

# Should I Raise The Red Flag?

## A comprehensive survey of anomaly scoring methods toward mitigating false alarms

Zahra Zohrevand and Uwe Glässer

Simon Fraser University, Burnaby BC, Canada  
 {zzohreva, glaesser}@sfu.ca

**Abstract.** A general Intrusion Detection System (IDS) fundamentally acts based on an Anomaly Detection System (ADS) or a combination of anomaly detection and signature-based methods, gathering and analyzing observations and reporting possible suspicious cases to a system administrator or the other users for further investigation. One of the notorious challenges which even the state-of-the-art ADS and IDS have not overcome is the possibility of a very high false alarms rate. Especially in very large and complex system settings, the amount of low-level alarms easily overwhelms administrators and increases their tendency to ignore alerts.

We can group the existing false alarm mitigation strategies into two main families: The first group covers the methods directly customized and applied toward higher quality anomaly scoring in ADS. The second group includes approaches utilized in the related contexts as a filtering method toward decreasing the possibility of false alarm rates. Given the lack of a comprehensive study regarding possible ways to mitigate the false alarm rates, in this paper, we review the existing techniques for false alarm mitigation in ADS and present the pros and cons of each technique. We also study a few promising techniques applied in the signature-based IDS and other related contexts like commercial Security Information and Event Management (SIEM) tools, which are applicable and promising in the ADS context. Finally, we conclude with some directions for future research.

**Keywords:** Anomaly detection, Anomaly scoring, False alarm mitigation, Cyber Security Systems, Intrusion detection, Time series forecasting, Critical infrastructure protection

## 1 Introduction

As we know, adversarial settings systems like Cyber Security Systems (CSSs), Intrusion Detection Systems (IDSs) or Insider Threat Detectors should be able to handle very high volume of heterogeneous log data under influence of user behaviour and network status. Therefore, taking advantage from strong Anomaly Detector (AD) is crucial to track the existing heterogeneous characteristics (e.g., rate, size, type) by applying ensemble methods of individual trackers or a method which handles correlation of detectors (a feature or group of features) in almost real time.

Nowadays, deep learning based state of the art methods have proved their unique capabilities in constructing or forecasting the next observations, subsequently can lead into more accurate ADS. In other words, having a better estimation of what should we expect to observe, helps to detect abnormal cases.

But, as Figure 1 illustrates, ADS is beyond model behaviour component; finding the right approach of scoring and ranking datapoints in respect to anomalous level, finding the optimal threshold to discretize labels, getting rid of noises or valueless outliers, mitigating false alarm rate in the presence of high recall, and finding the story behind anomalies are some of the challenges which should be addressed. Unfortunately, even advanced models are not end-to-end anomaly detectors to be able to solve this problem and its challenges by taking raw input data and finalize the anomalous sets.

Most of the aforementioned tasks need extra optimization which should be performed in the inference phase. This stage can be as simple as comparing the estimation errors versus a fixed threshold or as complicated as applying another complex model to infer the final scores and labels based on contextual information and the result of behaviour predictor.

The main goal of this study is reviewing the strategies taken by different methods in literature, which may help to mitigate false alarms rate. To this aim, some other issues like scoring approaches, setting threshold techniques, and collective analysis which may help in reducing the false alarm rate is addressed, too.

Surprisingly, even in the simplest scenarios like one-detector based ADS, finding the optimum threshold for discretizing score and raising alarm for anomalies is very tricky. But, in real world, usually ADSs should handle heterogeneous set of models based on different detectors, which makes it more complicated, because a very slight mistuned threshold can unexpectedly lead into an unemployable system which produces huge false alarm rate or misses most of the attacks by labeling them as normal [11].

