technische universität
dortmund

Seminar paper

Unsupervised Anomaly Detection of Data Streams in Devices

Mengcheng Jin
02.2022

Supervisors:
First supervisor:
Prof. Dr. Emmanuel Müller
Second supervisor: Bin Li

Technical University of Dortmund
Faculty of Computer Science
LS - 9
http://ls9-www.cs.tu-dortmund.de/

Contents

Chapter 1

Introduction

Nowadays, automation has covered many areas, many jobs. Complex devices are used every day and generate massive data streams of multidimensional measurement data in real-time to characterize their real-time status. Therefore, it is important to detect the data stream generated by the devices, which ensures the daily life of people nowadays. Occasionally, certain devices fail, causing a system outage, and postmortem analysis shows that these devices produced anomalies prior to the outage. If the technicians are notified in time, they can repair the device in time to minimize the damage before it breaks down. Finding anomalies with traditional supervised machine learning methods is often impractical or impossible because the failure is too complex, the device is too new, or the data changes too quickly and is not updated in a timely manner. Therefore, methods for unsupervised anomaly detection for data streams are of particular interest. Then due to the characteristics of unsupervised learning, when dealing with these problems, unsupervised learning is more suitable.

Unsupervised learning applies to situations where you have a data set but no labels or a few normal data labels. Relying on the model itself through continuous exploration, summarizing and summarizing the knowledge, trying to discover the inherent laws or features in the data, to label the training data. (including K-Means, PCA, AE, etc.). When the proportion of abnormal data in our data is small and the value of abnormal data is quite different from the normal value, we usually adopt unsupervised learning. However, if the proportion of abnormal data is high, it may lead to unsatisfactory final results.

Several factors should be considered when developing an unsupervised multidimensional data stream anomaly detection solution:

- **Multidimensional**
  Data streams are always multidimensional. Devices generate data streams with high dimensionality, where the anomaly is observable in a subset of dimensions, but masked in noise dimensions. Unfortunately, with increasing dimensionality, many conventional anomaly detection methods fail to work effectively.

- **Correlated**
  Data streams are always correlated, not independent. For example, under normal conditions, the desired set point temperature and the observed temperature are always highly correlated.

- **Multimodal**
  They are also always multimodal. Like some devices have an active mode and standby mode. However, there are none or only a few intermediate data points for transitions between different modes, like concept drift in time series data.

These characteristics of the data stream make anomaly detection more difficult. But at the same time, we also have to consider the different types of anomalies. There are 2 most frequently occurring anomaly types: Point Anomaly and Contextual Anomaly.

- **Point Anomaly**
  It can also be called a global anomaly, that is, a certain point is different from most of the global points, then this point constitutes a single-point anomaly.
  In the Figure 1.1[5], we can find that the three parts of O_1, O_2 and O_3 are different from most of the global points (N_1, N_2), and O_1, O_2 and O_3 are the point anomalies in this data.

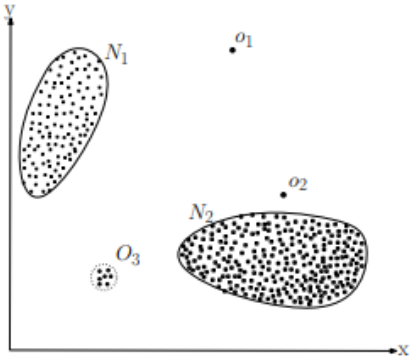


Figure 1.1: Point Anomalies O_1, O_2, O_3

- **Contextual Anomaly**
  This type of anomaly is mostly an anomaly in time series data, that is, there is a big difference between the performance at a certain point in time and the time period before and after then the anomaly is a contextual anomaly.
  In the Figure 1.2[5], note that the temperature at time t_1 is the same as that at time t_2 but occurs in a different context and hence is not considered as an anomaly[5].

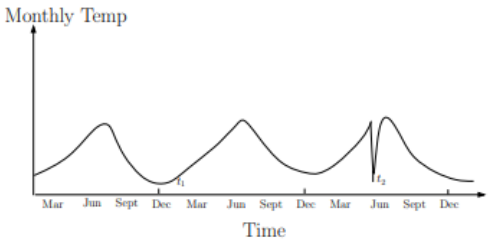


Figure 1.2: Contextual anomaly t_2 in a temperature time series.

