Deep Multi-Instance Contrastive Learning with Dual Attention for Anomaly Precursor Detection

Dongkuan Xu*† Wei Cheng‡ Jingchao Ni‡ Dongsheng Luo* Masanao Natsumeda‡
Dongjin Song§ Bo Zong‡ Haifeng Chen‡ Xiang Zhang*

Abstract

Prognostics or early detection of incipient faults by leveraging the monitoring time series data in complex systems is valuable to automatic system management and predictive maintenance. However, this task is challenging. First, learning the multi-dimensional heterogeneous time series data with various anomaly types is hard. Second, the precise annotation of anomaly incipient periods is lacking. Third, the interpretable tools to diagnose the precursor symptoms are lacking. Despite some recent progresses, few of the existing approaches can jointly resolve these challenges. In this paper, we propose MCDA, a deep multi-instance contrastive learning approach with dual attention, to detect anomaly precursor. MCDA utilizes multi-instance learning to model the uncertainty of precursor period, and employs recurrent neural network with tensorized hidden states to extract precursor features encoded in temporal dynamics as well as the correlations between different pairs of time series. A dual attention mechanism on both temporal aspect and time series variables is developed to pinpoint the time period and the sensors the precursor symptoms are involved in. A contrastive loss is designed to address the issue that annotated anomalies are few. To the best of our knowledge, MCDA is the first method studying the problem of 'when' and 'where' for the anomaly precursor detection simultaneously. Extensive experiments on both synthetic and real datasets demonstrate the effectiveness of MCDA.

1 Introduction

Complex physical systems are prevalent in modern manufacturing industry. Monitoring behaviors of these large-scale systems generates massive time series data, such as the readings of sensors distributed in a power plant, and the flow intensities of system logs from the cloud computing facilities in Google, Yahoo! and Amazon [1]. The unprecedented growth of the monitoring data increases the demand for automatic and timely detection of incipient anomalies as well as precise discovery of precursor symptoms. It has been reported that 1 minute of downtime in an automotive manufacturing plant could result in as much as $20,000 cost [2]. Hence an early detection and diagnosis of system anomaly is crucial to avoid serious money waste and business loss.

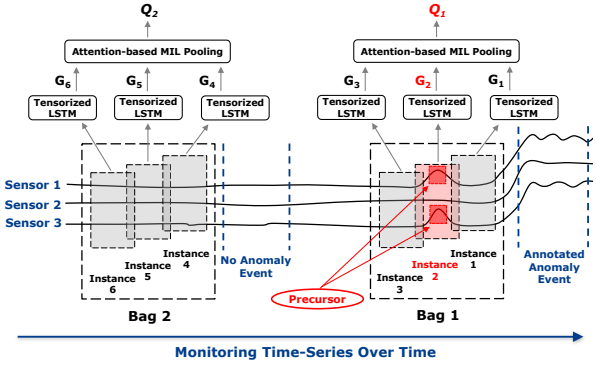


Figure 1: The framework of MCDA.

Due to its practical importance, there have been intensive interests in developing algorithms to detect time series anomalies [3–8]. However, the vital task of *anomaly precursor* detection is underexplored. An anomaly precursor represents early symptoms of an upcoming anomaly. Both anomaly and precursor are actually anomalies. The major difference is the severity from the system operators' point of view. When the precursors show up, the system usually still runs properly. In the precursor, the time series values of the involved variables change mildly. As time goes by, the minor abnormal devices will propagate their effects to more devices and trigger the observable anomaly. Anomaly precursor detection aims to answer the question that in which particular *time periods* and on which exact *sensors* the early symptoms show up. Detecting precursor is useful for early prediction of anomalies, which can effectively facilitates the circumvention of serious problems. An example is shown in Fig. 1. There are three

*Penn State University. {dux19, dul262, xzz89}@psu.edu
†Work done during an internship at NEC Labs America
‡NEC Labs America, Inc. {weicheng, jni, mnatsumeda, bzong, haifeng}@nec-labs.com
§University of Connecticut. {dongjin.song}@uconn.edu

sensors and their signals are monitored over time. An anomaly event is reported in the right-most blue dashed interval. The precursor shows up in the red dotted rectangle and Sensor 1 & 3 are involved in the precursor (as illustrated by the red boxes). Only the anomaly is annotated. It is unknown which small time series segment contains the precursor.

The detection of anomaly precursor is challenging. First, obtaining precise annotation of precursor period is infeasible in practice. Usually, only visible anomaly events can be reported by the system operators or security reporting systems [9], and they are scarce. Second, it is hard to answer which sensors are involved in the precursor symptoms, especially for the complex systems with a large amount of sensors. Finally, previous studies [10] suggest that in addition to the temporal dynamics in the raw multivariate time series, the correlations (interactions) between pairs of time series (sensors) are essential to characterize the system status. How to consider both temporal dynamics and correlations between time series for characterizing precursors is also challenging.

Some recent progresses have been made on anomaly precursor detection [11, 12]. However, none of these precursor detection methods can pinpoint both the time period and the sensors the symptoms are involved in. Usually, the anomaly precursor only occurs on some particular sensors rather than all of them. One approach that can be used to capture precursors is multi-instance learning (MIL) [13–15]. MIL assumes that a set of data instances are grouped in the forms of bags and the bag-level labels are available but the instance-level labels are not. As shown in Fig. 1, a small time series segment is considered as an instance. MIL can be utilized to detect the instances that contain the precursors by utilizing the labels of annotated anomalies. However, the MIL itself does not consider the temporal pattern of time series data.

