRICKET-X dataset, the gap is very marginal.

6 Conclusion

We have introduced EA-ConvNets, a deep model based on a multi-scale convolutional neural network architecture for early classification of time series data. EA-ConvNets leverages the information at different scales and captures the interpretable features (shapelets) at a very early stages, which also makes it a robust model when each time series is inherently of a short length. Our experiments indicate that EA-ConvNets yields lower or comparable test errors given a pre-defined observed length budget at test time, while maintaining comparable classification accuracies to the state-of-the-art time series classification models that use full time series. To interpret the effectiveness of EA-ConvNets, we present the learned local features (shapelets) and visualize the nonlinear feature representations in each layers of the deep model. Our experiments clearly show that the learned feature representation is highly discriminative and can achieve effective early classification.

There are a number of interesting future directions. One direction is to extend our earliness-aware framework using other neural models such as LSTMs, which are natural tools for deal with time series data. Moreover, although in this paper we only considered EA-ConvNets for univariate time series, the framework can be extended to multivariate time series data sets. Finally, our framework is not only limited to classification tasks, and can be extended to more general (early) time series prediction tasks.

References

- [1] Shin Ando and Einoshin Suzuki. Time-sensitive classification of behavioral data. In *SDM*, pages 458–466. SIAM, 2013.
- [2] Shin Ando and Einoshin Suzuki. Minimizing response time in time series classification. *Knowledge and Information Systems*, 46(2):449–476, 2016.
- [3] Anthony Bagnall, Jason Lines, Jon Hills, and Aaron Bostrom. Time-series classification with cote: the collective of transformation-based ensembles. *Knowledge and Data Engineering, IEEE Transactions on*, 27(9):2522–2535, 2015.
- [4] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.
- [5] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/.
- [6] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376, 2011.
- [7] Zhicheng Cui, Wenlin Chen, and Yixin Chen. Multi-scale convolutional neural networks for time series classification. *arXiv preprint arXiv:1603.06995*, 2016.

- [8] Asma Dachraoui, Alexis Bondu, and Antoine Cornuéjols. Early classification of time series as a non myopic sequential decision making problem. In *Machine Learning and Knowledge Discovery in Databases*, pages 433–447. Springer, 2015.
- [9] Alberto Dainotti, Antonio Pescapé, and Carlo Sansone. *Early classification of network traffic through multi-classification*. Springer, 2011.
- [10] Grupo de Sistemas Inteligentes. Boosting interval-based literals: Variable length and early classification. *Data Mining in Time Series Databases*, page 149, 2004.
- [11] Christos Faloutsos, Mudumbai Ranganathan, and Yannis Manolopoulos. *Fast subsequence matching in time-series databases*, volume 23. ACM, 1994.
- [12] Mohamed F Ghalwash, Vladan Radosavljevic, and Zoran Obradovic. Extraction of interpretable multivariate patterns for early diagnostics. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 201–210. IEEE, 2013.
- [13] Mohamed F Ghalwash, Vladan Radosavljevic, and Zoran Obradovic. Utilizing temporal patterns for estimating uncertainty in interpretable early decision making. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 402–411. ACM, 2014.
- [14] Mohamed F Ghalwash, Dušan Ramljak, and Zoran Obradović. Early classification of multivariate time series using a hybrid hmm/svm model. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–6. IEEE, 2012.
- [15] Josif Grabocka, Nicolas Schilling, Martin Wistuba, and Lars Schmidt-Thieme. Learning time-series shapelets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 392–401. ACM, 2014.
- [16] Guoliang He, Yong Duan, Rong Peng, Xiaoyuan Jing, Tiejun Qian, and Lingling Wang. Early classification on multivariate time series. *Neurocomputing*, 149:777–787, 2015.
- [17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [18] Lu Hou James T Kwok and Jacek M Zurada. Efficient learning of timeseries shapelets. 2016.
- [19] Kang Li, Sheng Li, and Yun Fu. Early classification of ongoing observation. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 310–319. IEEE, 2014.
- [20] Yu-Feng Lin, Hsuan-Hsu Chen, Vincent S Tseng, and Jian Pei. Reliable early classification on multivariate time series with numerical and categorical attributes. In *Advances in Knowledge Discovery and Data Mining*, pages 199–211. Springer, 2015.
- [21] Jason Lines and Anthony Bagnall. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 29(3):565–592, 2015.
- [22] Yi Mao, Wenlin Chen, Yixin Chen, Chenyang Lu, Marin Kollef, and Thomas Bailey. An integrated data mining approach to real-time clinical monitoring and deterioration warning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1140–1148. ACM, 2012.
- [23] Amy McGovern, Derek H Rosendahl, Rodger A Brown, and Kelvin K Droegemeier. Identifying predictive multi-dimensional time series motifs: an application to severe weather prediction. *Data Mining and Knowledge Discovery*, 22(1-2):232–258, 2011.

- [24] Akihiro Mori, Seiichi Uchida, Ryo Kurazume, Rin-ichiro Taniguchi, Tsutomu Hasegawa, and Hiroaki Sakoe. Early recognition and prediction of gestures. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 560–563. IEEE, 2006.
- [25] Usue Mori, Alexander Mendiburu, Eamonn Keogh, and Jose A Lozano. Reliable early classification of time series based on discriminating the classes over time. *Data Mining and Knowledge Discovery*, pages 1–31, 2016.
- [26] Nathan Parrish, Hyrum S Anderson, Maya R Gupta, and Dun Yu Hsiao. Classifying with confidence from incomplete information. *The Journal of Machine Learning Research*, 14(1):3561–3589, 2013.
- [27] Oren Rippel, Michael A Gelbart, and Ryan P Adams. Learning ordered representations with nested dropout. In *ICML*, 2014.
- [28] Patrick Schäfer. The boss is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery*, 29(6):1505–1530, 2015.
- [29] Mit Shah, Josif Grabocka, Nicolas Schilling, Martin Wistuba, and Lars Schmidt-Thieme. Learning dtw-shapelets for time-series classification. In *Proceedings of the 3rd IKDD Conference on Data Science, 2016*, page 3. ACM, 2016.
- [30] T. Tieleman and G. E Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. Technical report, Coursera: Neural Networks for Machine Learning, 2012.
- [31] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–656, 2015.
- [32] Seiichi Uchida and Kazuma Amamoto. Early recognition of sequential patterns by classifier combination. In *ICPR*, pages 1–4. IEEE, 2008.
- [33] Markus Weber, Marcus Liwicki, Didier Stricker, Christopher Scholzel, and Seiichi Uchida. Lstm-based early recognition of motion patterns. In *ICPR*. IEEE, 2014.
- [34] Zhengzheng Xing, Jian Pei, and Eamonn Keogh. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, 12(1):40–48, 2010.
- [35] Zhengzheng Xing, Jian Pei, and S Yu Philip. Early classification on time series. *Knowledge and information systems*, 31(1):105–127, 2012.
- [36] Zhengzheng Xing, Jian Pei, S Yu Philip, and Ke Wang. Extracting interpretable features for early classification on time series. In *SDM*, volume 11, pages 247–258. SIAM, 2011.
- [37] Jian Bo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI), Buenos Aires, Argentina*, pages 25–31, 2015.
- [38] Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956. ACM, 2009.
- [39] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657, 2015.

- [40] Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J Leon Zhao. Time series classification using multi-channels deep convolutional neural networks. In *Web-Age Information Management*, pages 298–310. Springer, 2014.