Anomaly detection in multivariate temporal data: a cross-borders analysis.

Rui Maia, *Instituto Superior Técnico, Portugal* Cláudia Antunes, *Instituto Superior Técnico, Portugal*

Abstract—A growing number of data collection systems have been developed, leveraging multiple research fields and applications to explore datasets with millions of samples. From cibersecurity to life sciences, anomaly detection in temporal data is considered crucial as it often incorporates critical information unveiling events such as cyber attacks or helping with cancer early detection or heart-diseases prevention. Temporal data is commonly studied from an univariate perspective where only one data dimension can change over time. Few works have been dedicated to multivariate temporal data anomaly detection. In this work we provide a comprehensive and structured analysis of the main definitions, methods and approaches focusing on multivariate anomaly detection. Starting from basic methods experimentation and state-of-art approaches, we propose to research and contribute with new multivariate temporal data anomaly detection methods that can deal with variable series length with millions af samples, presence of noise and multiple feature categories, either static or dynamic, real or categorical valued. We experiment over maritime vessel tracks datasets, exploring unusual behaviour, including dynamic contextual data. Finally we plan to generalize lessons to other application domains such as car traffic behaviour.

Index Terms—Multivariate, Time-Series, Symbolic, Anomaly, Outlier

1 Introduction

We start by describing this work'thesis motiv 1.1 regardingon multivariattemporal data anomaly detection. The application domain and the problem statement are presented in the next section. Fundamental concepts regarding anomaly detection definitions and main approaches are included in section 2. A comprehensive related research work analysis is done in section 3 and sub chapters are dedicated to different approach types. Finally we analyze the most common evaluation metrics in unsupervised anomaly detection context and draw high level remarks.

1.1 Motivation

The behaviour study of temporal data have been crossing multiple research fields in the last decades. The unusual behaviours of isolated samples or groups of samples are commonly designated as anomalies or outliers, and have early been described by Frank Grubbs in 1969 [1] as observations (or samples) that deviate markedly from other members of the series where they occur. These anomalies often reveal significant information in their specific domains and should be carefully analysed.

Anomalies can be divided in two main groups: (1) the univariate anomalies, occurring in datasets where each sample is described by only one dimension, and (2) multivariate anomalies concerning samples that can vary on multiple dimensions simultaneously.

Extensive overviews on univariate anomalies detection methods and on their application domains have been provided by Hodge et al. [2], Chandola et al. [3], Aggarwal et al. [4] and Gupta et al. [5].

Other domains or approach specific reviews were published, namely by Zhang et al. [6] covering time-series datasets obtained using networks of sensors; Ahmed et al. [7] and Kwon et al. [8] published overviews on anomaly detection for network intrusion and cyber attacks detection leveraging deep learning approaches; Fanaee-T et Gama [9] centered their analysis on tensor-based anomaly detection techniques and finally Zhen [10] published an extensive overview on trajectory data mining, targeting land moving objects and persons.

Despite all these relevant contributes only a single and brief overview was dedicated specifically to multivariate temporal data anomaly detection, the one published by Tsay et al. [11]. This work highlighted from their statistical perspective the differences between four types of univariate anomalies and generalized the analysis to multivariate anomalies.

This work is though firstly motivated by the lack of research focusing specifically anomaly detection in large multivariate temporal data datasets and the need for updated and focused overviews that can specifically contribute for the data science perspective. Table 1 compares this work against previous overviews underlining that none of previous have specifically targeted multivariate temporal data anomaly detection. Moreover most references in surveys do not explicate if they aim univariate or multivariate anomalies detection. The table is also organized regarding basic approach groups, some of them included the experimental phase of this work.

1.2 Problem Statement

Some well known anomaly detection applications aim the identification of unusual events in environmental monitoring sensors data, diagnosing automotive brake systems malfunctions, satellite communications interferences, new diseases migration patterns, early detection of cyber-security threats based on network traffic or even to anticipate finan-

TABLE 1: Comparison of our work and related most relevant generic surveys. The abbreviate column names correspond to: (1) Hodge et al. [2], (2) Chandola et al. [3], (3) Aggarwal et al. [4] and (4) Gupta et al. [5]

Mutlivariate TS Base Approach	(1) Hodge-2004	(2) Chandola-2009	(3) Aggarwal-2013	(4) Gupta-2013	Our Study
Distance	[12]	[12] [13] [15] [16]			[12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [27] [32] [34] [35] [33] [36] [37] [38]
Density				[15] [16] [17]	
Components				[18] [19] [20] [21] [22]	
Time					
Neural Networks					
Other				[33]	

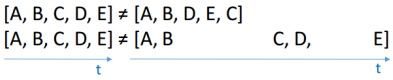


Fig. 1: Illustration of difference between time dependent observation sequences. Time variable may represent unusual temporal pattern in a list of observations, even other dimensions maintain expected values.

cial and economical crashes. Research works typically concentrate on a specific application domain, making domain dependent assumptions for data preprocessing, cleaning and imputation, but also for modeling, assuming previous knowledge about values normal distribution.

Each temporal data (or time-series) observation is associated with the moment in time where it was registered. This means that observations have in fact $n+1$ dimensions: n observation dimensions, associated to a moment in time when the observation was registered. The observation moment in time can be a crucial factor, which can be intuitively demonstrated by the differentiation between two sample sequences as illustrated in Figure 1. The first two lists are obviously different due to the different observations order. The second pair of lists are also different as a result of contrasting sample registration moments. Despite this basic notion most anomaly detection approaches do not incorporate meaningful time knowledge.

Series of samples that vary only in one dimension can be designated as **univariate temporal data**. These where first studied by Fox back in 1972 [39] however most nowadays anomaly detection research works still focus on univariate temporal data [40] supporting the tendency of using by adaptation univariate approaches in multivariate datasets.

Multivariate temporal datasets have multiple dimensions varying at the same moment in time as illustrated in Figure 2. Each dimension might vary in dependent or independent way concerning other dimensions.

A model that correctly represents a dataset regarding the expected multivariate temporal data behaviour is intuitively connected to accurate anomaly detection, since anomalies are unexpected observations contrasting with the normal behaviour. Nonetheless authors have been pointing out that multivariate temporal data modeling and anomaly detection is still a complex and multi-faceted challenge [5].

Regarding novelty and anomaly detection, both try to figure out the subset of new or abnormal observations from the set of normal observations. Novelty detection approaches tend to learn and update the model with new

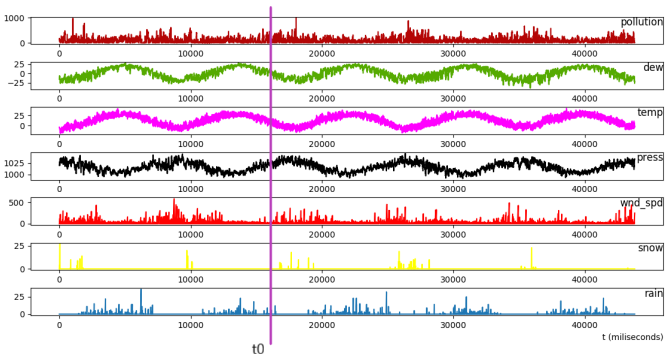


Fig. 2: Multivariate temporal data example. Each observation in a specific time moment is associated with multiple dimension values that vary independently of each others. In this example each observation is associated with seven dimension values.

observations while anomaly detection methods typically do not use outliers to update the base model. From other perspective, noise identification approaches try to identify and remove unwanted observations from the data, since they are considered as non relevant for the research domain. Despite different detection goals, either in anomaly, novelty or noise detection, research approaches commonly share methods and solutions [3]. Later in this work we further describe the relation between behaviour modeling, anomaly detection, novelty detection and noise identification.

Anomaly detection approaches heavily depend on the size and characteristics of the datasets [3] [4]. **Other multivariate temporal data complexities should be taken into account when defining this research problem**:

- It is frequently assumed that the major part of datasets observations behaves normally, and just a small subset are anomalies. Although this may not always be the case. Domain knowledge is typically involved supporting preliminary assumptions about observation and dimension values distribution.
- Most time dependent datasets are unlabeled, meaning that observations are not labeled as normal or abnormal. Being so, most of the available datasets do not directly support supervised learning approaches rather they support unsupervised ones.
- Anomaly detection methods have to deal with resource intensive computational tasks specifically regarding contexts were data is being continuously captured by multiple sensors.
- Temporal data may require online processing methods considering that anomaly detection is frequently associated to critical systems faults, life support systems or cyber threat detection, for example.
- Differences between observation novelties, anomalies and noise can be hard to defined or distinguish either conceptually and computationally.
- There is the need to research and develop new time dependent anomaly detection methods that can integrate heterogeneous datasets were very different types of information are registered (for example, weather conditions are registered in zones, in very unbalanced

data quantity distribution while vessels positions can be either frequent or infrequent).

2 Fundamental Concepts

This chapter introduces fundamental concepts for the understanding of anomaly detection in multivariate temporal data. Different learning and detection modes are discussed. The basic concept of anomaly is reviewed and defined for the remaining research work. Each different type is listed and detailed. All presented fundamental concepts are supported by relevant research works were different authors propose slightly different definitions either on univariate or multivariate anomaly detection contexts. This chapter closes with a summarized description of the expected outputs commonly found on published research papers.

2.1 Anomaly Detection Generic Framework

Anomaly detection approaches heavily depend on datasets characteristics, from observations frequency, feature space dimension to the overall dataset size (i.e. number of registered observations). Independently of these characteristics the main goal of an anomaly detection method is to identify a typically small group of observations or observations sequences considering as a preliminary assumption that the majority of the samples are normal [5].

Most anomaly detection methods can divided in two main steps: model building and anomaly identification. Regarding the (1) first step, a model can be trained and built using previously annotated data - enabling *supervised* learning approaches - although the chance is that it will be built upon non annotated or only partially annotated data, allowing either *unsupervised* or *semi-supervised* approaches. Differences between these are summarized in subsection 2.2. The (2) second step is heavily dependent on the choice of the approach type (made in the previous step).