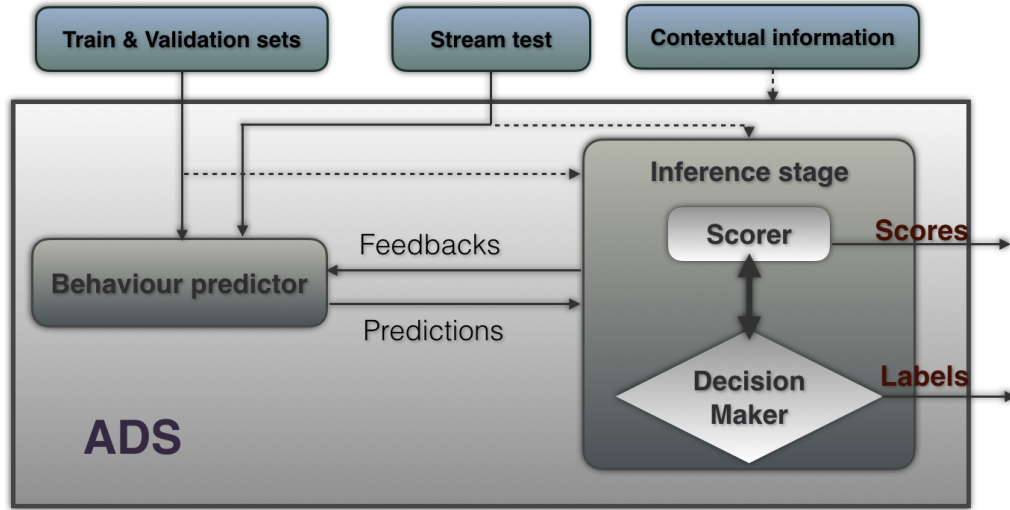


Fig. 1: Anomaly detection system structure

Our literature review showed that usually anomaly detection and malicious detection have been used interchangeably. But, there is a fine difference between these two concepts that distinguishes them and barricades applying some anomaly detectors toward finding the malicious events.

The main aim of anomaly detection is raising the flag in case of observing unknown, unusual or rare events. As we know, malicious events are very likely to be rare and unusual, but simply applying anomaly detection to perform malicious may cause a huge amount of false alarm rate. Thus, we can claim that the main demand of anomaly detection is puzzling out the unusual events, while the mission of malicious detection is finding unusual events which may hurt system or have cascading effects. Therefore, this definition should be modified to be able to capture the target of malicious detection [78]. Moreover, to find malicious events based on anomaly detection, we should prune results to keep anomalies of interest, which may unavoidably, reduces the true alert rate, as well. From another point of view and in a nutshell, this study is looking for the existing approaches to customize ADS result to be able to find the malicious events in IDSs or other critical domains.

One aspect of adversarial settings is that adversaries try to blend in with the distribution of normal points [22]. When the attacks (targets of interest) that are not confined to extreme outliers, or when the extreme outliers are not anomalies, the anomalies of interest will be confused with normal points or with uninteresting outliers as swamping and masking happens.

Even though, we believe that anomaly points should be almost isolated, but in some contexts, the possibility of having multiple processes generating anomalies is more likely. For example, in a cyber-security setting, there exist different kinds of attacks and many different methods for stealing information; therefore, some of them may cover the others. So the group of ADSs which focus on uniqueness of the observations to score and find the anomalies would be inefficient to detect attacks, in the aforementioned settings.

On the other hand, the risk of having clustered anomalies is feasible, which happens if one process is generating many instances of anomalies, those observations may not appear as statistical anomalies, anymore; particularly, if these anomalous points would be tightly clustered, they form a very dense region, which causes the density based methods be doomed to failure. So, we are interested in finding the strategies that some methods have considered to address these challenges.

Another reason that motivated us to perform this study is how these matters have been addressed in AD algorithms to reduce useless positives events, which may happen because of having wrong and abrupt spikes in the error rate, due to different reasons like noisy data or model error. For example, abrupt changes in values are often not perfectly predicted and resulted in sharp spikes in error values even when this behavior is normal [75]. Conventionally, this matter may happen in data driven methods, which do not consider any confidence interval for their prediction.

## 1.1 Motivation and Challenges

A comprehensive survey on anomaly detection should provide the readers enough information regarding end-to-end steps of anomaly detection and the intuition behind applying each particular sub-step, including the behaviour modeling method, scoring technique, score discretization approach, so that they would be able to pick and combine the methods based on their strength and context. To the best of our knowledge, most of the existing review and survey papers are just concerned about the behaviour modeling method. So, the reader may get confused about what to do with the final obtained result from the predictive model and how to decrease false-alarm rate.

Moreover, a comparative analysis of various techniques and their pros as cons is required. But, since, most of the studies and methods have been customized based on different datasets and their scoring approach are not unified or compatible, so they are not fully comparable. Thus, we have gathered a set of measures, which may be suitable to evaluate the performance of future methods even based on the different datasets.