To address the challenges, we propose a deep multi-instance contrastive learning approach with dual attention (MCDA). MCDA aims to locate and learn the representations of precursors and then uses them to detect the precursors in future time series data. To utilize the label of annotated anomalies for tracking precursors, the MIL framework is applied. Specifically, we organize the time series data in the forms of instances and bags. For each annotated anomaly, its immediately preceding bag is regarded as positive[1]; other bags are regarded as negative. For example, in Fig. 1, Bag 1 is considered as positive and Bag 2 is negative. Based on the standard assumption of MIL, the positive bag contains at least one positive instance (i.e., a segment containing precursors), while instances in a negative bag should all be negative. In this manner, we basically try to locate the instance that are close to the upcoming anomalies to model the precursors. To model the temporal behavior of time series data of each instance, an LSTM network with tensorized hidden states is developed. The time series data of an instance is fed into this LSTM network to extract the features of the instance. Our LSTM network incorporates a time-dependent correlation module to learn features encoding both temporal dynamics and the correlations between pairs of time series. Moreover, a dual attention mechanism on both temporal aspect and time series variables on top of the hidden states in the LSTM network is developed. It can pinpoint during which time instances the precursor symptoms show up and what sensors are involved. To address the issue that annotated anomalies are few, we leverage the idea of contrastive learning [16] to design a bag pair contrastive loss. Its basic idea is to make the representations of the bags from system normal period be dissimilar from the ones of annotated anomaly bags. After the model is trained, the future time series data can be fed to it for automatically learning representations of precursors, which can then be immediately used for determining whether an anomaly event will happen. The major contributions are summarized as follows.

- We study a novel problem of anomaly precursor detection, with the aim to detect both when (what time periods) and where (which sensors) the precursors are.
- We propose MCDA. MCDA combines MIL and tensorized LSTM with a time-dependent correlation module for the end-to-end learning of precursors. It is also developed with a dual attention module and a contrastive loss to produce robust and interpretable results. MCDA is the first method studying the problem of 'when' and 'where' for the anomaly precursor detection simultaneously.
- We perform extensive experiments on both real-world and synthetic datasets. The results validate the effectiveness of MCDA and its superiority over other competitors.

2 The Problem

N time series is denoted by $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T) \in \mathbb{R}^{N \times T}$, where $\mathbf{x}_t = (x_t^1, \cdots, x_t^N)^\top \in \mathbb{R}^N$ and T is the number of time steps. A time series segment is defined as an instance denoted by $\mathbf{E}_k = (\mathbf{x}_{t_k}, \cdots, \mathbf{x}_{t_k+I-1}) \in \mathbb{R}^{N \times I}$, where t_k is its starting time step and I is its time

[1] Please see discussion in the supplementary materials for dealing with the case where no precursor exists.

interval length. A bag is a set of instances denoted by $\mathbb{B} = \{\mathbf{E}_1, \cdots, \mathbf{E}_n\}$. As shown in Fig. 1, there are six instances. Bag 1 and Bag 2 contain three instances respectively. The bag label is denoted by Y. Y=1 indicates the bag is positive; otherwise, negative. For simplicity, the length of all instances and the instance number in different bags are set to fixed values in this paper. A precursor $\mathbf{Z}$ is the time series data from involved variables during the interval of a specific instance prior to the anomaly that satisfies

$$P(Y = 1|\mathbf{Z}) > \xi, \tag{2.1}$$

where ξ is a threshold defined by users or learned from data. Formally, the precursor is formulated as

$$\mathbf{Z} = (\mathbf{x}_{t_z}^{[l_1 \cdots l_M]}, \cdots, \mathbf{x}_{t_z+I-1}^{[l_1 \cdots l_M]}) \in \mathbb{R}^{M \times I}, \tag{2.2}$$

where $\mathbf{x}_t^{[l_1 \cdots l_M]} = (x_t^{l_1}, \cdots, x_t^{l_M})^\top \in \mathbb{R}^M$. M is the number of involved variables for the anomaly. $l_1 \cdots l_M$ are the indexes of these variables. $1 \leqslant l_1, , l_M \leqslant N$. t_z is the starting time step of the instance that contains the precursor. Given time series data $\mathbf{X} \in \mathbb{R}^{N \times T}$ and bag label Y, the anomaly precursor detection is to detect the precursor $\mathbf{Z}$, i.e., to detect $l_1 \cdots l_M$ and t_z.

3 Deep Multi-Instance Contrastive Learning with Dual Attention

The framework of MCDA is shown in Fig. 1. The time series data is organized in the forms of instances and bags based on MIL. The immediately preceding bag for an annotated anomaly is considered as positive, i.e., $Y = 1$; otherwise, negative. The transformed representation of an instance $\mathbf{E}_k$ is generated by

$$\mathbf{G}_k = f(\mathbf{E}_k) = f(\mathbf{x}_{t_k}, \cdots, \mathbf{x}_{t_k+I-1}) \in \mathbb{R}^{N \times d}, \tag{3.3}$$

where $f(\cdot)$ represents the tensorized LSTM and d is the hidden dimensionality for each sensory variable. The transformed representation of a bag $\mathbb{B}$ is generated by

$$\mathbf{Q} = g(\mathbb{B}) = g(\mathbf{G}_1, \cdots, \mathbf{G}_n), \tag{3.4}$$

where $g(\cdot)$ represents the attention-based MIL pooling. The bag representations with their labels are further utilized to optimize the objective function. Then the future data is fed into the model to generate representations of the testing bags. MCDA calculates how similar they are to the representations of positive bags extracted in the training phase. If the distance between a testing bag and the labeled positive bag, such as Euclidean distance, is lower than a threshold δ, we regard the testing bag as positive; otherwise as negative.

3.1 Time Series Data in the MIL Framework

Anomaly precursor detection is a weakly-supervised learning task, i.e., only anomaly event label is available but the corresponding precursor period remains uncertain. MIL is a good match for dealing with this task. In MIL setting, data instances are organized in different groups called bags. A binary label $Y \in \{0, 1\}$ is associated with a bag $\mathbb{B} = \{\mathbf{E}_1, \cdots, \mathbf{E}_n\}$. $Y = 1$ indicates the bag is positive and $Y = 0$ indicates negative. MIL assumes that each instance in $\mathbb{B}$ is with an individual label, i.e., $y_1, \cdots, y_n$, and $y_k \in \{0, 1\}$, but the instance labels are not available. Based on the assumption of MIL, the relation between the bag label and the instance labels is described in Eq. (3.5). We apply the MIL technique to the bag of time series segment instances to infer which instances contain the precursor, i.e., when exactly the precursor symptoms show up.

$$Y = \begin{cases} 0, \text{ if } \sum_{k=1}^{n} y_k = 0; \\ 1, \text{otherwise.} \end{cases} \tag{3.5}$$

A simple example of applying MIL to the time series data is shown in Fig. 1. In Fig. 1, the time series data is generated from three sensory variables and an anomaly period is annotated. The time interval of a bag is chopped into a bunch of time series segments (i.e., instances) with a sliding window. The bag immediately preceding a labeled anomaly period is considered positive (Bag 1 in Fig. 1), which indicates it contains at least one anomaly precursor; otherwise, the bag is considered negative (Bag 2 in Fig. 1). The instance that contains anomaly precursor symptoms is considered as positive. We can utilize MIL to infer the instances that contain the precursor, but MIL itself does not consider the temporal pattern of time series data.