Approaches are also characterized by their ability to focus on the complete sequences (or series) or subsequences of multivariate temporal observations. Anomaly detection procedures can therefore need to split data onto multiple subsequences also found as windows, motifs or fragments.

A multivariate temporal data sequence can be interpreted as an array of univariate time dependent series, all having equal length. Due to the complexity of multivariate temporal data anomaly detection, multiple authors search for anomalies in calculating anomaly scores in univariate series, after decomposing a multivariate into a set ou univariate series. Finally they compute an anomaly score for the complete multivariate one. Also in order to avoid multivariate complexity, other authors prefer to test only a carefully selected set of features (or dimensions) which requires a preliminary feature selection process.

Regarding spatio-temporal anomaly detection field, most approaches start by finding spatial anomalies in order to produce a subset of sequences to analyze. These are then verified over other dimensions. Besides the complexity of anomaly detection in multivariate time dependent contexts, additional challenges have to be dealt with when defining research approaches: multivariate temporal data including spatio dimension is commonly captured using different and

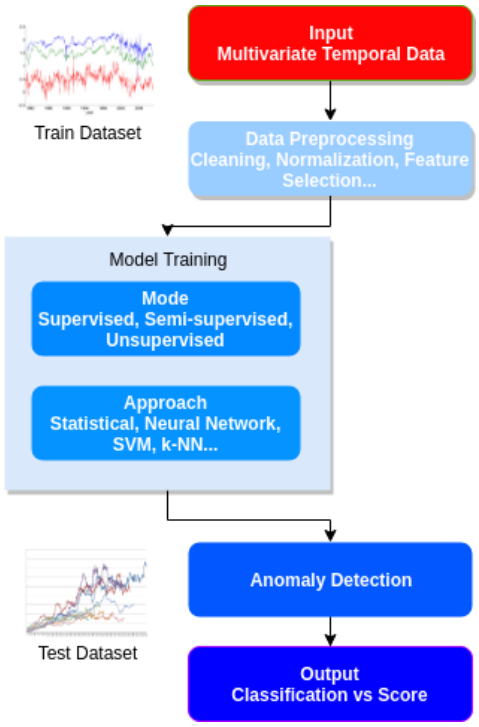


Fig. 3: Multivariate Time-Series anomaly detection generic framework.

distributed sensors which affects series sizes or rates, for example. Sensors installed in vessels or cars, their properties and capabilities, their data generation methods or transmission failures, all introduce additional dynamic complexities that should to be taken into account when dealing with anomaly detection.

Several research fields have dedicated to anomaly detection, from statistics to machine learning, sharing high level approach steps. Most of them can be mapped into a three steps meta architecture: (1) data preprocessing, to clean the dataset or to normalize observations and series; (2) construction of a model describing the normal or expected behaviour of the series, and (3) finally the anomaly detection process considering unseen observations. Figure 3 illustrates this generic architecture.

The anomaly detection process starts with input data preprocessing, frequently including normalization and cleaning. Series length normalization is one of the common methods used in multiple approaches considering that most of them can only analyze same length data series. The removal of disposable observations or estimation and imputation of observations - by interpolation or averaging, for example - can also be frequently found in normalization phase.

With reference to spatio-temporal data series, relevant in the scope of this work, positions measured in Latitude and Longitude have decimal representations such as 38.898556, authors also apply value normalization procedures in order to be able to compare geographic positions. Positions generally rely on Global Positioning System (GPS) which accuracy varies on multiple factors such as satellites geometry, signal blockage or atmospheric conditions. Considering that 1 degree of Latitude is equivalent to $111,111$ meters

at Equador, and consequently the sixth decimal place in one degree represents approximately 11 centimeters, there is room for lowering values precision to fourth decimal digit while maintaining real accuracy. In such a context GPS positions normalization can be applied assuming an accepted error rate of 11 meters.

Another context were data normalization takes its place is the distributed sensors scenario. Communication failures [5] affect samples existence or quality, attracting researchers to data cleaning aiming to remove erroneous observations as they can affect model training depending on the recurrence and type of failures.

After the preprocessing phase a model representing the normal behaviour is acquired based on clean data. This will be finally used in classification or scoring phase over test data - normally a subset the main dataset - allowing each unseen sample to be marked as normal or abnormal (or either scored). Approaches that calculate scores typically compare these against a threshold that represents a boundary between a normal or abnormal behaviour.

Univariate and multivariate anomaly detection are intrinsically connected, pulling authors to adapt univariate based approaches to the multivariate case, for example calculating anomaly scores of each different feature or dimension, and finally computing the final anomaly score for the multivariate series. Other authors prefer to test only a subset of multivariate series features. Both perspectives are limited and suggest that multivariate temporal data anomaly detection has a full set of open challenges.

2.2 Learning Modes

Anomaly detection approaches are shaped by input dataset properties. The existence of labeled data, were a target variable or set of variables is associated with each input sample, defines if the model can be trained and built in supervised, unsupervised or semi-supervised mode. A machine learning approach extending these perspectives is reinforcement learning, although, this will not be covered in this study since it was not possible to find published papers on multivariate anomaly detection supported by this approach methods. Figure 4 illustrates the three approach categories analysed in this work.

Unsupervised Learning Approaches based on unsupervised learning do not require labeled data to define train and build their models. They don't need an association between each input sample - univariate or multivariate - and a classification or score regarding a target variable or group of variables. There is no structural distinction between train and test datasets, both are not labeled. To overcome the lack information regarding which samples are abnormal these approaches frequently assume that the majority of samples are normal, frequently requiring the definition of normal and abnormal samples proportion or percentage. Unsupervised approaches frequently use statistical, density or distance based calculations to define what is the normal behaviour of the series, and what is abnormal. Unsupervised approaches do not require labeled training data which makes them the most common group of the approach categories.

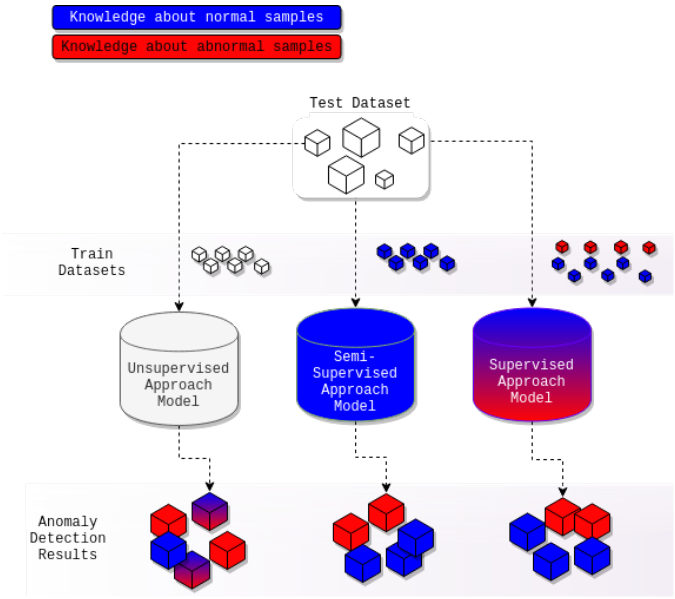


Fig. 4: Approach categories and model learning differences.

Semi-supervised Learning In this learning mode the train dataset contains only normal samples. Without any anomaly in observations the trained model will not contaminated by information from abnormal samples. Consequently the model will detect an abnormal sample whenever the deviation from the base model normal behaviour is above a certain threshold. These models can be typically found as "one-class" models. One of the major issues of semi-supervised approaches is that it does not distinguish between novelty and anomaly, regarding the former is a new sample behaviour that should influence model base definition (taking it to an update). By the contrary, an anomaly is a sample associated with a faulty or not acceptable behaviour that should be identified but not influence the base model definition. Semi-supervised approaches are more common than supervised ones. As well as unsupervised learning methods these do not require a labeled dataset for training.

Supervised Learning Training a model in supervised mode requires a labeled dataset for normal and abnormal classes. The model is then used to predict if an unseen input sample belongs to either normal or abnormal class. The training of such a model involves two main issues: (1) the number of abnormal samples in a dataset is typically significantly lower than the normal ones, which turns it difficult to build a model that can classify accurately both normal and abnormal samples. Moreover, some supervised approaches such as decision trees do not deal well with the highly unbalanced datasets. (2) To overcome the lack of abnormal labeled samples in datasets researchers tend to generate new abnormal samples. This is typically an ad-hoc process highly dependent from the application domain. Supervised learning methods for multivariate anomaly detection are less common regarding they require fully labeled train and test datasets.

2.3 Anomaly Definition

The definition of anomaly - as a class - has been revisited since 1969, when Frank Grubbs [1] identified it as unusual behaviours of isolated samples or groups of samples. Anomalies can also be frequently found as *outliers*, *exceptions*, *faults or discordant observation*s, depending on the research field and application domain. Typically in Computer Science, Health or Medicine research, community tend to use *Anomaly* or Outlier terms to identify an abnormal behaviour of a sample or group of samples. In other research areas, such as Mechanical or Automotive for example, authors tend to use the *Fault* term since they are frequently investigating engine or sensor faults using multivariate temporal datasets. Although *Discordant Observations* can also be found in multiple contexts, it is mainly associated with univariate data anomaly detection, which is not in the scope of this work.

Chandola et al. [3] defined *Anomaly* as a pattern that do not follows an expected behaviour. This generic definition can be intuitively illustrated in one, two or three dimensional spaces, as illustrated in Figures 5 and 6. Anomalies occurring in bigger dimensional spaces will escape human capacity visualization capabilities, resulting in human intuition loss.