Since, the boundary between normal and anomalous (erroneous) behavior is often not precisely defined and is continually evolving because of the data drift. So, finding the best representative boundary based on the result of predictive models is a challenge, which need to be addresses.

## 1.2 Organization

The remainder of the article is organized as follows. Section 2, first sets out the problem definition, then provides an informal description of the required concepts related to anomaly detection on which the rest of the paper relies. Section 3, then formulate and justifies a collection of requirements which should be addressed in anomaly detection. In the next step, Section 4 analytically and comprehensively reviews the existing methods to scale the false alarms rate, from the initial scoring steps till decision making steps for raising the red flag. Section 5 continues the false positive mitigation topic in post-hoc analysis. Section 5 is devoted to the study of concept drift detection methods, which is highly influential in false alarm rate. ultimately, Section 7 brings up some research question and directions for future research.

## 2 Related works

For a long time, anomaly detection has been the topic of many surveys and books. A considerable amount of research on outlier and anomaly detection comes from statistics contexts like [70,31,8]. Afterward, some other studies from computer science field have reviewed and surveyed the anomaly detection concepts more or less concerning the computational aspects [2,58,34]; which most of them have been covered in the comprehensive survey by Chandola and et al. [12] that have deeply analyzed the pros and cons of anomaly detection methods. Thereafter, some other studies have collected and reviewed the state-of-the art methods in various contexts. For example, [30] have thoroughly analyzed and categorized time-series anomaly detection methods based on their fundamental strategy and the type of data taken as input. [44] provides a comprehensive survey of deep learning-based methods for cyber-intrusion detection. A broad review of deep anomaly detection techniques for fraud detection is presented by [1] and Internet of Things (IoT) related anomaly detection has been reviewed by [52].

On the other hand, [35] performed a very extensive review of the applied techniques in the signature-based IDSs toward mitigating false alarm rate. The primary difference of this study with ours is that Hubballi and et al. have targeted signature-based intrusion detection, while we are focusing on mitigation techniques applicable on anomaly detection methods toward finding adversarial settings.

### 2.1 Our contribution

We follow the survey approach taken in [12,35] for false alarm mitigation in anomaly detection context. Our survey presents a very comprehensive and structured overview of extensive research related to anomaly scoring and their influence on false-alarm rate.

As aforementioned, most of the existing surveys on anomaly detection [2,12] focus on the behaviour modeling technique applied in training anomaly detector and have ignored the post pruning, threshold setting performed by system's decision function to score anomaly level of the data-points and confirm the labels. But, the anomaly detector should be considered as an end-to-end system and the finalizing steps are more important than being ignored as a hack. This concept is especially, more important in profiling and prediction-based anomaly detectors that the level of their

accuracy may be heavily influenced by small changes in their threshold value or the metric to score anomalies. We summarize our main contributions as the following:

1. This survey builds upon the extensive research and surveys on prediction and profiling based anomaly detection by significantly expanding the discussion toward anomaly scoring and false alarm mitigation. We not only focus on anomaly false-alarm mitigation related techniques, but also identify unique assumptions regarding the nature of anomalies made by the techniques. The combination of these kinds of assumptions and pruning steps are critical for determining the failure and successful contexts for that technique. We also present the challenges, advantages, and disadvantages of each technique.
2. While, the existing surveys discuss anomaly detection as a whole concept by focusing on finding anomalies or under the general representation of point, collective and contextual anomalies; we have gathered a broad overview of the criteria that anomaly detection and anomaly scorer methods should address to be applicable in real world settings. Thus, we distinguish these technique in a more precise way and based on their ability to find the anomalies, which are more complex in nature.

### 3 Basic definitions

As a first go, we start off with describing the anomaly, anomaly scoring and recalling other key concepts and requirements in anomaly detection involved in false alarm mitigation. But, problem definition has been provided, above all.

#### 3.1 Problem definition

This survey addresses a challenging but important question for ADS target domains like IDS, CSS and fraud detection. When and how the ADS should raise a red (flag) alert and announce the possibility of suspicious cases. In another words, how should the alert threshold be set in the domain spaces with a large number of highly dynamic and heterogeneous features and detectors. Especially, in the era of big data, the ADS should find a very reasonable and finite set of the most abnormal events.