3.2 Interpretable LSTM Networks The LSTM network is a powerful approach to capture the temporal behavior of sequential data. However, the typical LSTM cannot learn the independent representation for each sensory variable exclusively based on the data from that variable. To learn which sensory variables are responsible for the anomaly, and at the same time encode the correlation information between different pairs of time series, a novel LSTM with tensorized hidden states is proposed. The idea of tensorizing LSTM has been used in some recent work [17] and has shown its advantages for sequential tasks. Our model differs from them in two crucial ways. First, the hidden state explicitly contains the correlation information between sensory variables, which helps MCDA detect the precursor of the anomaly resulting from the correlation change between sensory variables, which is verified in Sec. 4.2.3. Second, our model is able to detect both the time period and the

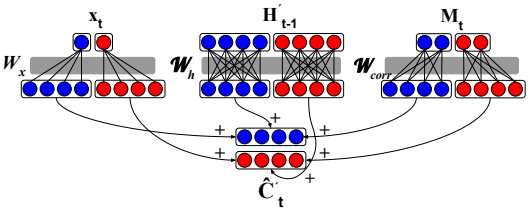


Figure 2: The derivation of cell updating matrix $\tilde{\mathbf{C}}_t$. There are two sensory variables. The dimensionality of the hidden state for each variable is four.

sensors that the precursor symptoms are involved in.

The intuition behind our tensorized LSTM is that we use a state matrix $\mathbf{H}_t = (\mathbf{h}_t^1, \cdots, \mathbf{h}_t^N)^\top \in \mathbb{R}^{N\times d}$ to represent the hidden state of all variables. $\mathbf{h}_t^l \in \mathbb{R}^d$ is the hidden feature for the l-th variable. We ensure that all the data used to generate $\mathbf{h}_t^l$ is exclusively related to the l-th variable. To consider the correlation, we feed a variable correlation matrix $\mathbf{M}_t = (\mathbf{m}_t^1, \cdots, \mathbf{m}_t^N)^\top$ into our model, where $\mathbf{m}_t^l \in \mathbb{R}^N$ indicates the correlation between the l-th variable and others at time t. Our LSTM unit is described in Eqs. (3.6)-(3.9). Given the input data $\mathbf{x}_1, \mathbf{x}_2, ... \in \mathbb{R}^N$ and $\mathbf{M}_t$, a cell matrix $\mathbf{C}_t \in \mathbb{R}^{N\times d}$ and a state matrix $\mathbf{H}_t$ are calculated.

(3.6)
$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{W}_x * \mathbf{x}_t + \boldsymbol{\mathcal{W}}_h \otimes_N \mathbf{H}_{t-1} + \boldsymbol{\mathcal{W}}_{corr} \otimes_N \mathbf{M}_t + \mathbf{B}_c)$$

$$\begin{bmatrix}\mathbf{f}_t\\ \mathbf{i}_t\\ \mathbf{o}_t\end{bmatrix} = \begin{bmatrix}\sigma\\ \sigma\\ \sigma\end{bmatrix} \{\mathbf{W}[\mathbf{x}_t \oplus \mathrm{vec}(\mathbf{H}_{t-1}) \oplus \mathrm{vec}(\mathbf{M}_t)] + \mathbf{b}\}, \tag{3.7}$$

$$\mathbf{C}_t = mat(\mathbf{f}_t \odot \mathrm{vec}(\mathbf{C}_{t-1}) + \mathbf{i}_t \odot \mathrm{vec}(\tilde{\mathbf{C}}_t)), \tag{3.8}$$

$$\mathbf{H}_t = mat(\mathbf{o}_t \odot \tanh(\mathrm{vec}(\mathbf{C}_t))), \tag{3.9}$$

where $\mathbf{W}_x \in \mathbb{R}^{N\times d}$, $\boldsymbol{\mathcal{W}}_h \in \mathbb{R}^{N\times d\times d}$, $\boldsymbol{\mathcal{W}}_{corr} \in \mathbb{R}^{N\times d\times N}$, $\mathbf{W} \in \mathbb{R}^{3Nd\times(N+Nd+NN)}$, $\mathbf{B}_c \in \mathbb{R}^{N\times d}$, $\mathbf{b} \in \mathbb{R}^{3Nd}$ are parameters. $\mathbf{f}_t$, $\mathbf{i}_t$, $\mathbf{o}_t \in \mathbb{R}^{Nd}$ are forget, input, output gates respectively, and their values are in the range of [0,1]. $\sigma(\cdot)$ represents the element-wise *sigmoid* activation function. $\oplus$ denotes concatenation operator, $\odot$ denotes element-wise multiplication. $\mathrm{vec}(\cdot)$ concatenates the rows of a matrix into a vector. $mat(\cdot)$ reshapes a vector into a matrix with size $N \times d$. $\mathbf{W}_x = (\mathbf{w}_x^1, \cdots, \mathbf{w}_x^N)^\top$ is a transition matrix. $\boldsymbol{\mathcal{W}}_h = (\mathbf{W}_h^1, \cdots, \mathbf{W}_h^N)^\top$ is a transition tensor, where $\mathbf{W}_h^l \in \mathbb{R}^{d\times d}$. $\boldsymbol{\mathcal{W}}_{corr} = (\mathbf{W}_{corr}^1, \cdots, \mathbf{W}_{corr}^N)^\top$ is a transition tensor, where $\mathbf{W}_{corr}^k \in \mathbb{R}^{d\times N}$. $\mathbf{M}_t$ is a localized correlation matrix at time t, which is figured out by the input data around time t represented by $(\mathbf{x}_{t-s}, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_{t+s})^\top \in \mathbb{R}^{N\times(2s+1)}$, where $2s+1$ is the size of the local input data. Note that $2s+1 <= I$. In this paper, we use Pearson correlation coefficient to calculate $\mathbf{M}_t$. Note that other correlation metrics can also be adopted. Eq. (3.6) calculates the cell updating matrix $\tilde{\mathbf{C}}_t = (\tilde{\mathbf{c}}_t^1, \cdots, \tilde{\mathbf{c}}_t^N)^\top$, where $\tilde{\mathbf{c}}_t^l \in \mathbb{R}^d$. $\tilde{\mathbf{c}}_t^l$ is generated exclusively from the data related to the l-th variable. The calculation process is shown in Fig. 2. $\mathbf{W}_x * \mathbf{x}_t$ captures the information from the input data and is defined by

$$\mathbf{W}_x * \mathbf{x}_t = (\mathbf{w}_x^1 x_t^1, \cdots, \mathbf{w}_x^N x_t^N)^\top. \tag{3.10}$$