According to the difficulties we are facing now, we need to use unsupervised anomaly detection methods to find anomalies in the data stream. And just as important is how to explain why this data is an anomalous, or further what caused the anomalous. This al-

lows these failures or other situations caused by anomalous to be dealt with in a timely manner.

In this paper, we will introduce and discuss several unsupervised anomaly detection methods and anomaly interpretation methods for the data stream.

Chapter 2

Related Works

2.1 Unsupervised anomaly detection approaches

Anomaly detection has been studied for decades. And so far, there have been many unsupervised anomaly detection methods for data streams. They can be divided into 4 broad categories, like in Figure 2.1.

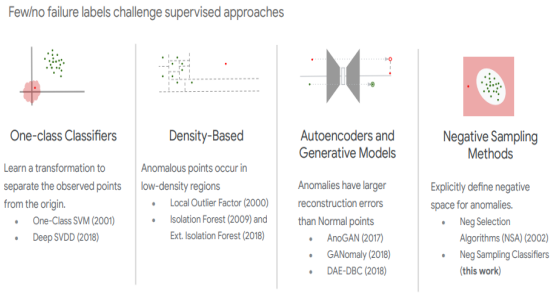


Figure 2.1: Different unsupervised anomaly detection categories

In this chapter, we want to introduce some common methods and related work for solving anomaly detection.

2.1.1 One-class Classification

One-class classifiers are trained to learn a transformation function $f : X \rightarrow c$ that generates a scalar value as a class score $c \subset C$, when the input data resembles the observed, and mostly normal, data stream[1]. Then, we consider some data points as outliers when their c values are clearly different from the values in C. Two of the representative models are One-class SVM(OCSVM)[14] and Deep Support Vector Data Description(DSVDD)[10].

- **DSVDD**
  SVDD is an unsupervised learning model based on SVM. Unlike OCSVM, SVDD tries to find the smallest possible hypersphere that contains most of the training set[16]. However, because SVM struggles with high-dimensional data[9], and when the training data only has normal data, it is impossible to control the false alarm rate by selecting hyperparameters[17].

- **OCSVM**
  In OCSVM only normal data are used to train the model and a hyperplane is trained that circle the normal data in the sample, like Figure 2.2. Prediction is the use of this hyperplane to make decisions, and samples that are within the circle are considered normal data and outside the boundaries are identified as anomalies. But it also has the same problems as SVDD.

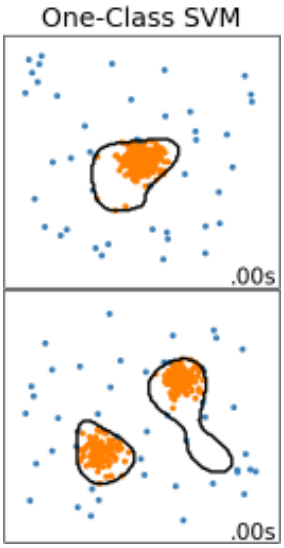


Figure 2.2: OCSVM schematic diagram

Consider the example in the Figure 2.3 above, where the distance-based approach is able to detect o3, but for o1 and o2, it is not as obvious. The idea of density-based is that we need to compare the density around an object with the density around its local neighbors. The basic assumption of the density-based outlier detection method is that the density around a non-anomalous object is similar to the density around its neighbors, while the density around an anomalous object is significantly different from the density around its neighbors.

2.1.2 Density-based

Density-based outlier detection methods study the densities of objects and their neighbors. Here, an object is identified as an outlier if its density is relatively lower than that of its neighbors. Many real-world datasets demonstrate a more complex structure, where objects may be considered outliers relative to their local neighborhoods rather than relative to the global data distribution.

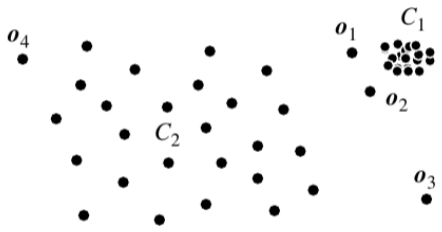


Figure 2.3: Density-based anomaly detection method

- **Isolation Forest**
  One widely used approach originally proposed[11], called Isolation Forest, is an ensemble of random trees that recursively partition the data space until all points are isolated. In the process of detecting, if some samples reach the leaf nodes soon, they may be regarded as abnormal points. To reduce the influence of randomness, a forest is formed by constructing multiple trees, and then the average path length of the sample in all trees is calculated to find anomalies. If there is a lot of noise or irrelevant dimensions in the data, it will affect the construction of the tree and reduce the reliability of the algorithm.