#### 3.2 What is an anomaly?

One common definition of anomaly (outlier) is based on Hawkins' state in [31]: "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism". Several other definitions of anomaly and anomaly detection have been originated based on different interpretation of anomaly; and the applied description, as a statement of goal, can strongly manipulate ADS results. Some of the important definitions are as the following:

1. **Anomalies are rare events.** Since, unusual cases do not happen frequently, so anomalies are rare events. An ADS, which focuses on this definition should find a soft or hard threshold of frequencies, so many anomalies will be missed if the estimation of frequency would not be correct, like points *A* and *B* which have been missed in the second scenario of Figure 2 because of the existence of other more anomalous points. Moreover, the result of the methods which focus on rarity score of data points are not comparable, because the rarity depends on ADS algorithm. On the other hand, sometimes the rarity assumption may increase false negative and false positive rate, if the data comes from several generator processes.
2. **Anomalies are distinct and different events.** Based on this definition, whatever, which is odd is anomalous. Being different is not meaningful without having some implicit probability that shows the rank of belonging to a distribution or model. So again a threshold is required for determining "to what extent?" For example, in clustering methods like [62], the anomaly score can be computed based on the distance to clusters or the most less dense (rarest) clusters may be detected as anomalies. So, implicitly it consider the difference measure to decide.
3. **Anomalies are abnormal events.** The observations which diverge from normal expectations are anomalous. So, normal data labels should be available, while it is not the case in the most of the target domains like IDSs [25].

But, none of the above interpretations of anomaly detection are acquisitive enough in all the application domains; For example in network IDS, fraud detection, the comprehensive definition as goal should capture some other specifications like the observation rarity in the presence of potential clusteredness phenomena, not being noise and etc. In other words,

each ADS based IDS should be able to locate and report a manageable subset of events which are divert enough from normal to be suspicious and worth further investigation.

Based on [25], an ideal definition of anomaly should be applicable on all of the possible distributions without any extra effort. Moreover, it should provide the possibility of comparing anomalous degree of one target variable versus the anomalousness values of the other variables. Therefore we provide our definition of anomaly in suspicious detection context as a combination of the existing ones. **An observation or collection of observations can be anomaly if they would be a very distinct event in terms of feature values or clustered but unobserved previously and stable or frequently close to the previous suspicious cases to reject the noise likelihood.**

### 3.3 What is anomaly scoring?

ADS aims to order anomalies based on their anomalous score that AD algorithm assigns to the data points. One of the very basic methods is transferring the real order in feature space through some scoring function  $S_{AD} : X \rightarrow R+$ . Thus, more anomalous points get smaller score. In this regard, some statistical depth function have been applied to perform scoring [97]. Since, false alarm rate and scoring approach are mutually dependent, in Section 4 we describe the applied anomaly scoring techniques.

### 3.4 Behavior predictor model

These kinds of methods maintain a model as a normal profile and compare the value of the upcoming datapoints against this profile to decide their divergence from the expectations. For example, [90] keeps track of a normal profile, based on the smoothed version of the historical TS, and also a variance vector; A new point will be compared with the normal profile and the variance vector to be assigned as anomaly score. A huge variety of machine learning based models like SVR, Auto Regression, Random Forest (RF), Hierarchical Temporal Memory (HTM) and Deep learning have been used to maintain and track the normal behaviour and estimate the upcoming expectations [27,49,33].

### 3.5 Concept drift and abrupt evolution of data

The term of concept drift implies that the statistical properties of the target variable or even input data has changed in the way that predictive model is not able to forecast them and loses its accuracy level. For example, the behavior of customers in power consumption may change over time because of many reasons like restructuring their internal network; consequently the consumption predictor is likely to become less and less accurate over time, because of this concept drift.

In general, it is hard to determine the exact rate of drift in the data and some strategies have been proposed that we introduce some of them in Section 7. Even though, some literature consider seasonality or cycles as concept drift [94], but in this study our interpretation of concept drift focuses on shifts, which permanently change the normal behaviour.

## 4 Challenging Criteria for ADs

As aforementioned, anomaly detection and scoring is challenging because of many reasons including high dimensionality spaces, potential data drifts, seasonality and highly irregular rate of data observation, highly noisy data, mixed data types (continuous data, integer data, lattice data), bounded data, lags between the emergence of anomalous behaviour in different under-investigation dimensions and non-Gaussian distributions. Therefore, a strong method in this context should be capable and flexible enough to address these properties in the observed data [83].