$\boldsymbol{\mathcal{W}}_h \otimes_N \mathbf{H}_{t-1}$ captures the information from the previous hidden state and is defined by

$$\boldsymbol{\mathcal{W}}_h \otimes_N \mathbf{H}_{t-1} = (\mathbf{W}_h^1 \mathbf{h}_{t-1}^1, \cdots, \mathbf{W}_h^N \mathbf{h}_{t-1}^N)^\top, \tag{3.11}$$

where $\otimes_N$ indicates the tensor product along the axis of N. $\boldsymbol{\mathcal{W}}_{corr} \otimes_N \mathbf{M}_t$ captures the information from the correlation, defined by

$$\boldsymbol{\mathcal{W}}_{corr} \otimes_N \mathbf{M}_t = (\mathbf{W}_{corr}^1 \mathbf{m}_t^1, \cdots, \mathbf{W}_{corr}^N \mathbf{m}_t^N)^\top. \tag{3.12}$$

Eqs.(3.7) calculate the three gates and the process is shown in Fig. 3, where $\mathbf{W}_f = \mathbf{W}_{1:Nd,*}$,$\mathbf{W}_i = \mathbf{W}_{Nd+1:2Nd,*}$,$\mathbf{W}_f = \mathbf{W}_{2Nd+1:3Nd,*}$. The calculation of all three gates utilizes $\mathbf{x}_t$, $\mathbf{H}_{t-1}$ and $\mathbf{M}_t$, so as to utilize the cross-correlation between input variables. Eq. (3.8) updates the cell state $\mathbf{C}_t$. Eq. (3.9) calculates the new hidden state. The hidden state $\mathbf{H}_t$ at the last time step is used as the transformed representation for the input instance. Note that the gates only scale $\mathbf{C}_{t-1}$ and $\tilde{\mathbf{C}}_t$, so the variable-wise data organization is kept in $\mathbf{H}_t$.

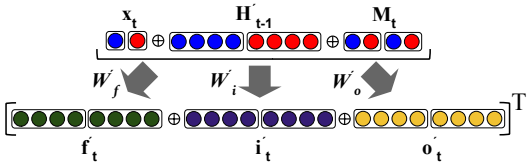


Figure 3: The diagram of the gate calculation process for the forget gate $\mathbf{f}_t$, input gate $\mathbf{i}_t$ and output gate $\mathbf{o}_t$.

3.3 Dual Attention Mechanism Based on our LSTM, we can get the transformed representations of all instances with the exclusive feature for each sensory variable, but we do not know which instances and which variables the anomaly precursor is more involved in. Because it is capable of adaptively capturing the pertinent information [18] attention mechanism is a natural fit for achieving these two goals. Thus we propose a dual attention mechanism and its diagram is shown in Fig. 4. Our work is the first one to apply attention mechanism to find out both 'when' and 'where' for the anomaly precursor.

Assume the transformed representation of instance $\mathbf{E}_k$ is denoted by $\mathbf{G}_k = (\mathbf{g}_k^1, \cdots, \mathbf{g}_k^N)^\top$, where $\mathbf{g}_k^l \in \mathbb{R}^d$. Inspired by [14] that combines $\tanh(\cdot)$ with the gating mechanism to enhance the non-linearity, we use the following attention mechanism to extract the attention values for different instances.

(3.13)
$$\alpha_k = \frac{\exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\,\mathrm{vec}(\mathbf{G}_k)^\top) \odot \sigma(\mathbf{U}\,\mathrm{vec}(\mathbf{G}_k)^\top))\}}{\sum_{i=1}^n \exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\,\mathrm{vec}(\mathbf{G}_i))^\top \odot \sigma(\mathbf{U}\,\mathrm{vec}(\mathbf{G}_i))^\top)\}},$$

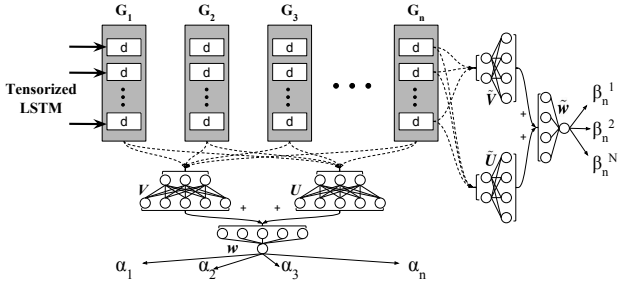


Figure 4: The diagram of the dual attention mechanism. The white rectangle with the number d inside indicates the feature representation for each variable.

where $\mathbf{w} \in \mathbb{R}^S$, $\mathbf{V} \in \mathbb{R}^{S\times(Nd)}$, $\mathbf{U} \in \mathbb{R}^{S\times(Nd)}$ are parameters. n is the instance number of a bag and S is a hyper-parameter. $\sigma(\cdot)$ is the gating mechanism part. To extract the attention values for different sensory variables, we design the following attention mechanism.

$$\beta_k^l = \frac{\exp\{\tilde{\mathbf{w}}^\top(\tanh(\tilde{\mathbf{V}}(\mathbf{g}_k^l)^\top)) \odot \sigma(\tilde{\mathbf{U}}(\mathbf{g}_k^l)^\top))\}}{\sum_{i=1}^N \exp\{\tilde{\mathbf{w}}^\top(\tanh(\tilde{\mathbf{V}}(\mathbf{g}_k^l))^\top \odot \sigma(\tilde{\mathbf{U}}(\mathbf{g}_k^l)^\top)\}}, \tag{3.14}$$

where $\tilde{\mathbf{w}} \in \mathbb{R}^{\tilde{S}}$, $\tilde{\mathbf{V}} \in \mathbb{R}^{\tilde{S}\times d}$, $\tilde{\mathbf{U}} \in \mathbb{R}^{\tilde{S}\times d}$ are parameters. N is the variable number and $\tilde{S}$ is a hyper-parameter. β_k^l indicates the attention values of the l-th variable for the k-th instance.