2.1.3 Autoencoder and generative methods

Autoencoder and generative methods only use normal data to train the encoder-decoder architecture. Then the raw data inputs into the autoencoder and the autoencoder outputs the reconstruction data. If some points have greater reconstruction error than other points. These points tend to anomalies.

- **Auroencoders**
  In 2002, Hawkins et al. proposed an Autoencoder[8] that detects anomalies by reconstructing data errors. In 2016, Malhotra et al. proposed a Long Short Term Memory(LSTM) network-based Encoder Decoder (Autoencoder) scheme for Anomaly Detection (EncDec-AD) that learns to reconstruct 'normal' time-series behavior, and thereafter uses the reconstruction error to detect anomalies [12]. And in 2018, on the basis of EncDec-AD, Zong et al. proposed in the paper to combine the deep autoencoder(Including LSTM autoencoder and vanilla autoencoder) with a Gaussian Mixture Model(DAGMM), by generating a low-dimensional representation and reconstruction error for each input data point, which is further fed into a GMM[18]. Instead of using the standard Expectation-Maximization (EM) algorithm to realize its data anomaly detection. DAGMM solves the problem that the system performance is limited due to the low convergence speed when applying the EM algorithm.

- **Generative methods**
  In 2018, Akcay et al. proposed GANomaly[2] based on the GAN and Autoencoder. Different from the general method based on self-encoder, it adopts the structure of Encoder1-Decoder-Encoder2. At the same time, learn the two mapping relationships of "original data to reconstructed data" and "original data encoding to reconstructed data encoding". Finally, the difference between the latent space features generated by the first part of the encoder (the encoding of the original data) and the latent space features (the encoding of the reconstructed data) generated by the second part of the encoder is used to pay attention to the small changes in the data[2]. That solves the problem that the encoder is susceptible to noise.

2.1.4 Negative Selection Algorithms (NSA)

NSA were initially pro as a biologically inspired method of detecting computer viruses[6]. Most NSAs apply search algorithms that attempt to emulate how antibodies distinguish pathogens from body cells. In 2020, Sipple proposed a new classifier with Negative Sampling. He used the existing dataset to build a hypercube V. And the edge length of each dimension is the

difference between the maximum and the minimum values of this dimension, which is Δv. And the data within this hypercube are treated as positive samples (also known as normal data). Then he added or subtracted a δ to each dimension to obtain a strictly greater hull U to bound negative samples(also known as outliers) with edge length Δu. By taking normal data and anomalies uniformly in V and U a binary classification model can be trained, like in Figure 2.4.

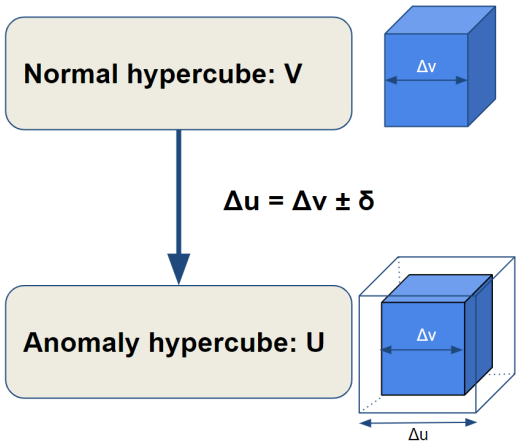
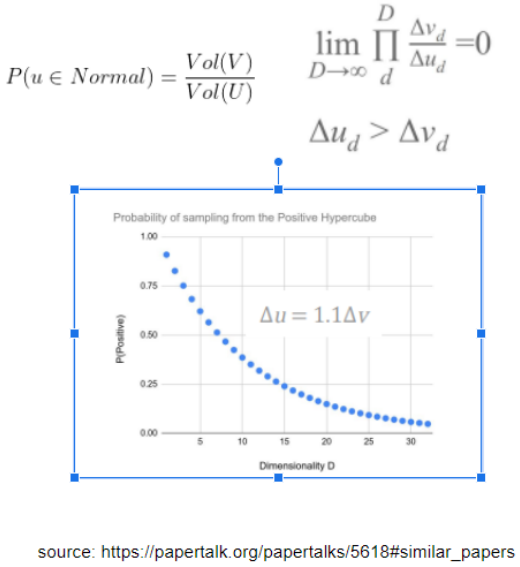