In addition to detection rate (as the ratio between the number of correctly detected anomalies and their total number) and false alarm (as the ratio between the number of normal data points that are incorrectly misclassified and the total number of normal points), there are some other measures that can help to improve the quality of AD based cyber protection. Based on the meta learning for anomaly detection presented in [21,22] and some other contexts, we have gathered a collection of requirements which should be considered in anomaly scoring. In addition, we have summarized the measures which can evaluate the strength of the method in addressing the specific circumstances in cyber protection.

#### 4.1 Masking Effect

It is said that one or an anomaly point "Masks" a second anomaly, if the second one can be considered as an anomaly only by itself, but not in the presence of the first anomaly. Masking can occur when the mean and the covariance estimates have been skewed toward a cluster of outlying observations, so the distance of the outlying point from the mean is not large enough [9]. Figure 3 illustrates a toy example of this scenario about a method which update its threshold value, dynamically. Therefore, an attacker by knowing the ADS's strategy, feeds system with fake values, similar enough to the process to be in the confidence interval, but different enough to make ADS to shift its threshold value. Later on, attacker will be able to perform the masked attack (A), without being captured by the manipulated ADS. Another example of masking has been displayed in Figure 2 that rarity-based ADS has missed A and B, because of limiting of the number of anomalous points. In other words, if test looks for too few anomalies, the additional ones may influence the statistics so that no points be declared as anomalies [48]. One statistical representation of this phenomena is when  $\sigma_A^2 \leq \sigma_N^2$  by considering  $\sigma_N^2$  and  $\sigma_A^2$  as the normalized sample variance of the selected normal points and the selected anomaly points, respectively. **Semantic Variations** is one of the proposed measures which is able to evaluate ability of ADS in handling masking effect. A measure of the degree to which the anomalies are generated by more than one underlying process, or, alternatively, the degree to which the anomalies are dissimilar from each other.

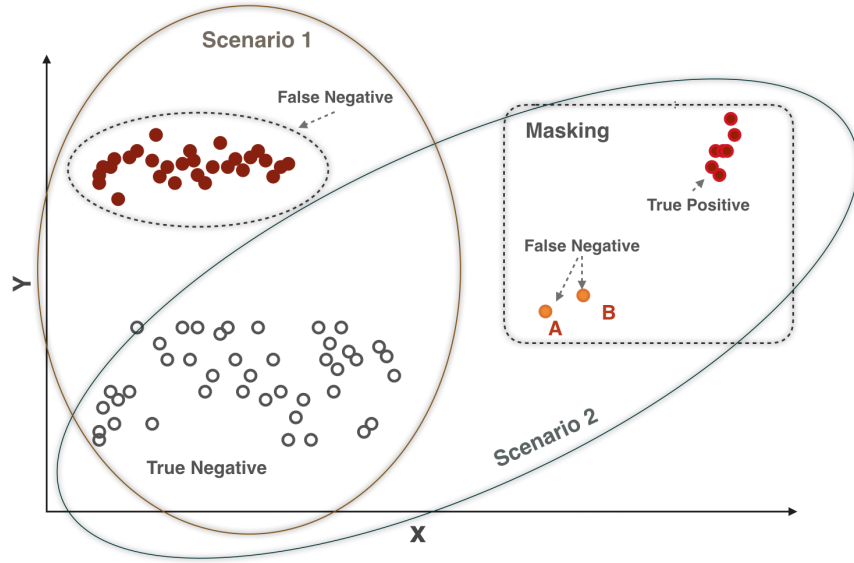


Fig. 2: An example of rarity assumption based masking

#### 4.2 Swamping Effect

This phenomena is reverse of Masking and may happen if the swamped datapoint can be considered as an anomaly only under the presence of another datapoint(s), like false positive cases: B, C, D in Figure3, which have been swamped by orange points. In other words, after deletion of the first outlier the second observation becomes a non-outlying observation. If a group of outlying instances skews the mean and the covariance estimates toward themselves can lead to swamping some non-outlying instances, and the resulting distance from these instances to the mean will be large, making them look like anomaly [9]. If the statistical test overestimates the number of anomalies in dataset, it can be influenced by swamping effect. So, there is a tricky tradeoff between ADS abilities to avoid falling in swamping and masking traps.