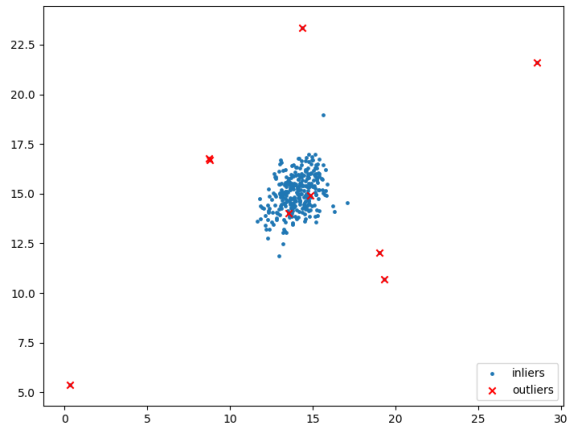


Fig. 5: Two dimensional anomalies identified by red crosses.

Aggarwal and Yu [41] defined *Anomaly* as an observation which is very different from the dataset behaviour considering a specific measure calculation. Authors considered that this observation has useful information for the system comprehension. In their work noise was also defined, which can also be clustered or identified.

This work adopts these base definitions, from Chandola et al. [3] and Aggarwal and Yu [41], considering their complementarity.

2.4 Anomaly Types

The anomaly detection overview by Chandola et al. [3] identified three main types of anomalies: Observation, Context, Sequence and Collective.

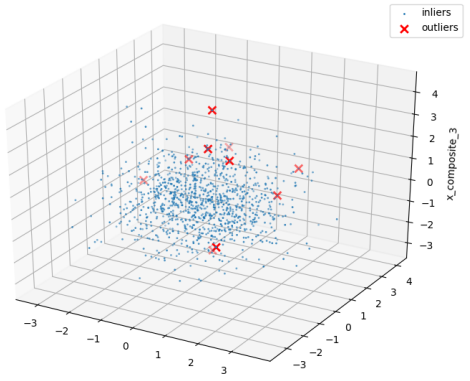


Fig. 6: Three dimensional anomalies identified by red crosses.

Observation Anomaly - An observation is considered abnormal when one or more dimension values - categorical or real valued - are outside the expected intervals or sets (as illustrated in Figure 5). This is the simplest type of multivariate temporal anomaly and derives directly from the univariate concept.

Unidimensional or Multidimensional - Depending on the identified abnormal values this can be a Unidimensional or Multidimensional anomaly.

> **Practical Example** - A vessel track may be considered abnormal if the registered speed is very high compared the average vessel speed, for that specific type of vessel. If all other parameters are considered as normal, this would be an Unidimensional Observation anomaly.

Contextual Anomaly - An observation can be considered abnormal despite all dimension values are according the expected intervals or sets. Song et al. [42] introduced a statistical conditional anomaly detection approach by defining each observation as a pair of attribute sets (x, y) where x is the set of contextual dimension values and y is the set of indicator dimensions values (see Figure 7). An observation is anomalous if y values are not expected considering x values. Typically multivariate temporal data dimensions are not classified as Contextual or Indicator, and this classification is not easy to define [3] as it strongly depends on the problem domain.

> **Practical Example** - Contextual dimensions can be vessel latitude and longitude in a spatial dataset or transmission time in a time dependent domain. The vessel speed can be the dimensional indicator. This anomaly type relies on the separation between contextual and indicator dimensions which in turn essentially depends on the application domain.

Sequence Anomaly - A sequence of observations is considered abnormal when, even if each observation dimension values are normal, the relation between observations is unexpected as illustrated on Figure 8. The highlighted sequence express an anomaly because mul-

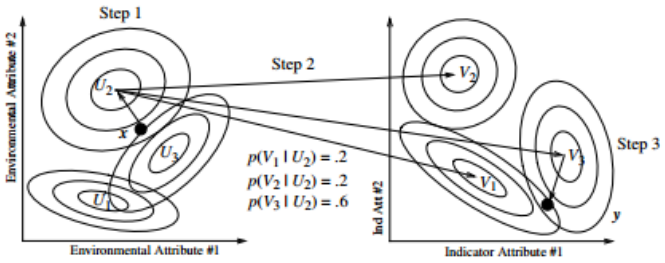


Fig. 7: Conditional Anomaly Detection as presented by Song et al. [42]. Observations are conceptually represented in two different meta-dimensions, Contextual and Indicator. Each of these meta-dimensions is associated with a sub-set of the observations dimensions. More intuitively, each contextual attribute limits the normal space for all the indicator variables.

tiple observations were detected earlier than expected (Premature Atrial Contraction) in a normal heartbeat pattern.

Fig. 8: A Premature Atrial Contraction (PAC) is an irregular heartbeat pattern that can be classified as a Multivariate Sequence Anomaly since although values are not abnormal the relation between data dimensions is found unexpected.

Practical Example - None of the observation dimensions in Figure 8 is considered abnormal by itself. This anomaly depends on the existence of a dimension that relates observations, for example *time*, in multivariate temporal data. Chandola et al [3] considering Song et al. [42] definition, noting that Sequence anomalies can be observed as Context anomalies if the context is defined by one or more time related features.

Collective Anomaly - Searching for collective anomalies involves exploring the hidden structure of data for abnormal relations between observations or sequences of observations. Each multivariate series behaviour is not necessarily an individual, sequence neither context anomaly. In fact there will be no series anomalies but a change in the dataset behaviour, considering multiples series in the same temporal reference. It is common to find Collective anomalies being defined as previously Sequence or Context anomalies, but this would exclude Collective anomalies as defined in this work, i.e. from the collective behaviour perspective.

Practical Example - Two vessels behaving normally from observation and sequence perspectives, can change the relation between them. If they agree to exchange illegal products during their normal missions they can still maintain normal track parameters (such as speed or GPS position, for example) and they will also behave normally according to their seasonal or temporal context, the admitted regions and the expected the route. Although, these vessels might have a change in the relation between their temporal series normal behaviour, which might indicate they are orchestrating irregular activities.

2.5 Anomaly Detection Outputs

In multivariate temporal data, like in the univariate case, an anomaly detection process tries to identify the unexpected observation or sequences of observations. Chandola et al. [3] identified two output types: labeling and scoring.

These two types are typically found as sufficient for univariate and multivariate anomaly identification but there is the need to extend the interpretation to the multivariate case, specifically the Collective anomaly. Here different series may be considered normal but their inter relation may be found abnormal.

Label - Anomaly detection approaches classify one observation or sequence as normal or anomaly. It is a binary classification process taking generally less computational complexity [7] than the score calculation.

Score - An anomaly score is calculated for each observation or sequence. This value can be used to rank observations and sequences according to their odds of being abnormal while it can be compared against a threshold that distinguishes between normal ou abnormal values.

Regarding Collective anomalies, this work assumes that observations or sequences of observations belonging to different data series can - independently of their individual series abnormal factor - be characterized by a specific Score or Label describing normal or abnormal relation according the normal behaviour.

3 Related Work

Anomalies or outliers in temporal data can be seen as samples whose values deviate considering expected values. Univariate time series anomalies were first studied by Fox in 1972 [39] and even today most time-series evaluation, comparison and anomaly detection methods continue to focus on univariate time series [40]. These fact may justify the common practice of using univariate detection techniques in multivariate time-series. Figure 9 illustrates one univariate and one multivariate anomaly on an multivariate temporal data.

Regarding large and different data series sizes, most anomaly detection approaches choose to segment series analysis using sliding windows [43] and offline processing [5]. This means that the window length is a determinant feature in different approaches. Dimensionality reduction and feature selection techniques are the basis of different approaches, although some authors recently underlined the

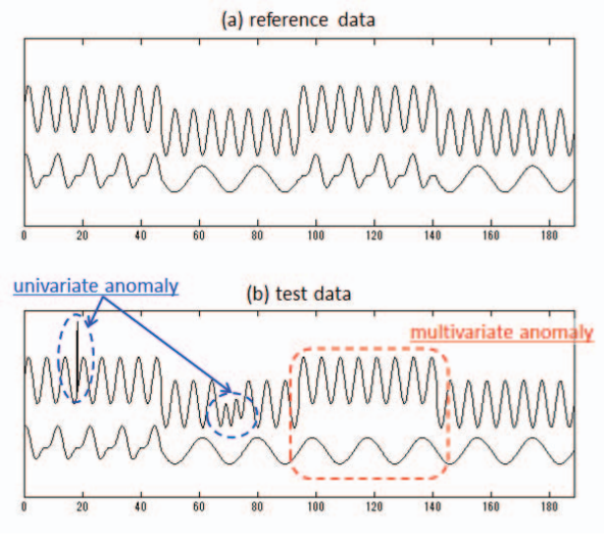


Fig. 9: Time-Series anomalies example. One univariate anomaly marked with blue color, and one multivariate anomaly in red as illustrated by Takeishi and Yairi [35].

possible correlation variation between features in dynamic multivariate temporal data [44].

Five main types of univariate anomalies can be found in literature: (1) Additive, affecting only one sample, (2) Innovational, changing the time series behaviour, (3) Level Shift, changing abruptly and permanently the trend of the time series, (4) Temporary Change, changing the behaviour of a time series for a period of time, and finally (5) Seasonal Level Shift, changing the behaviour of the time series in cyclic manner. There is no similar work of classification on multivariate anomalies.

The presented five types of univariate anomalies can affect each series of multivariate series. It is worth mentioning that multivariate anomalies can also be caused by relation changing between its individual series, rather than unexpected values or independent anomalies. This means that even if all series of a multivariate temporal data series are normal when considered individually there could anomalies in multivariate perspective as long as the relation between series is different from the normal behaviour. It seems therefore reasonable to conclude that the use of univariate anomaly detection methods to multivariate data do not to cover the more complex task of multivariate anomaly detection.

Figure 10 illustrates a simplified view over anomaly detection approaches. They can be analysed in three base perspectives: (1) from the Base Approach perspective, (2) the Training Mode and (3) the detection goal. Next sections will describe anomaly detection reference approaches for each of the identified families.

3.1 Statistical Approaches

First time-series outliers detection approaches came from statistics field and are still frequently adopted in data science research. They try to model time-series probability distributions in order to classify observations as normal or abnormal or to calculate an anomaly score. These approaches are generally applicable to datasets with quantitative ordinal observations, some of them requiring relevant preprocessing and classification procedures [2]. The involved computational complexity in statistical approaches is considered to be generally high.