Figure 2.4: Negative sampling Algorithm

Ideally, we would like to train a binary classifier F on data that has as few labeling errors as possible. Intuitively, it makes sense to develop an algorithm that carefully selects the negative sample to avoid the Normal space. For example, Gonzalez proposes a type of region growing approach that avoids choosing points close to the positive sample[7]. However, such a sampling approach is difficult and/or computationally expensive in high dimensions because we are not able to characterize the positive volume[15]. However, Sipple proposed an easier method. Since the sampling is uniform, the probability of getting a true anomaly (true negative) in Normal Hypercube V (probability of false positive) is the ratio of the volumes of the two hypercubes. The volumes can be obtained by multiplying each edge. Due to every Δu is a little bigger than Δv, so if there is enough high dimensional data, the probability of false positive will converge to 0.



source: https://papertalk.org/papertalks/5618#similar_papers

Figure 2.5: Solve false positive problem

In Figure 2.5, it is an example for this. Even though each u is only 10% bigger than v, the probability of false positives is converging to 0 when the data exceeds 30 dimensions.

2.2 Anomaly Interpretation

Many of these anomaly detection methods have been applied in people's daily life. Therefore, the interpretation of anomalies

is also essential. If the technical staff can find the cause of the anomaly faster and discover where the anomaly is generated, the anomaly can be solved more easily. There are currently 2 major categories of interpretation methods: Integrated Gradients and Layerwise relevance propagation(LRP).

2.2.1 Integrated Gradients

For Integrated Gradient, when we attribute the occurrence of something to a cause, we should take the absence of that cause as the baseline. For example, for an image recognition system, the baseline can be an all-black picture, while for a NLP system, the baseline can be a word vector with all values of 0. Therefore, if we want to explain anomalies, we should compare anomalies to the most normal data. In Figure 2.6, Sipple defined a baseline set U^*, which is a subset from U[15]. Unlike U, the data in U^* are the most likely to be normal. After that, he used euclidean distance to calculate the distance between the data point and each point in U^*, and chose the closest u^* as the baseline of data point x. Finally, used Integrated Gradient to compute and integrates the gradient for each dimension from a baseline point to the observed point. That is, quantifying each dimension by the Integrated Gradient, which can indicate which dimensions are contributing the most to the data being normal or being abnormal. Sipple refers to this as Blame(B). If a data is abnormal, then the sum of the blames of each dimension should be close to 1.

$$U^* \subset U : \forall_{u \in U^*} F(x) \approx 1$$

$$u^* = argmin_{u \in U^*} \{dist(x, u)\}$$

$$B_d(x) \equiv (u_d^* - x_d) \times \int_{\alpha=0}^{1} \frac{\partial F(x + \alpha \times (u^* - x))}{\partial x_d} d\alpha$$

$$\sum_{d \in D} B_d(x) \approx 1$$

Figure 2.6: Baseline choose and Integrated Gradient

For example, in Figure 2.7, it's an anomaly interpretation of an anomalous point x with $F(x) = 0$.(in this work he used the term positive to refer to the space that is sampled by observation and is mostly Normal, and negative to refer to an unobserved complement space from which a labeled sample has to be generated. So $F(x) = 0$ means x is anomaly) Three dimensions assigned most of the blame. We can assume that the cause of the anomaly is mainly due to these dimensions.

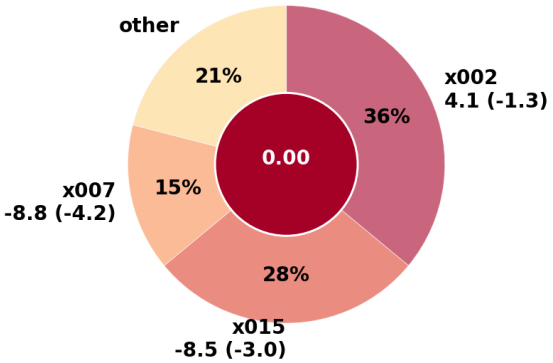


Figure 2.7: Result by Integrated Gradient[15]

2.2.2 Layerwise relevance propagation(LRP)

LRP[3] produces a heatmap in the input space indicating the importance/relevance of each part contributing to the final classification outcome. In contrast to susceptibility maps produced by guided back propagation ("Which change would change the outcome most?"), the LRP method is able to directly highlight positive contributions to the network classification in the input space[4]. The authors in paper [4] used LRP to explain the disease problem in the medical brain view and obtained relatively good results. However, LRP has only been shown to outperform gradient methods on images[13], so the results are not as good as gradient methods in data streams.