Based upon the representations of instances and the attention values, we can construct the transformed representation for a bag using an attention-based MIL pooling. The attention values of the instances in bag $\mathbb{B} = \{\mathbf{E}_1, \cdots, \mathbf{E}_n\}$ are denoted by $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^\top$. Here, we use the representation of the instance with the largest instance attention value to represent the whole bag. Suppose the index of the largest α is

$$k^* = \arg\max(\alpha_k) \quad s.t. \sum_k \alpha_k = 1, \tag{3.15}$$

where $k = 1, \cdots, n$. The variable values for instance $\mathbf{E}_{k^*}$ are denoted by $\boldsymbol{\beta}_{k^*} = (\beta_{k^*}^1, \cdots, \beta_{k^*}^N)^\top$. If the representation of $\mathbf{E}_{k^*}$ is $\mathbf{G}_{k^*} = (\mathbf{g}_{k^*}^1, \cdots, \mathbf{g}_{k^*}^N)^\top$, then the representation of bag $\mathbb{B}$ can be derived by [2]

$$\mathbf{Q} = \mathbf{G}_{k^*} * \boldsymbol{\beta}_{k^*} = (\mathbf{g}_{k^*}^1\beta_{k^*}^1, \cdots, \mathbf{g}_{k^*}^N\beta_{k^*}^N)^\top \in \mathbb{R}^{N\times d}. \tag{3.16}$$

In case that multiple instances jointly characterize a type of precursor symptom, we only need to reformulate the bag representation to $\mathbf{Q} = \alpha_1(\mathbf{G}_1 * \boldsymbol{\beta}_1) + \alpha_2(\mathbf{G}_2 * \boldsymbol{\beta}_2) + \cdots + \alpha_n(\mathbf{G}_n * \boldsymbol{\beta}_n)$.

3.4 Objective Function Given the transformed representations of bags denoted by $\mathbf{Q}_1, \cdots, \mathbf{Q}_M$, where M is the number of bag, and the bag labels $Y_1, \cdots, Y_M$, the objective function of MCDA is

$$\min J = J_{cont} + \lambda J_{reg}. \tag{3.17}$$

$J_{cont} = \sum_{i,j}\{(1 - P_{ij})\frac{1}{2}D_{ij}^2 + P_{ij}\frac{1}{2}\{\max(0, \eta - D_{ij})\}^2\}$ is the bag pair contrastive loss. i, j are the bag indexes. P_{ij} is the pair label. $P_{ij} = 1$ if $Y_i = Y_j$; otherwise 0. $D_{ij} = D(\mathbf{Q}_i, \mathbf{Q}_j)$ is the bag distance. η is a threshold. It makes the representations of two bags with the same label be similar and the ones with different labels be dissimilar by minimizing J_{cont}. Here, we use contrastive loss considering its superiority on cases where the labeled data is few[3], which is quite common for anomaly detection task. Note that other loss functions can also be adopted, such as triplet loss [19]. J_{reg} is an regularization term and prevents our model from overfitting. λ is a hyperparameter.

Table 1: Description of the datasets for Task 1

Dataset	SC28C1	SC28C2	SC34C1	SC34C2	SC29C1
# positive bag	125	165	175	165	170
# negative bag	125	165	175	165	170
# features	42	42	42	42	42
length of bag	200	200	200	200	200

4 Experiments

4.1 Datasets and Baseline Methods We evaluate MCDA on both synthetic and real datasets. One real dataset is Showcase provided by a major retail chain company[4]. The data are real-time sensor monitoring time series of refrigerator system in different showcases from different stores. Each showcase contains 42 sensory time series monitoring temperature, air pressure, humidity, etc, from Feb 12, 2013 to Sept 15, 2016. Each of them includes 814,535 time-step records. Three showcases, SC28, SC29 and SC34, are reported with the same type of anomaly event with precursor periods annotated by domain experts. Another is a real cyber-physical system data [3] generated from manufacturing industry. This dataset contains time series collected from 1625 electric sensors installed on different components of the cyber-physical system. The total length is 1400 time steps. The anomaly occurred at the 210-th time step.

To verify the advantage of MCDA, we conduct experiments on three tasks. Task 1 is *precursor time series detection* (T1) to test whether MCDA can outperform the baseline methods for detecting the time series segment containing anomaly precursors. Task 2 is *anomaly*

[2]The operation $*$ is defined the same as the one in Eq. (3.10).

[3]Please see discussion in the supplementary materials for dealing with the case where annotated anomalies are few.

[4]For privacy, we remove sensitive descriptions of the data.

Table 2: Precursor detection accuracy (%) results.

Dataset	SC28C1	SC28C2	SC34C1	SC34C2	SC29C1
Kmeans	57.2±3.1	62.7±1.2	72.1±1.7	63.7±2.2	64.5±0.9
GAKK	54.0±2.4	50.6±1.0	50.9±1.8	52.3±1.4	56.4±0.6
LSTM-AE	54.0±0.0	58.7±2.1	64.9±1.6	62.1±1.5	63.8±0.4
DAGMM	35.6±1.1	63.9±2.5	61.4±0.8	58.2±3.8	47.6±2.9
DTW	45.2±1.6	69.7±1.3	59.7±2.3	73.7±1.8	81.5±2.1
L2	44.8±1.2	66.4±2.2	67.1±2.8	66.3±2.2	65.2±1.7
SAX	58.8±1.6	42.1±3.9	37.6±1.9	36.6±2.1	69.9±5.6
SVM	49.2±2.3	76.1±0.9	72.8±2.7	65.0±1.7	74.2±2.8
MI-SVM	46.8±0.7	84.5±0.7	74.3±0.4	74.2±0.5	81.2±0.2
MCDA-V	52.4±2.5	67.0±1.8	65.0±2.2	63.6±1.4	77.1±1.7
MCDA-MIL	49.6±1.9	50.6±1.3	72.6±1.1	66.7±1.5	75.9±1.5
MCDA-M	51.6±2.1	50.0±1.0	64.3±1.9	62.7±2.3	83.2±2.4
MCDA	**59.2±2.5**	**85.2±1.4**	**77.4±2.1**	**76.1±1.9**	**92.6±1.7**

Table 3: Precursor detection F1-measure (%) results.