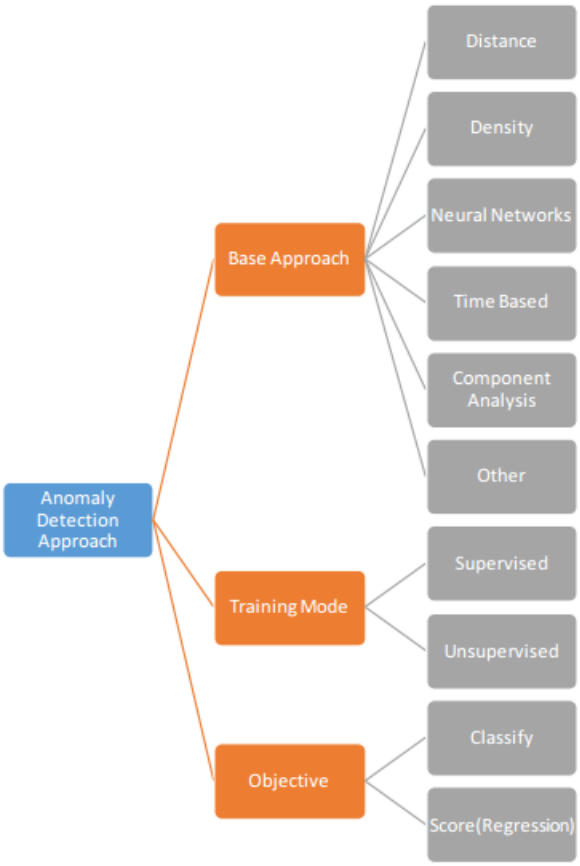


Fig. 10: Anomaly Detection approach categories, training mode and output types.

Parametric statistical approaches involve prior knowledge on the series behaviour so that the correct observation probability distribution is applied [45]. This approaches rely on having prior knowledge on the expected behaviour of the series, so that a specific probabilist model can be chosen - a Gaussian distribution, for example - and the correct parameters estimated for the training dataset. Different techniques such as Maximum Likelihood Estimatation (MLE) can be applied in parameters estimation.

By the contrary, non-parametric statistical approaches require no previous behaviour knowledge. They try to infer a statistical model from the given training data. Relevant approaches from statistics field include Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) and adaptations to online multivariate processing [46] [47], Vector Autoregression (VARMA), and CUmulative SUM Statistics (CUSUM).

3.2 Distance Based Approaches

One the most well-know distance based methods is the k-Nearest Neighbor (or k-NN). It uses a measure - commonly Euclidean or Mahalanobis distance - to calculate distances

between observations. It is a supervised learning algorithm that classifies each new observation according to the minor distance to k neighbors. k-NN classification was proposed by Cover et al. [48] in 1967, and continues to be tested and extended by many authors. Ramaswamy et al. [12] proposed to rank potential outliers based on the distance of each point from its k^th nearest neighbor. The raking top n observations are classified as outliers. Considering the model requires the entire distance matrix must be calculated, the authors also proposed data partitioning for faster processing, clustering points in *cells*, in order to minimize computational complexity. Knorr et al. [13] proposed another distance based method where an observation o is classified as an outlier if there are less than k points within a pre-specified distance R. The authors considered that an outlier in a multi-dimensional space (or multivariate series) is an observation that has at least one outlier dimension behaviour. They experimented their approach in three scenarios including one very small [13] dataset of people trajectories (from a video-surveillance system). They summarized each complete trajectories as a single matrix (as described on Equation 1).

$$T = \begin{bmatrix} T_{start}(x,y) \\ T_{end}(x,y) \\ T_{heading}(average, minimum, \\ andmaximumheading) \\ T_{velocity}(average, minimum, \\ andmaximumvelocity) \end{bmatrix} \quad (1)$$

Distance between trajectories is calculated using Euclidean distances of each trajectory feature. For example, the distance between $TA_start(x,y)$ and $TB_start(x,y)$, being TA, TB different trajectories, is given by the Euclidean distance between points axes. The model depends on a feature weight vector w whose values are determined by domain experts. The global distance between trajectories can be represented as defined in Equation 2. In the published research, authors only used one feature or pair of features in their distance function to determine outliers, namely: velocity, velocity and/or heading, and start and end positions. This can be considered a univariate approach applied to a multivariate time-series, also not taking into account the temporal feature or relations.

$$D(TA,TB) = \begin{bmatrix} Dstart(P1;P2) \\ Dend(P1;P2) \\ Dheading(P1;P2) \\ Dvelocity(P1;P2) \end{bmatrix} \quad (2)$$

$$* \\ [w_{start}, wend, wheading, wvelocity]$$

Jankov et al. [14] presented a real time anomaly detection system for big data sensor streams. The work, proposed in the scope of the 7th ACM Distributed and Event-Based Systems (DEBS) Grand Challenge - 2017 - was based on sequential three step workflow: (1) the approach starts by clustering the multivariate sensor series using k-Means. (2) The second step uses the found clusters to train a Markov Model that describes the transitions between the predefined number of clusters. Finally (3), for each sliding window on analysis, the approach calculates state transitions probabilities for the current samples.

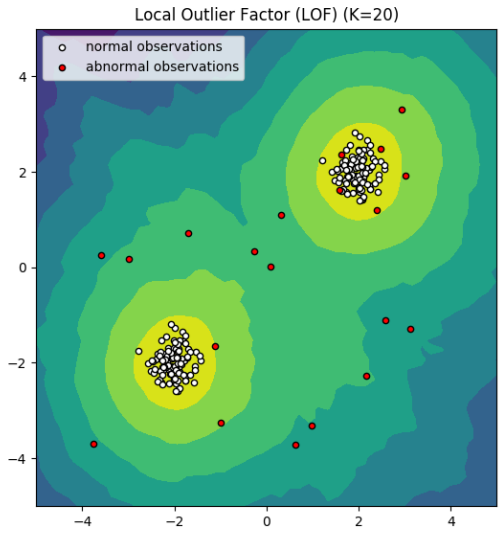


Fig. 11: LOF method applied considering k=20 k^th nearest neighbors.

3.3 Density Based Approaches

Density based approaches try to distinguish outliers based on the dataset density distribution. Breunig et al. [15] considered that the probability of an observation being an outlier should have an associated a degree (a real value), instead of a binary classification. For that they introduced the Local Outlier Factor (LOF) as a degree of how isolated an observation is in the dataset, considering the n^th most near neighbors as illustrated in Figure 11.

The concept of Local Reachability Density (LRD) was introduced by the authors in order to stabilize statistical fluctuations of the method, by considering that if two observations are sufficiently close, their distance will be considered as $k-distance$ (see Figure 12). Otherwise, the distance will be the actual distance. LOF Values similar to 1 indicate similar density, higher than thus a probably normal observation. LOF values lower than 1 indicate outliers. There is no possibility to know *a priori* what is the rational for the inlier and outlier thresholds, which strongly depends on dataset characteristics. In the same way, the k^th neighbors to consider also has to be experimentally defined, to tune the calculation of LOF and LRD values. Breunig et al. [15] experimented LOF approach in a 64 dimensional dataset of tv snapshots color histograms. Pokrajac et al. [16] further developed this work by incrementally calculating a new observation LOF and LRD values as soon as it is inserted in the dataset (in a data streaming context). The observation can be immediately classified as normal or outlier and LOF and LRF values of previously existing observations can be updated if needed.

Zhange et al. [17] proposed another density based approach named Stream Projected Outlier deTector (SPOT). Aiming at detecting outlying subspaces in high-dimensional

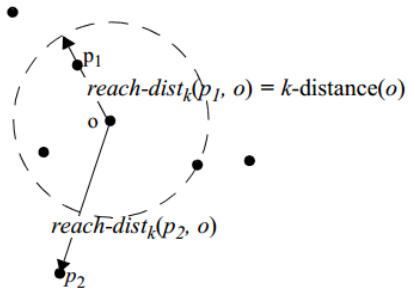


Fig. 12: Local Reachability Distance (LRD) as illustrated by Breunig et al. [15]. If objects p and o are far away from each other the reachability distance between the two - LRD - is their actual distance (point $p2$). If they are at a calculated range (depending on the neighbors) the considered distance is k-distance of the point. The LRD of the three closest points to o is this the same.

data streams, the authors proposed a window based method where cell summaries are updated as highly dimensional data arrives. The proposed cell summaries - Base Cell Summary (BCS) and Projected Cell Summary (PCS) - statistically describe the current window state by including simple measure such as Relative Density and Inverse Relative Standard Deviation as defined on 3 and 4.

$$BCS(c) = D_c, LS_c, SSc \tag{3}$$

$$PCS(c, s) = RD, IRSD \tag{4}$$

Specifically, D_c, LS_c and SS_c are respectively the number of points in a cell c and the sum and squared sum of each point dimension values while the point is in c. RD and $IRSD$ the Relative Density and Inverse Relative Standard Deviation of all points existing in c considering a subspace s of an hypercube representation. Hypercubes, also found as n-cube or n-dimensional cube, can be seen as a generalization of a three dimensional cube. They are used to represent data in multiple dimensions, typically more than three. An n-cube of d dimensions has n points, where $2^d = n$, therefore, an hypercube representing data in 4 dimensions will have 14 vertices. SPOT approach calculate fixed, clustering and outlier based Sparse Subspace Templates (SSTs) either through unsupervised or supervised learning, using cells BCS and PCS summaries. These templates are finally used in a Multi-Obejctive Generic Algorithm (MOGA) to find outlying subspaces. In the detection stage, a point is classified as outlier if it belongs to one of the cell supspaces outlying SST subspaces.

Kanamori et al. [49], Sugiyama et al. [50] and Hido et al. [51], from the statistics research domain, also referred the possibility of anomaly or novelty detection by the use of density functions. The density score, which some designate as *importance*, is directly related to the probability of sample being abnormal.