Chapter 3

Studied Models

In the paper[15], author implemented the Negative Sampling Algorithm(NSA) with Random Forest(NSRF) and Neural Networks(NSNN) and used them to compare with 4 other unsupervised anomaly detection methods, namely OCSVM, DSVDD, Isolation Forest(ISO) and Extended Isolation Forest(EIF). He compared them on 5 standard benchmark datasets and a multi-dimensional, multimodal dataset from a real climate control device.

Data set	Size	Dim	Anomaly
Forest cover (fc)	286,048	10	2,747 (0.9%)
Shuttle (sh)	49,097	9	3,511 (7%)
Mammography (mm)	11,183	6	260 (2.3%)
Mulcross (mc)	262,144	4	26,214 (10%)
Satellite (sa)	6,435	36	2,036 (32%)
Smart Buildings (sb)	60,425	7	1,921 (3.2%)

Figure 3.1: Datasets

	OCSVM	DSVDD	ISO	EIF	NSRF	NSNN
FC	53±20	69±7	85±4	**93±1**	80±2	86±4
SH	93±0	88±9	**96±1**	91±1	**93±7**	**96±5**
MM	71±7	78±6	77±2	**86±2**	**85±4**	84±2
MC	90±0	54±17	88±0	66±4	94±1	**99±1**
SA	51±1	62±3	67±2	**71±3**	65±4	**73±3**
SB	76±1	60±7	71±7	80±4	**95±1**	93±1

Figure 3.2: Result in AUC

Before conducting record trial runs, he performed hyperparameter optimization on AUC for each algorithm and selected the highest performing parameters. He conducted four formal trials with five-fold cross validation for each dataset and anomaly detector, which generated a total of twenty AUC results per detector algorithm and dataset against a held-out 20% validation slices. The mean and standard deviation of each dataset-detector combination are presented in Figure 3.2 as percentages. For each of the six detectors, he performed a pairwise Wilcoxon rank-sum test of significance and highlighted top performing algorithms, using a significance threshold of 5% [15].

In Figure 3.2, we can find that a relatively good result can be obtained by using the NSA and Density-based approaches on 5 standard benchmark datasets. And NSA performs best in real-world data.

Chapter 4

Discussion and Conclusion

According to the previous introduction and experiments, we can find that different anomaly detection methods have their own advantages and disadvantages. For example:

- **One-class Classification**
  Because SVM struggles with high-dimensional data[9], and when the training data only has normal data, One-class Classification methods can not control the false alarm rate by selecting hyperparameters[17].

- **Density-based Approaches**
  The computational complexity is also high because of the need to traverse the data to calculate the distance, which is not suitable for online applications or for high-dimensional data. In addition, only anomalies can be found, not clusters, and manual tuning is required.

- **Autoencoder and generative methods**
  When there are more outliers in the training data, the model may not work particularly well, and what we want to do is unsupervised anomaly detection (using only normal data learning), so the training allows a small amount of outliers, but when the outliers account for a relatively large amount, Autoencoder may overfit (learn the anomaly pattern).

- **Negative Sampling Algorithm**
  As shown in Figure 2.5, False positive occurs when taking positive samples, and the probability is particularly high in low dimensions. Moreover, this Algorithm is currently only applicable to detect point anomalies, while in real life, most data streams are time dependent, in other words, time series data, where the probability of contextual anomalies is high. So how to combine this algorithm with time series in the future is an important work.

Thus, when these methods are actually put into real life, we need to consider the advantages and disadvantages of each method as well as our own actual situation before choosing the right one.