Dataset	SC28C1	SC28C2	SC34C1	SC34C2	SC29C1
DTW	9.3±1.6	62.7±1.4	36.3±2.6	69.3±2.9	80.4±1.9
L2	8.0±1.2	56.8±2.1	54.5±2.8	55.0±2.8	52.7±2.1
SAX	60.8±1.9	41.5±5.2	37.2±2.7	40.0±2.4	67.9±3.2
SVM	36.2±2.3	77.9±0.9	73.8±2.7	65.7±1.7	73.8±2.8
MI-SVM	46.8±0.5	86.2±0.9	73.6±0.3	73.7±0.8	81.1±0.3
MCDA-V	67.6±2.5	75.1±2.5	74.0±1.9	73.2±2.0	81.3±1.2
MCDA-MIL	66.8±1.9	50.6±1.4	76.3±0.8	64.1±0.8	80.5±1.7
MCDA-M	62.1±2.7	66.5±2.1	72.5±1.5	66.4±1.8	85.6±2.2
MCDA	**70.9±1.7**	**87.0±2.4**	**78.7±2.3**	**78.3±2.2**	**93.1±1.5**

detection (T2) to test whether MCDA can detect the anomaly effectively and earlier compared to the baseline methods. Task 3 is *interpreting precursor* (T3) to test whether MCDA can detect the time period and the sensory variables of the precursor.

The baseline methods are summarized in the supplementary materials. Kmeans and GAKK [20] group time series segments into two clusters. The predicted label of a segment is the same as the label of its corresponding cluster and the cluster label is determined by the major segments. DTW [21] and L2 implement the kNN vote for time series based on dynamic time warping and euclidean distance respectively. SAX realizes the nearest neighbor classification based on SAX representation [22]. SVM classifies the segments into the abnormal ones and the normal ones. MI-SVM [13] is a popular MIL approach. It modifies the standard SVM formulation so that the constraints on instance labels correspond to the MIL assumption. Kmeans, GAKK, DTW, L2, SAX and SVM are implemented by the tslearn package[5]. LSTM-AE [23] uses the reconstruction error to detect anomalies. DAGMM [5] regards the time series data with high energy as anomalies.

To evaluate the different components in MCDA, we study its variants. MCDA-V is the variant without applying attention mechanism to the sensory variables. It generates the transformed representations of instances by tensorized LSTM and constructs bag representations based on multi-instance pooling. MCDA-MIL is the variant without applying attention mechanism to the instances. It extracts bag representations by tensorized LSTM directly, instead of utilizing MIL. MCDA-M does not consider the correlation information between sensory variables by removing $\mathbf{M}_t$.

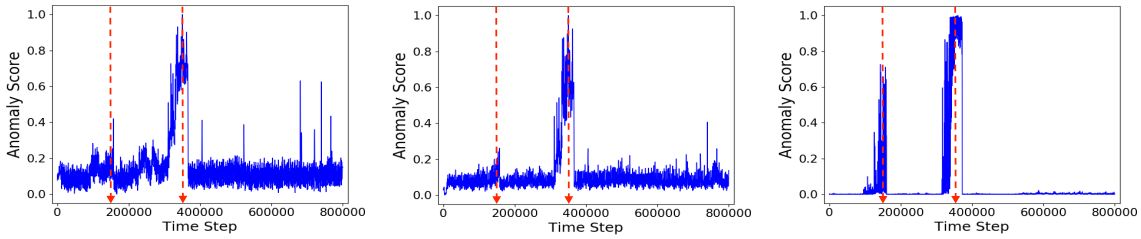
In our experiments, for Showcase and Cyber-physical system, the instance length is set to 50 and 5 time steps respectively, the bag length is set to 200 and 20 time steps respectively and the size of the sliding window to generate instances is set to 10 and 1 time step respectively. The learning rate is set to 10^{-3} initially and decreases during the training. λ is set to 0.5. η is set to 10. d, S, and $\tilde{S}$ are set to 15. s is set to 1. They are determined by grid-search from $\{0.125, 0.25, 0.5, 1, 2, 4\}$, $\{1, 2, 5, 10, 15, 20\}$ and $\{5, 10, 15, 20\}$ respectively. The ℓ_2 regularization is adopted for J_{reg}. For all supervised methods, we randomly select 1/5 of the training set as validation set to determine the best hyper-parameters. MCDA is optimized by Adam [24]. The code of MCDA and the data used are available[6].

4.2 Experimental Results

4.2.1 Performance of Precursor Time Series Detection (Task 1) We compare MCDA and other methods in terms of their performance of detecting the bags containing the precursors, i.e., positive bags. The performance is evaluated by accuracy and F1-measure [5]. Based on SC28, SC29 and SC34 that are marked with precursors, we generate positive bags (time series segments with length 200 time steps) by sampling the sliding windows on the original time series data. For each of them, we randomly sample the same number of negative bags with the same length in system normal period. We use sliding window with stride 10 to chop a bag into instances, which results in 16 instances in each bag. There are two precursor cases in SC28 and SC34, and one in SC29, thus we generate 5 kinds of positive bags. The dataset description is summarized in Table 1. The name of SC28C1 indicates it is based on precursor case 1 in showcase 28. The naming convention is similar for others. For the supervised methods, we use SC28C1, SC28C2 as the training set and test on SC34C1, SC34C2, SC29C1, and use SC34C1, SC34C2 as training and then test on SC28C1, SC28C2. The threshold δ for MCDA to detect the positive bag is set

[5]https://github.com/rtavenar/tslearn

[6]https://tinyurl.com/yd8tn76b



(a) LSTM-AE (b) DAGMM (c) MCDA

Figure 5: Anomaly detection results on Showcase 28. Red dots indicate the time of anomalies labeled by users.