3.4 Components Analysis Approaches

Principal Component Analysis (PCA) [52], Singular Value Decompostion (SVD), Independent Component Analysis

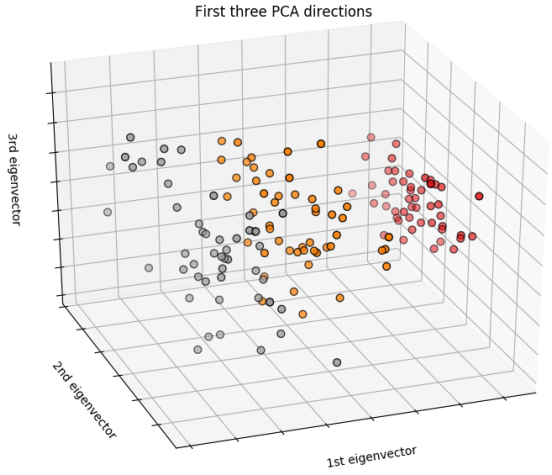


Fig. 13: Observations found in a four dimensional dataset are here represented using three principal components, capturing the most relevant information, i.e., the vectors with the higher variance that can better explain dataset behaviour.

(ICA) or Support Vector Machines (SVM), are dimensionality reduction methods which can be used to represent a multi variable dataset onto few uncorrelated coordinates while retaining data variability. Principal Component Analysis (PCA), proposed by Wold et al. [52], extracts the dominant or latent components by representing time-series variables using a predefined and limited number of vectors, the principal components or eigenvectors (in Figure 13 a multivariate time-series is represented using only three dimensions).

Drosdowsky et al. [18] [19], Lu et al. [20] and Lasaponara [21] used PCA based approaches to identify abnormal behaviours in different environmental or natural areas. Either to discover Australian different areas such as drought areas, deep fertility loss areas and hurricane areas [18] [19] or to evaluate vegetation interannual anomalies [20], PCA was applied to enhance abnormal regions of multivariate datasets.

Aiming anomaly detection in scenarios of massive moving objects - specifically vessels in the experimented dataset - Li et al. [22] - proposed a motif-based representation of vessel tracks. Each motif includes temporal and non-temporal features. After the representation of the tracks as motifs, a Support Vector Machine based approach classifies each vessel track as normal or abnormal. Each vessel path P is an ordered sequence of points $p_1, p_2, ..., p_n$ where each point is associated with a time-stamp, a latitude-longitude position and a set of non spatiotemporal attributes such as vessel length, heading or flag. To represent each a sequence of movement motifs is extracted, based on a predefined prototypical representation of vessels movements. This predefined set of motifs can be described as the known vessel operations which were previously defined by experts (eg. straigh line, right turn, loop, ...). Thereby, each track representation is transformed onto a sequence of motifs - from the predefined set - with a starting

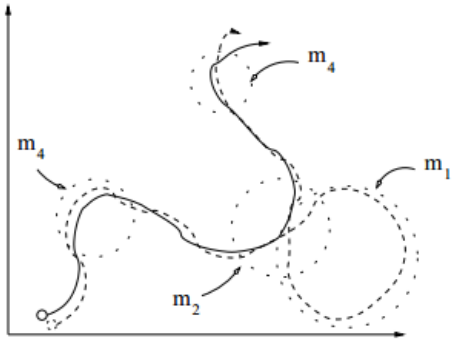


Fig. 14: Two vessel tracks are represented and the motifs identified as presented by Li et al. [22]. Both vessels share the same $1xm_2$ and $2xm_4$ motifs, except an extra m_1 in the dashed track. The small differences in vessels positions are considered as semantically irrelevant.

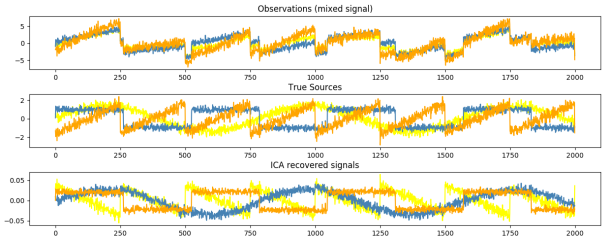


Fig. 15: A time-series is decomposed onto three independent sources capturing the most relevant information. ICA represents the original series using high variance components that can explain dataset behaviour.

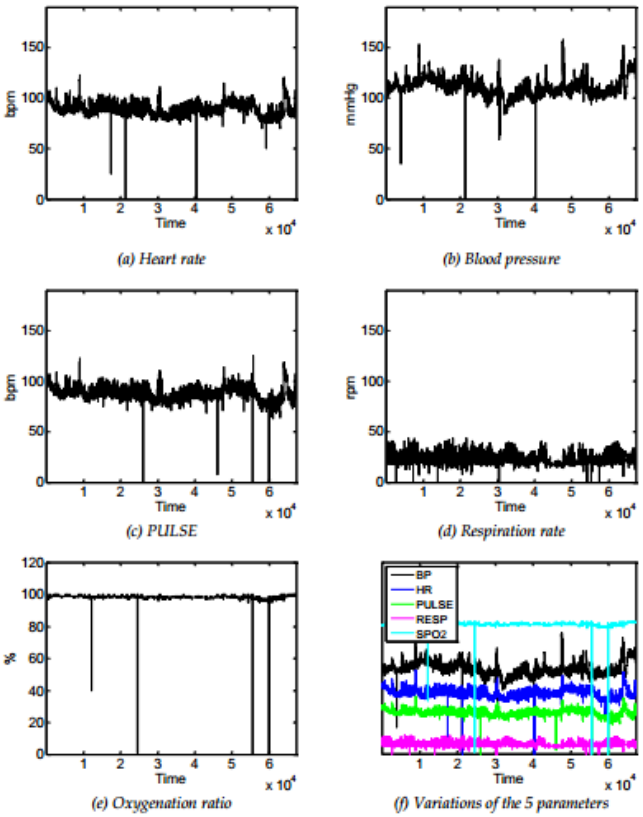


Fig. 16: Multiple time-series present relevant patient information captured by Salem et al. [24]. Linear Regression is applied to predict values using the related time-series when abnormal sequences are detected.

and end time, as well as a start and end location $<(mot_i, t_{start}, t_{end}, loc_{start}, l_{end}), (mot_j, t_{start}, t_{end}, l_{start}, l_{end})>$.

Baragona and Battaglia [23] used Independent Component Analysis (ICA) to identify which components of an multivariate time-series contributed for an anomaly. ICA, also referred as blind-source separation, can be applied to decompose time-series into non-Gaussian (non-normal) components that are statistically independent (see Figure 15). Although the approach performed better regarding single components outliers detection and was considered has less relevant by the authors when in the presence of multiple component outliers. It also did not take into account the time dynamics of series, which is crucial in the scope of this work.

Salem et al. [24] proposed an SVM and Linear Regression based framework for anomaly detection in Wireless Body Area Network (WBAN) devices in health care context. WBNAs are miniaturized wireless sensors transmitting time-series data to external units such as laptops, mobile phones or tablets. These time-series are then analysed separately for anomaly detection, and jointly when the approach predicts, for abnormal detections, what would be the normal or expected values. More specifically, the approach analyzes irregular sensor measures to distinguish between faulty sensors or critical patient situations. Linear SVM is first used to detect abnormal sensor readings in each time-series. When an abnormal sequence is detected the approach reconstruct the expected values applying Linear Regression and the other time-series readings. It does not updates the SVM model, which reduces the complexity in online processing mode but may be considered a drawback regarding data update process.

In [25] Kane and Shiri propose to apply dimensionality reduction and correlation analysis of whole multivariate time-series sequences. The first step of their method in the representation of multivariate time-series as univariate time-series. Then, they apply Singular Values Decomposition (SVD) to uncover the most relevant latent features, their weights and eventual correlations in a multivariate series. The authors approach also requires the time-series to have the same number of samples, which requires data series pruning. They aimed at measure similarity between series using simple methods such as Pearson's product-momment coefficient.

On a different computer science perpective, Liu et Rebentrost [26] proposed an adaptation of SVM and PCA approaches to detect anomaly in quantum states, stating that quantum computation, simulation and communication, will tend to integrate anomaly detection methods.

3.5 Time Based Approaches

Shokoohi-Yekta et al. [28] analysed and compared the representation of multivariate time-series using two different Dynamic Time Warping(DTW) approaches: considering all dimensions of an multivariate time-series as independent - DTW_i - and considering that the dimensions are related - DTW_d. The authors proposed a method to verify which of the DTW approaches is more adequate in a specific context since the results are dataset dependent, even for datasets in the same domain. They studied the behaviour of DTW_i and DTW_d considering two time series Q and C when introducing lag and loose coupling effects and used the common definition of DTW_i as the cumulative distance of all independent time series $DTW_i(Q,C) = \sum_{m=1}^{k} DTW(Q_m, C_m)$, and DTW_d by redefining the DTW base cost function $d(q_i, c_i)$ as the cumulative squared Euclidean distances of k data points of each time series. Considering p_i, k as the i^th sample of the k^th dimension of Q and r_j, k in the same dimension of C time series, then the cost-function of DTW becomes $d(q_i, c_i) = \sum_{m=1}^{k}(q_{i,m} - c_{j,m})^2$.

Still in DTW approach field, Seto et al. [29] proposed to avoid feature extraction in multivariate time-series by hierarchically clustering series based on the distance between series sub-sequences. A template is then built for each cluster by averaging the correspondent set of series. Finally, for each test sample, a distance vector to each cluster template is calculated. Vectors are used as feature vectors in the final classification step, using Principal Component Analysis (PCA) and a Support Vector Machine (SVM) for classification.

To be able to use different distance measures the authors *z-normalized* every dimension of the analysed time series before computing DTW distances. Depending on the used approach the error might double, having non trivial interpretation on the factors that might lead to these differences. The authors also underlined that very different warping paths might result in identical DTW distances, which in the scope behaviour analysis and anomaly detection might be a problem.

3.6 Neural Networks Based Approaches

The frequent use of neural networks is due to neuroscience and computer science advances since 80's decade, where computing systems built upon large number of interconnected layers and cells were proposed to mimic the human brain structure and dynamic behaviour.