In paper[15], Sipple shows that NSNN(Negative Sampling Neural Networks) is already being used in the real world. In 2019, actively monitors over 15000 power and climate control devices installed in 145 office buildings in Google. Each NSNN instance is associated with a

single cohort and periodically retrains a model over a sliding historical window to adapt to seasonal changes, and predicts an anomaly score to each new state vector. And he used Integrated Gradient to help technicians understand the anomalies by assigning a proportional Blame to individual dimensions. Since it is going in 2019, over 44% of all device-level anomalies result in calling technicians support. However, these are not enough to show that he has been good enough, because he has only experimented in Google, and different companies may have different equipment and produce different data.

Finally, anomaly interpretation is particularly important for anomaly detection, especially for devices. Because there is no point in anomaly detection if the anomaly is found but not resolved. Therefore, in anomaly detection, it is necessary to choose a suitable anomaly detection method as well as to find a suitable anomaly interpretation method in order to finally deal with the anomaly and make the devices operate normally and work smoothly.

List of Figures

Bibliography

[1] Aggarwal, Charu C: *An introduction to outlier analysis.* In *Outlier analysis*, pages 1–34. Springer, 2017.

[2] Akcay, Samet, Amir Atapour-Abarghouei and Toby P Breckon: *Ganomaly: Semi-supervised anomaly detection via adversarial training.* In *Asian conference on computer vision*, pages 622–637. Springer, 2018.

[3] Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller and Wojciech Samek: *On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.* PloS one, 10(7):e0130140, 2015.

[4] Böhle, Moritz, Fabian Eitel, Martin Weygandt and Kerstin Ritter: *Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification.* Frontiers in aging neuroscience, page 194, 2019.

[5] Chandola, Varun, Arindam Banerjee and Vipin Kumar: *Anomaly detection: A survey.* ACM computing surveys (CSUR), 41(3):1–58, 2009.

[6] Forrest, Stephanie, Alan S Perelson, Lawrence Allen and Rajesh Cherukuri: *Self-nonself discrimination in a computer.* In *Proceedings of 1994 IEEE computer society symposium on research in security and privacy*, pages 202–212. Ieee, 1994.

[7] Gonzalez, Fabio, Dipankar Dasgupta and Robert Kozma: *Combining negative selection and classification techniques for anomaly detection.* In *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*, volume 1, pages 705–710. IEEE, 2002.

[8] Hawkins, Simon, Hongxing He, Graham Williams and Rohan Baxter: *Outlier detection using replicator neural networks.* In *International Conference on Data Warehousing and Knowledge Discovery*, pages 170–180. Springer, 2002.

[9] Huang, Fu Jie and Yann LeCun: *Large-scale learning with svm and convolutional for generic object categorization.* In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 284–291. IEEE, 2006.

[10] Lee, HB, J Dy and A Krause: *Proceedings of the 35th International Conference on Machine Learning, Volume 80 of Proceedings of Machine Learning Research*. 2018.

[11] Liu, Fei Tony, Kai Ming Ting and Zhi-Hua Zhou: *Isolation forest*. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[12] Malhotra, Pankaj, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal and Gautam Shroff: *LSTM-based encoder-decoder for multi-sensor anomaly detection*. arXiv preprint arXiv:1607.00148, 2016.

[13] Samek, Wojciech, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin and Klaus-Robert Müller: *Evaluating the visualization of what a deep neural network has learned*. IEEE transactions on neural networks and learning systems, 28(11):2660–2673, 2016.

[14] Schölkopf, Bernhard, John C Platt, John Shawe-Taylor, Alex J Smola and Robert C Williamson: *Estimating the support of a high-dimensional distribution*. Neural computation, 13(7):1443–1471, 2001.

[15] Sipple, John: *Interpretable, multidimensional, multimodal anomaly detection with negative sampling for detection of device failure*. In *International Conference on Machine Learning*, pages 9016–9025. PMLR, 2020.

[16] Song, Hongchao, Zhuqing Jiang, Aidong Men and Bo Yang: *A hybrid semi-supervised anomaly detection model for high-dimensional data*. Computational intelligence and neuroscience, 2017, 2017.

[17] Zhang, Chunkai and Yingyang Chen: *Time series anomaly detection with variational autoencoders*. arXiv preprint arXiv:1907.01702, 2019.

[18] Zong, Bo, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho and Haifeng Chen: *Deep autoencoding gaussian mixture model for unsupervised anomaly detection*. In *International Conference on Learning Representations*, 2018.