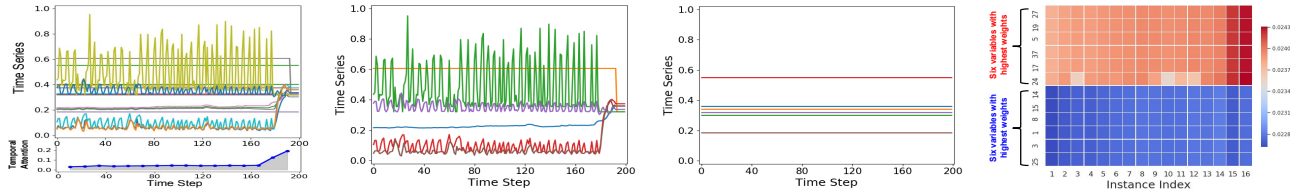
to the median value of all the distances between testing bags and the labeled positive bag.

Tables 2 and 3 show the results. From Table 2, MCDA achieves the best performance, which verifies the advantage of MCDA to detect the anomaly precursor. The supervised methods outperform the unsupervised ones in general. This is because the label information utilized in supervised methods can help extracting discriminative precursor features more accurately. MCDA outperforms MI-SVM, which indicates focusing on the relevant sensory variables can improve the detection performance. MCDA outperforms MCDA-V, indicating the importance of attention mechanism on learning better precursor feature. MCDA outperforms MCDA-MIL, indicating the effectiveness of multi-instance technique in capturing most relevant time period when precursors show up. By comparing MCDA with MCDA-M, we see the necessity of leveraging the correlations between sensory variables. Without considering the correlation change, the performance degrades. Similar observations are observed in Table 3. Note that there is no F1-measure results of Kmeans, GAKK, LSTM-AE and DAGMM because they are not supervised methods.

4.2.2 Performance of Anomaly Detection (Task 2) We evaluate the effectiveness of MCDA on the task of on-line early anomaly detection. We compare MCDA with LSTM-AE and DAGMM. Figs. 5 shows the results on the Showcase data. The red dots indicate the time of anomalies labeled by users. Higher anomaly score indicates higher probability of anomaly. We use Showcase 34 as the training data for MCDA and show the detection results on the data of Showcase 28, which has two reported anomalies. The results on other showcases are similar thus are omitted. According to Figs. 5(a) and 5(b), both LSTM-AE and DAGMM can only detect the second anomaly. In Fig. 5(c), MCDA succeeds in detecting incipient faults of both anomalies with less false positives, which verifies its effectiveness to detect anomaly precursors. Similar observations can be made on Cyber-physical system dataset that are included in the supplementary materials.

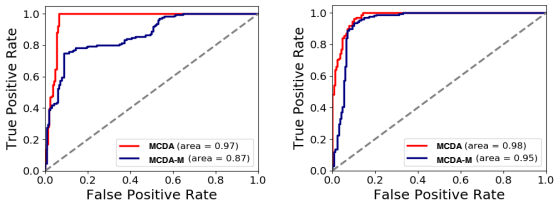
4.2.3 Interpreting When and Where of Precursors (Task 3) We use the Showcase 28 data to evaluate the ability of MCDA to detect the anomaly precursor. One positive bag is shown in the upper part of Fig. 6(a). The precursor annotated by domain experts is located in the last several time steps. The blue line at bottom of Fig. 6(a) indicates the attention values generated by MCDA for different instances. We can see that the attention value of the last period is significantly larger than others, which demonstrates MCDA detects the time location of precursor successfully. The six variables with the highest attention values in the last instance are shown in Fig. 6(b). They changed sharply during the interval of precursor, which indicates they are highly correlated with the anomaly afterwards. The six variables with the lowest are shown in Fig. 6(c) and they keep constant. The variable attention of these twelve variables is shown in Fig. 6(d). The attention values of the six variables in Fig. 6(b) are larger than that of the variables in Fig. 6(c), especially in the last two instances. Based on the annotation of domain experts, the six variables in Fig. 6(b) are the sensors related to the anomaly, which verifies the ability of MCDA to detect the sensor location of precursor. Results on another positive bag and synthetic data are shown in the supplementary materials.

4.2.4 Effectiveness of $\mathbf{M}_t$ The term $\mathbf{M}_t$ in the cell structure enables MCDA to use the variable correlation to improve its anomaly detection performance. To verify it, we applied MCDA-M that does not utilize $\mathbf{M}_t$. We use Showcases 28 and 34 as the training set respectively and test the precursor case from Showcase 29. The ROC (receiver operating characteristic) curves of MCDA and MCDA-M are shown in Fig. 7. It is observed the area under the ROC curve of MCDA is larger than that of MCDA-M, which verifies the effectiveness of $\mathbf{M}_t$. Besides, from Tables 2, MCDA outperformed MCDA-M, which also indicates the correlation information between sensory variables utilized in MCDA can improve the anomaly detection performance.



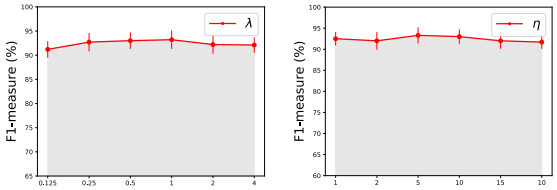
(a) Temporal attention distribution across time steps (b) Six time series (variables) with highest attention values (c) Six time series (variables) with lowest attention values (d) Variable attention distribution across instances

Figure 6: The precursor detection result on a positive bag from the Showcase data. The precursor is located in the last several time steps.



(a) Showcases 28, 29 (b) Showcases 29, 34

Figure 7: Comparison of precursor detection results.



(a) Varying λ (b) Varying η

Figure 8: Parameter sensitive analysis.

4.2.5 Parameter Sensitivity We select two important parameters, the regularization term λ in the objective function and the threshold η in the contrastive loss, and study how they will affect the performance of MCDA. We study their influence on the precursor time series detection. The result is shown in Fig. 8. We conduct the experiments on the SC29C1 dataset. We take λ=0.5 and η=10 as the basic setting, which obtains the best performance. Figs. 8(a)-8(b) show the results based on different values of λ and η. It is observed that MCDA is not sensitive to both λ and η.