Samantha et al. [30] [31] [27] applied Feedforward Neural Networks in conjunction with Genetic Algorithms to classify the time-domain vibration series from a rotating machine as normal or fault. Authors used genetic algorithms to optimize feature selection from the input signals varying the number of selected features. Finally they compared the use of a Support Vector Machine (SVM) classifier against the results of a Feedforward Artificial Neural Network. This type of network propagates the input values directly to the hidden layers, were each connection between neurons has it's own weight (see Figure 17).

To overcome limitations of common multivariate temporal data anomaly detection approaches such as statistical, time window based techniques or Recurrent Neural

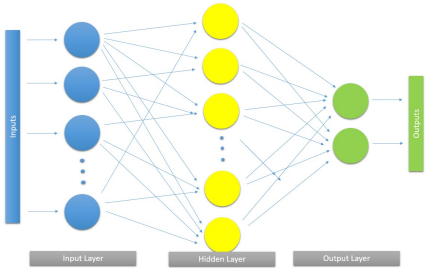


Fig. 17: Feedforward used by Samantha et al. [30] [31] [27] to model a rotation machine vibration Multivariate Time Series.

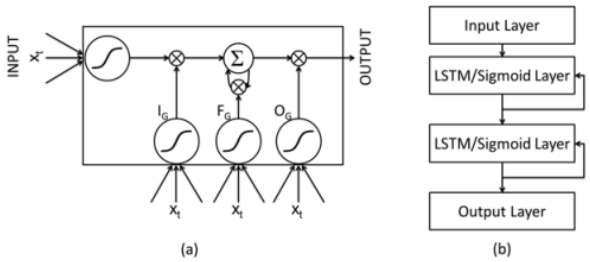


Fig. 18: Long-Short Term Memory cell stacked architecture used by Malhotra et al. [32] to model the normal behaviour of a Multivariate Time Series. The error distribution of the prediction is used to identify anomalies.

Networks (RNN), Malhotra et ak. [32] proposed a Long Short-Term Memory (LSTM) (see Figure 18) neural networks based approach. The authors aimed at learning time series long term correlations without the need of pre-specifying time windows. Simultaneously they tried to overcome the vanishing gradient problem experienced in RNNs. The authors experimented on four datasets: Electrocardiograms (ECGs), Space Shuttle Marotta valve time series, Power demand dataset and Multi-sensor engine data. The four datasets were generated by non independent sources. By the contrary this thesis proposal experiments on datasets having multiple independent sources of multivariate temporal data anomalies. Additionally this research have to deal with sources that may act together and in orchestrated form in order to camouflage individual or collective anomalies.

Recently Li et al. [53] [54] proposed a framework for multivariate anomaly detection in unsupervised contexts (MAD-GAN) using a Generative Adversarial Networks (GAN) [55] built upon two Long Short-Term Memory Neural Networks. They considered that current techniques may not be adequate in dynamic systems where sensors and behaviours change frequently. The approach tries to model multiple features correlation underlining spatial-temporal correlations which authors stated as being not fully exploited in current approaches. GAN are two collaborating networks where one - the generator - generates observations trying to mimic the reality (of a specific context or domain) while the other network - the discriminator - classifies each observation as normal ou abnormal. Malhotra et al. [32], Zhang et al. [56], Hundman et al. [57] and Song et al. [58] proposed to detected anomalies in multivariate temporal data through the use of deep neural networks including

Long Short-Term Memory based networks.

3.7 Other Approaches

Gupta et al. [34] analysed Apache Hadoop logs and hardware usage metrics. The authors combined multiple data sources in order to get what they identify as multivariate time-series datasets. They proposed a two-step clustering approach for anomaly detection based on PCA and a modified version of k-Means algorithm. First they calculate the *context clusters*, one of main concepts of the authors work. These clusters consists in partitioning multivariate time-series by context variables values, the ones related with the nature and number of Hadoop tasks. In a second step, the some metrics are calculated for each *context cluster*, using values extracted from machine performance logs (eg. CPU and memory usage, disk read operation,...). K-Means algorithm is applied to create both *context clusters* and *metric clusters*, being the last a template or centroid representation of multivariate time-series values of each series belonging to a *context cluster*. A new multivariate time-series instance is only considered as abnormal if a *context cluster* that matches this instance can be found but the *cluster metrics* is very far away of the instance metrics. The anomaly score of the instance is thus dependent on finding the correct *context cluster* and the distance between instance metrics and *cluster metrics*.

Takeishi and Yairi [35] proposed an anomaly detection method for multivariate time-series based on the identification of the most common small subsequences (the authors called them *Elemental patterns*). These patterns form a dictionary used to describe each multivariate time-series, were each one is described by a unique feature vector, were each vector element is a coefficient associated to a dictionary element (en *Elemental Pattern*).

In [5], Gupta et al. conduct an extensive analysis of outlier detection methods depending on the considered data types such as continuous series obtained by the use of sensors, discrete series (using web logs, for example), multi-dimensional streams as news text data feeds or network data such as social or computer networks streams. Although this extensive work mainly identify univariate time-series research approaches, it points out that multivariate time-series anomaly detection commonly use univariate anomaly detection techniques by smartly selecting multivariate time-series features rather then using the complete multivariate time-series feature set to compute outliers. Having analysed more than 160 research works, Gupta et al. only referenced Lakhina et al. [33] for using multivariate time-series to discover different computer networks data outlier events such as Denial of Service or Host-Scanning, for example.

Multivariate series in computer systems networks are used for traffic characterization and intrusion or attack detection. Tan et al. [36] proposed a new Multivariate Correlation Analysis (MCA) approach for Denial-of-Service (Dos) attacks detection. Authors explored the geometrical correlations between network series features to build Triangle Area Maps (TAM) that describe the relations between pairs of observation features and the observation. The detection process involves three main steps: (1) extract features from data, (2) build the normal profiles (TAMs) for training observations and finally (3) detect attacks using the TAM representing test observations.

Other anomaly detection perspectives have been applied leveraging information collected in growing networks of sensors. Kong et al. [37] proposed LoTAD, a Visual Assessment of Tendency (VAT) hierarchical clustering based method that processes the trajectories of vehicles and pedestrians in a parking lot. Low Rank and sparse representation approaches have been used to preserve global data structures in multivariate systems. Xu et al. [38] introduced anomaly detection in Hyperspectral Images (HSI) using low rank and sparse representation. HSIs are, in the most simple form, a three dimensional data type. However, the number of dimensions can greatly increase, requiring multivariate anomaly detection methods.

Li et al. [59] proposed to decompose multivariate series in univariate cases in order to apply a group of Hidden Markov Models while Goix et al. [60] explored the dependence structure of multivariate series for the detection of extreme events based on the theoretical foundations of the Extreme Value Theory applied to the multivariate context.

4 Evaluation Metrics

Different approaches can be found in literature to evaluate anomaly detection methods. These have to deal with different contexts such as very imbalanced datasets that invalidate the use of simple statistical performance results or the unsupervised and supervised nature of anomaly detection problems. Most of anomaly detection problems are characterized by very imbalanced classes. This means that the number of abnormal samples is much smaller than the normal ones, invalidating the use simple statistical evaluation methods such as accuracy. True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) are used to define Accuracy as shown in equation 5. Consider for example that only 2% of the vessels have abnormal behaviours. An anomaly detection method that classifies all tracks as normal would have 98% of accuracy, although it would be difficult to use this model as a vessel behaviour anomaly detection method.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

Most anomaly detection methods are not pure binary classifier, instead they start by scoring the probability of each input sample being an abnormal one. Thr function or method used for score calculation is named Scoring Function. The obtained score is then compared against a threshold for binary classification, identifying the input sample as normal or abnormal. When all input samples of the test dataset are processed the global performance of the detection method is evaluated. Precision and Recall (PR) and Receiver Operator Characteristic (ROC) curves are two common evaluation methods found in literature.

These two methods depend on true and false positive anomaly identifications. A True Positive identification occurs when an abnormal sample is correctly labeled as anomaly, while items incorrectly classified as anomalies are identified as False Positives. The same logic applies to the

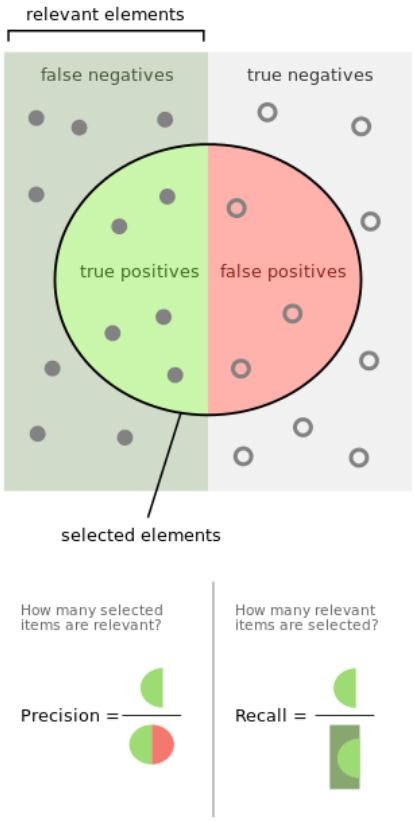


Fig. 19: Precision and Recall

Negative cases (non abnormal samples), as illustrated in Figure 19.

Precison can be described as the fraction of correctly identify anomalies - True Positives - considering the all classified anomalies - True Positives + False Positives, as described in equation 6. On other hand, Recall measures the ability to find all, or most, abnormal samples in the dataset, represented in equation 7. Intuitively, Precision represents the ability to find only the anomalies - having the least false alarms - while Recall describes the capability to find most of the anomalies.

$$Precision = \frac{TP}{TP\ +\ FP} \tag{6}$$

$$Recall = \frac{TP}{TP\ +\ FN} \tag{7}$$

Using both Precision and Recall measures, a PR curve shows the tradeoff between finding most anomalies in a dataset while not compromising the quality of the results with false positives, i.e. normal samples classified as abnormal ones. The curve evolves along a threshold. This threshold is the boundary where a sample is classified as normal or abnormal, and is directly related with Precision and Recall scores. The area under the PR curve is bigger when a method results have high precision and recall, meaning a low False Positive and False Negative rates (see Figure 20 The optimal method would get high Recall, returning all abnormal samples correctly identified, and high Precision, meaning no false anomalies would be identified.

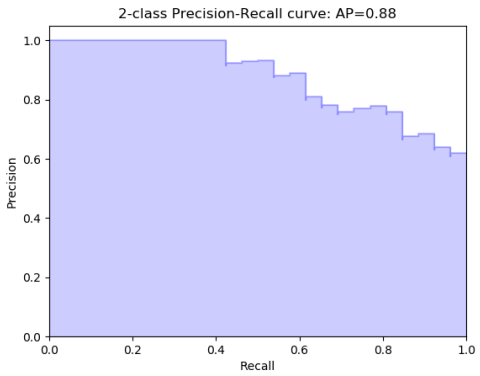


Fig. 20: Precision and Recall (PR) curve

The harmonic mean of the Precision and Recall scores, often identified as F1-Score or F-Measure (in equation 8), can be found as an aggregated performance measure for anomaly detection methods evaluation, considering equal contribution for Precision and Recall values.Other commonly found F measures include F_2, considering a bigger relative contribution of recall against precision and $F_{0.5}$, having more emphasis on precision than on recall.

$$F_1 Score = 2 * \frac{Precision\ *\ Recall}{Precision\ +\ Recall} \tag{8}$$

Like the PR curve, Receiver Operator Characteristic (ROC) curve is also frequently applied to represent the performance of binary classifiers. It uses the True Positive Rate (TPR) and False Positive Rate (FPR) - expressed in equations 9 and 10 - as axis (see Figure 21). Regarding unsupervised anomaly detection, a ROC curve represents the classification performance under different threshold settings. Here, threshold is a trade-off between more true and false positives - when the classifier tends to be optimistic - or less false and true positives - when it is more conservative. Differently from typical performance values where results fall into -1 to 1 or 0 to 1 intervals, the worst ROC result occurs when the Area Under Curve (AUC) is 0.5. This area increases with the class separation and the ability of the anomaly detection method to correctly distinguish between normal and abnormal samples.

$$TruePositiveRate = \frac{TruePositives}{TotalPositives} \tag{9}$$

$$FalsePoisitveRate = \frac{FalsePositives}{TotalNegatives} \tag{10}$$

The most commonly found evaluation methods for unsupervised anomaly detection are based on ROC and AUC curves. These were used in one of the few studies dedicated to the comparison of unsupervised anomaly detection methods, the one published by Goldstein and Uchida [61] where 9 algorithms were tested over 10 datasets. These curves are commonly found metrics in anomaly detection field, including in the maritime vessels anomaly detection [62].

Following an original statistical approach proposed by Muller et al [63], Clémençon and Jakubowicz [64] proposed the Excess-Mass curve as a scoring method for unsupervised

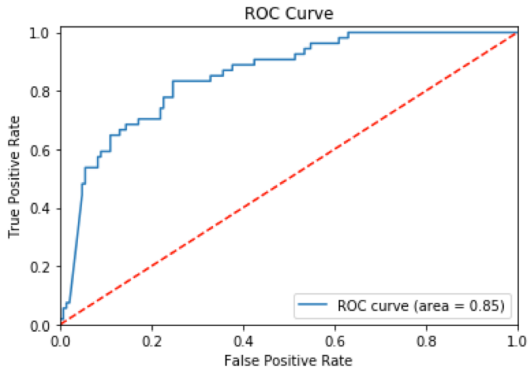
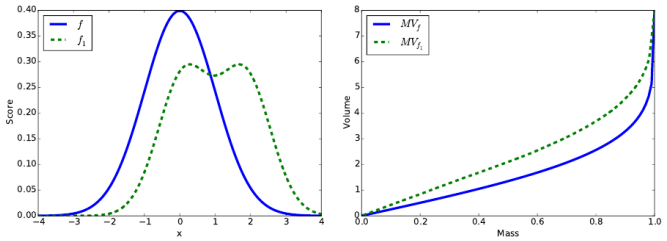


Fig. 21: Receiver Operator Characteristic (ROC) curve



(a) Scoring functions (b) Mass Volume curves

Fig. 22: Example of (a) two different scoring functions of the same normal density distribution function and their Mass Volume curves (b) [67]

problems using an M-estimation approach. The lower the density areas in the Excess-Mass curve, the higher the probability of a sample in that are being abnormal. Nicolas Goix [65] [66] et al. followed this perspective and compared the results measured by Excess-Mass (EM) curve against Receiver Operator Characteristic and Precision-Recall curves on 12 different problems. They run three reference anomaly detection methods - One-Class SVM, Isolation Forest and Local Outlier Factor - in an anomaly detection (by classification) task.

The Mass Volume curve was subsequently proposed by Clémençon and Thomas [67] as a performance metric specifically for unsupervised anomaly detection. This method is supported by the alignment between a probability density function and a scoring function. Hyperparameters of the scoring function are tuned by minimizing the area under the curve (see Figure 22).

Although dedicated to unsupervised machine learning contexts the proposed algorithm is based on iteratively splitting the dataset onto test and train sets following a typical iterative k-Fold cross validation methodology. In each iteration every hyperparameter of the scoring function is ajusted using the test set and the resulting Area of the Mass Volume (AMV) curve as illustrated in Figure 23. The final hyperparameter values are defined by averaging the hyperparameters got in all splits. Both Mass Volume and Excess Mass curves have been used int different research domains such astronomy, by [68] were the density based evaluation criteria is applied to unsupervised anomaly detection.

Further extending the approach to multivariate anomaly

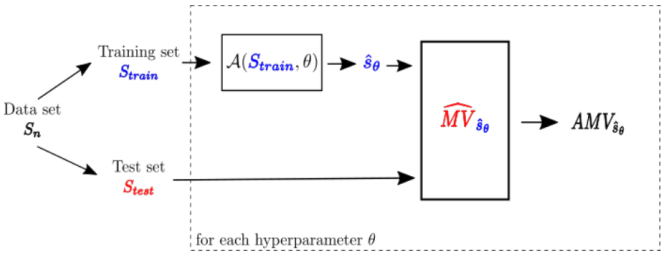


Fig. 23: Algorithm proposed by Albert Thomas on using Mass Volume curve analysis for unsupervised anomaly detection.

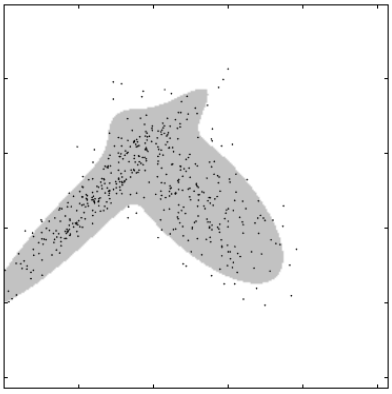


Fig. 24: Example of a Minimum Volume set (gray area) for a crossed two-component Gaussian mixture distribution as illustrated by Scott and Nowak [71].

detection in extreme regions - where anomalies are likely to be located far apart from the median or mean of the distribution - these authors proposed to rely on the Extreme Value Theory [69] (and Minimum Volume set) by finding the most likely directions of abnormal sample locations. These locations are inferred by estimating the Minimum Volume set and an angular measure. Together they define a finite limit for the tail of the distribution. Minimum Volume sets define regions where the mass of a distribution is mostly concentrated. These were investigated by Garcia et al. [70] who identified the application of Minimum Value sets to anomaly detection, multivariate confidence regions and clustering. Figure illustrates a Minimum Volume set for a two-component Gaussian distribution.

5 Concluding Remarks

In this paper we have reviewed different anomaly detection approaches focusing multivariate temporal data. We also analysed univariate based approaches since they are the foundation of multiple anomaly detection works. This research area is being given more attention from different research communities which lead us to present fundamental definitions and characteristics of anomalies and the anomaly detection field.

Anomaly detection approaches were grouped in six main groups, namely (1) distance based methods, (2) density based methods, (3) component analysis based methods, (4) time-based methods, (5) Neural Networks methods and

finally (6) a group for others not included in any of the former. We described the generic framework for anomaly detection as well as the typical training procedures.

This work explored some of the most common evaluation metrics in anomaly detection but also included measures that in our perspective should be taken into account regarding the unlabeled characteristic of most available real word datasets. This is most relevant for machine learning based methods were unsupervised approaches evaluation is still an open topic.

This work contributes to a consolidated view over some of the main research works in this field. In the future multivariate anomaly detection will have to deal natively with very big datasets where temporal data series can have very different characteristics, sizes or even data types. Sensors and software systems evolution as well as big tracks of historical data will raise the challenge for new anomaly detection approaches, these will certainly require new approaches and methods focused specifically in dealing with complex and large-scale anomaly detection tasks.

REFERENCES

[1] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

[2] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial intelligence review*, vol. 22, no. 2, pp. 85–126, 2004.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[4] C. C. Aggarwal, *Outlier Analysis*. Springer Science & Business Media, 2013.

[5] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250–2267, 2014.

[6] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 2, pp. 159–170, 2010.

[7] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, 2016.

[8] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, "A survey of deep learning-based network anomaly detection," *Cluster Computing*, pp. 1–13, 2017.

[9] H. Fanaee-T and J. Gama, "Tensor-based anomaly detection: An interdisciplinary survey," *Knowledge-Based Systems*, vol. 98, pp. 130–147, 2016.

[10] Y. Zheng, "Trajectory data mining: an overview," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, p. 29, 2015.

[11] R. S. Tsay, D. Peña, and A. E. Pankratz, "Outliers in multivariate time series," *Biometrika*, vol. 87, no. 4, pp. 789–804, 2000.

[12] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *ACM Sigmod Record*, vol. 29, no. 2. ACM, 2000, pp. 427–438.

[13] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 8, no. 3-4, pp. 237–253, 2000.