5 Related Work

There have been intensive interests in developing anomaly detection methods because of its practical importance [3–8]. The energy based deep models with two decision criteria for anomaly detection were proposed in [4]. Zong et al. [5] proposed a Gaussian mixture model based anomaly detection method. Precursor detection has drawn increasing research attention recently [11,12,25]. Ning et al. [25] proposed a multi-task spatio-temporal correlation graph model for precursor mining coupled with event forecasting. [12] employed the gated recurrent unit to detect the precursors for aviation safety incidents. However, none of these approaches is able to detect both the time period and the sensory variables the precursors are involved in.

Multi-instance learning (MIL) is popularly used to address weakly-supervised learning problem [11, 13–15, 26]. For example, MI-SVM [13] modifies the standard SVM formulation so that the constraints on instance labels correspond to the MIL assumption that at least one instance in a positive bag is positive. Ning et al. [11] utilized MIL to predict the protest events. Some researchers combined MIL with deep neural networks [14, 26]. Basically, none of these MIL approaches is developed for anomaly precursor detection in multivariate time series.

Nowadays, the attention technique gets growing popularity [14,15,18,27,28]. Qin et al. [27] combined the hidden and cell states of all sensory variables with the input of the k-th variable to produce the attention value for the k-th variable. A network architecture that solely relied on an attention mechanism was proposed in [18]. Our work is also related to some recent work on tensorizing LSTM [17,29]. For example, the authors in [29] utilized a tensorized hidden state to learn independent feature representation for different variables. Basically, these approaches are designed for time series prediction rather than anomaly precursor detection. Besides, the hidden state in our model explicitly contains the correlation information between sensory variables, which helps MCDA detect the precursor of the anomaly resulting from the correlation change between sensory variables.

6 Conclusion

In this paper, we propose a method MCDA to identify features of anomaly precursors, which enables detection of future anomalies in early stage. MCDA incorporates multi-instance learning to deal with the uncertainty of precursor period. With the tensorized RNN, MCDA is able to learn the hidden representations of different

variables capturing information from a certain variable of the input, meanwhile considers the correlation of pairwise time series. Its dual attention module enables MCDA to interpret when and where the precursor symptoms show up. Extensive experimental results demonstrate the effectiveness of MCDA.

References

[1] H. Chen, H. Cheng, G. Jiang, and K. Yoshihira, "Exploiting local and global invariants for the management of large scale information systems," in *ICDM*. IEEE, 2008, pp. 113–122.

[2] D. Djurdjanovic, J. Lee, and J. Ni, "Watchdog agent - an infotronics-based prognostics approach for product performance degradation assessment and prediction," *Advanced Engineering Informatics*, vol. 17, pp. 109–125, 2003.

[3] W. Cheng, K. Zhang, H. Chen, G. Jiang, Z. Chen, and W. Wang, "Ranking causal anomalies via temporal and dynamical analysis on vanishing correlations," in *SIGKDD*. ACM, 2016, pp. 805–814.

[4] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Deep structured energy based models for anomaly detection," in *ICML*, 2016, pp. 1100–1109.

[5] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *ICLR*, 2018.

[6] W. Yu, W. Cheng, C. C. Aggarwal, K. Zhang, H. Chen, and W. Wang, "Netwalk: A flexible deep embedding approach for anomaly detection in dynamic networks," in *SIGKDD*. ACM, 2018, pp. 2672–2681.

[7] C. Zhang, D. Song, Y. Chen, X. Feng, C. Lumezanu, W. Cheng, J. Ni, B. Zong, H. Chen, and N. V. Chawla, "A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data," *arXiv preprint arXiv:1811.08055*, 2018.

[8] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "Usad: Unsupervised anomaly detection on multivariate time series," in *SIGKDD*, 2020, pp. 3395–3404.

[9] N. B. Sarter and H. M. Alexander, "Error types and related error detection mechanisms in the aviation domain: An analysis of aviation safety reporting system incident reports," *The international journal of aviation psychology*, vol. 10, no. 2, pp. 189–206, 2000.

[10] D. Hallac, S. Vare, S. Boyd, and J. Leskovec, "Toeplitz inverse covariance-based clustering of multivariate time series data," in *SIGKDD*. ACM, 2017, pp. 215–223.

[11] Y. Ning, S. Muthiah, H. Rangwala, and N. Ramakrishnan, "Modeling precursors for event forecasting via nested multi-instance learning," in *SIGKDD*, 2016, pp. 1095–1104.

[12] V. M. Janakiraman, "Explaining aviation safety incidents using deep temporal multiple instance learning," in *SIGKDD*, 2018, pp. 406–415.

[13] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *NeurIPS*, 2003, pp. 577–584.

[14] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *ICML*, 2018.

[15] G. Nayak, R. Ghosh, X. Jia, V. Mithafi, and V. Kumar, "Semi-supervised classification using attention-based regularization on coarse-resolution data," in *SDM*. SIAM, 2020, pp. 253–261.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[17] Z. He, S. Gao, L. Xiao, D. Liu, H. He, and D. Barber, "Wider and deeper, cheaper and faster: Tensorized lstms for sequence learning," in *NeurIPS*, 2017, pp. 1–11.

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

[19] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, 2013.

[20] M. Cuturi, "Fast global alignment kernels," in *ICML*, 2011, pp. 929–936.

[21] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, 1994, pp. 359–370.

[22] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: a novel symbolic representation of time series," *DMKD*, pp. 107–144, 2007.

[23] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "Lstm-based encoder-decoder for multi-sensor anomaly detection," *arXiv preprint arXiv:1607.00148*, 2016.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] Y. Ning, R. Tao, C. K. Reddy, H. Rangwala, J. C. Starz, and N. Ramakrishnan, "Staple: Spatio-temporal precursor learning for event forecasting," in *SDM*. SIAM, 2018, pp. 99–107.

[26] M. Ilse, J. M. Tomczak, and M. Welling, "Deep multiple instance learning for digital histopathology," in *Handbook of Medical Image Computing and Computer Assisted Intervention*. Elsevier, 2020, pp. 521–546.

[27] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," *arXiv preprint arXiv:1704.02971*, 2017.

[28] M. Hahn, "Theoretical limitations of self-attention in neural sequence models," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 156–171, 2020.

[29] T. Guo, T. Lin, and N. Antulov-Fantulin, "Exploring interpretable lstm neural networks over multi-variable data," in *ICML*, 2019.