[14] D. Jankov, S. Sikdar, R. Mukherjee, K. Teymourian, and C. Jermaine, "Real-time high performance anomaly detection over data streams: Grand challenge," in *Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems*. ACM, 2017, pp. 292–297.

[15] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, no. 2. ACM, 2000, pp. 93–104.

[16] D. Pokrajac, A. Lazarevic, and L. J. Latecki, "Incremental local outlier detection for data streams," in *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on*. IEEE, 2007, pp. 504–515.

[17] J. Zhang, Q. Gao, and H. Wang, "Spot: A system for detecting projected outliers from high-dimensional data streams," in *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. IEEE, 2008, pp. 1628–1631.

[18] W. Drosdowsky, "An analysis of australian seasonal rainfall anomalies: 1950–1987. i: Spatial patterns," *International Journal of Climatology*, vol. 13, no. 1, pp. 1–30, 1993.

[19] ——, "An analysis of australian seasonal rainfall anomalies: 1950–1987. ii: Temporal variability and teleconnection patterns," *International Journal of Climatology*, vol. 13, no. 2, pp. 111–149, 1993.

[20] C.-T. Lu and L. R. Liang, "Wavelet fuzzy classification for detecting and tracking region outliers in meteorological data," in *Proceedings of the 12th annual ACM international workshop on Geographic information systems*. ACM, 2004, pp. 258–265.

[21] R. Lasaponara, "On the use of principal component analysis (pca) for evaluating interannual vegetation anomalies from spot/vegetation ndvi temporal series," *Ecological Modelling*, vol. 194, no. 4, pp. 429–434, 2006.

[22] X. Li, J. Han, and S. Kim, "Motion-alert: automatic anomaly detection in massive moving objects," in *International Conference on Intelligence and Security Informatics*. Springer, 2006, pp. 166–177.

[23] R. Baragona and F. Battaglia, "Outliers detection in multivariate time series by independent component analysis," *Neural computation*, vol. 19, no. 7, pp. 1962–1984, 2007.

[24] O. Salem, A. Guerassimov, A. Mehaoua, A. Marcus, and B. Furht, "Anomaly detection in medical wireless sensor networks using svm and linear regression models," *Int. J. E-Health Med. Commun.*, vol. 5, no. 1, pp. 20–45, Jan. 2014. [Online]. Available: http://dx.doi.org/10.4018/ijehmc.2014010102

[25] A. Kane and N. Shiri, "Multivariate time series representation and similarity search using pca," in *Industrial Conference on Data Mining*. Springer, 2017, pp. 122–136.

[26] N. Liu and P. Rebentrost, "Quantum machine learning for quantum anomaly detection," *arXiv preprint arXiv:1710.07405*, 2017.

[27] B. Samanta, "Gear fault detection using artificial neural networks and support vector machines with genetic algorithms," *Mechanical Systems and Signal Processing*, vol. 18, no. 3, pp. 625–644, 2004.

[28] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing dtw to the multi-dimensional case requires an adaptive approach," *Data Min. Knowl. Discov.*, vol. 31, no. 1, pp. 1–31, Jan. 2017. [Online]. Available: https://doi.org/10.1007/s10618-016-0455-0

[29] S. Seto, W. Zhang, and Y. Zhou, "Multivariate time series classification using dynamic time warping template selection for human activity recognition," in *Computational Intelligence, 2015 IEEE Symposium Series on*. IEEE, 2015, pp. 1399–1406.

[30] B. Samanta, K. Al-Balushi, and S. Al-Araimi, "Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection," *Engineering Applications of Artificial Intelligence*, vol. 16, no. 7, pp. 657–665, 2003.

[31] B. Samanta and K. Al-Balushi, "Artificial neural network based fault diagnostics of rolling element bearings using time-domain features," *Mechanical systems and signal processing*, vol. 17, no. 2, pp. 317–328, 2003.

[32] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proceedings*. Presses universitaires de Louvain, 2015, p. 89.

[33] A. Lakhina, M. Crovella, and C. Diot, "Characterization of network-wide anomalies in traffic flows," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*. ACM, 2004, pp. 201–206.

[34] M. Gupta, A. B. Sharma, H. Chen, and G. Jiang, "Context-aware time series anomaly detection for complex systems," in *WORKSHOP NOTES*, 2013, p. 14.

[35] N. Takeishi and T. Yairi, "Anomaly detection from multivariate time-series with sparse representation," in *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2651–2656.

[36] Z. Tan, A. Jamdagni, X. He, P. Nanda, and R. P. Liu, "A system for denial-of-service attack detection based on multivariate correlation analysis," *IEEE transactions on parallel and distributed systems*, vol. 25, no. 2, pp. 447–456, 2014.

[37] X. Kong, X. Song, F. Xia, H. Guo, J. Wang, and A. Tolba, "Lotad: long-term traffic anomaly detection based on crowdsourced bus trajectory data," *World Wide Web*, pp. 1–23, 2017.

[38] Y. Xu, Z. Wu, J. Li, A. Plaza, and Z. Wei, "Anomaly detection in hyperspectral images based on low-rank and sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 4, pp. 1990–2000, 2016.

[39] A. J. Fox, "Outliers in time series," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 3, pp. 350–363, 1972.

[40] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing dynamic time warping to the multi-dimensional case requires an adaptive approach."

[41] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *ACM Sigmod Record*, vol. 30, no. 2. ACM, 2001, pp. 37–46.

[42] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Conditional anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 5, pp. 631–645, 2007.

[43] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and Information Systems*, vol. 51, no. 2, pp. 339–367, May 2017. [Online]. Available: https://doi.org/10.1007/s10115-016-0987-z

[44] F. Gottwalt, E. Chang, and T. Dillon, "CorrCorr: A feature selection method for multivariate correlation network anomaly detection techniques," *Computers & Security*, vol. 83, pp. 234–245, 2019.

[45] Z. Niu, S. Shi, J. Sun, and X. He, "A survey of outlier detection methodologies and their applications," *Artificial intelligence and computational intelligence*, pp. 380–387, 2011.

[46] C. Liu, S. C. Hoi, P. Zhao, and J. Sun, "Online arima algorithms for time series prediction," in *Thirtieth AAAI conference on artificial intelligence*, 2016.

[47] F. Schmidt, F. Suri-Payer, A. Gulenko, M. Wallschläger, A. Acker, and O. Kao, "Unsupervised Anomaly Event Detection for VNF Service Monitoring Using Multivariate Online Arima," in *2018 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*. IEEE, 2018, pp. 278–283.

[48] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[49] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1391–1445, 2009.

[50] M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara, "Least-squares conditional density estimation," *IEICE Transactions on Information and Systems*, vol. 93, no. 3, pp. 583–594, 2010.

[51] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, "Statistical outlier detection using direct density ratio estimation," *Knowledge and information systems*, vol. 26, no. 2, pp. 309–336, 2011.

[52] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[53] D. Li, D. Chen, J. Goh, and S.-k. Ng, "Anomaly detection with generative adversarial networks for multivariate time series," *arXiv preprint arXiv:1809.04758*, 2018.

[54] D. Li, D. Chen, L. Shi, B. Jin, J. Goh, and S.-K. Ng, "MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks," *arXiv preprint arXiv:1901.04997*, 2019.

[55] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[56] 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.

[57] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 387–395.

[58] D. Song, N. Xia, W. Cheng, H. Chen, and D. Tao, "Deep r-th root of rank supervised joint binary embedding for multivariate time series retrieval," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 2229–2238.

[59] J. Li, W. Pedrycz, and I. Jamal, "Multivariate time series anomaly detection: A framework of Hidden Markov Models," *Applied Soft Computing*, vol. 60, pp. 229–240, 2017.

[60] N. Goix, A. Sabourin, and S. Clémençon, "Sparse representation of multivariate extremes with applications to anomaly detection," *Journal of Multivariate Analysis*, vol. 161, pp. 12–31, Sep. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0047259X17304062

[61] M. Goldstein and S. Uchida, "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data," *PLOS ONE*, vol. 11, no. 4, p. e0152173, Apr. 2016. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152173

[62] B. Auslander, K. M. Gupta, and D. W. Aha, "A comparative evaluation of anomaly detection algorithms for maritime video surveillance," in *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense X*, vol. 8019. International Society for Optics and Photonics, 2011, p. 801907.

[63] D. W. Müller and G. Sawitzki, "Excess mass estimates and tests for multimodality," *Journal of the American Statistical Association*, vol. 86, no. 415, pp. 738–746, 1991.

[64] S. Clémençon and J. Jakubowicz, "Scoring anomalies: a m-estimation formulation," in *Artificial Intelligence and Statistics*, 2013, pp. 659–667.

[65] N. Goix, "How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?" *arXiv:1607.01152 [cs, stat]*, Jul. 2016, arXiv: 1607.01152. [Online]. Available: http://arxiv.org/abs/1607.01152

[66] N. Goix, A. Sabourin, and S. Clémençon, "On anomaly ranking and excess-mass curves," in *Artificial Intelligence and Statistics*, 2015, pp. 287–295.

[67] S. Clémençon and A. Thomas, "Mass Volume Curves and Anomaly Ranking," *arXiv preprint arXiv:1705.01305*, 2017.

[68] D. Baron and D. Poznanski, "The weirdest SDSS galaxies: results from an outlier detection algorithm," *Monthly Notices of the Royal Astronomical Society*, vol. 465, no. 4, pp. 4530–4555, Mar. 2017. [Online]. Available: https://academic.oup.com/mnras/article/465/4/4530/2568826

[69] S. Kotz and S. Nadarajah, *Extreme value distributions: theory and applications*. World Scientific, 2000.

[70] J. Nuñez Garcia, Z. Kutalik, K.-H. Cho, and O. Wolkenhauer, "Level sets and minimum volume sets of probability density functions," *International Journal of Approximate Reasoning*, vol. 34, no. 1, pp. 25–47, Sep. 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888613X03000525

[71] C. D. Scott and R. D. Nowak, "Learning minimum volume sets," *Journal of Machine Learning Research*, vol. 7, no. Apr, pp. 665–704, 2006.