**Point difficulty** is one of the proposed measures in [22] to evaluate swamping effect. The anomalies of interest assumption breaks down as the target points become harder to distinguish from the normal points. Point difficulty of any point can be measured as the estimated probability that it belongs to the other class. So, an ideal ADS should be able to detect anomalies with higher point difficulty rates. But, as semantic variance, to find a realistic value for this

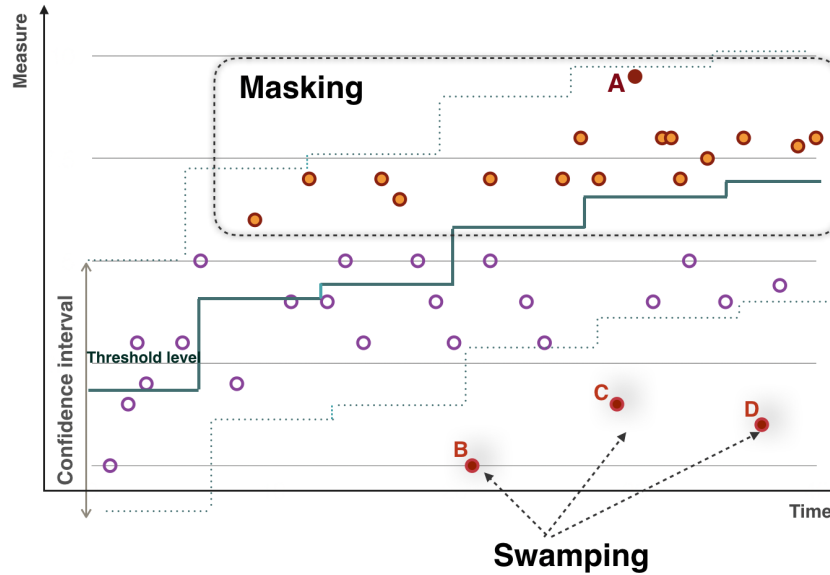


Fig. 3: A simple example of Masking and Swamping effects

measure, class labels are required, otherwise, anomaly scores can be applied to show the challenging boundaries of ADS. In addition the definition of distance between normal and anomaly distributions in this measure is very tricky and influential on the final result.

#### 4.3 Variable frequency of anomalies

If anomalies are very rare, methods which assume the majority of the training points are "normal" and fit a model to capture the rarity may perform very well. This assumption holds true if the frequency of anomalies would be usual and about 1 to 10 percent, but in some difficult configurations, data may include more than 30 percent of anomalous data [40,48]. Therefore, there are some anomaly scenarios which are more probable than specific irregular normal scenarios like DoS attacks; in such circumstances, the rarity assumption is a failure theory and methods which try to puzzle out normal and anomaly decision boundaries just based on the learned behaviour may perform better. As scenario 1 in Figure 2 illustrates one of these failure scenarios that ADS has missed a group of anomalies, just because of rarity assumption of data anomalous points.

So [22] have introduced **Relative frequency**, which is equal to *contamination rate* or *plurality* as the fraction of the incoming data points that are anomalies. Thus, the reliability of ADS under *variable frequency conditions* is measurable based on their tolerance level under different levels of *relative frequency* without losing the accuracy.

#### 4.4 High dimensionality curse

Having access to more features and detectors, decreases the risk of losing some influential information in performing the task of interest, but it may cause some other problems, which have been thoroughly surveyed in [93] like:

- Irrelevant features. A significant number of attributes may be irrelevant.
- Concentration of scores and distances. Similarity of numerical measures like distance
- Incomparable and non-interpretable scores. Obtained scores produced in each domain is incomparable with others and the final score is not strong and semantically meaningful
- Exponential search space. Not systematically analyzable search space, which lead into having enormous possible hypothesis for every observed significance.

Existing methods are able to tackle one or more of these problems, but some challenges have remained as open research questions. For example, from statistical perspective, each irrelevant feature increases the dimensionality of the space, and the sample size required by (naive) density estimation methods tends to scale exponentially with the dimension, which is not always possible. Moreover, having irrelevant features:

- Decreases the recall. As the dimensionality of the data increases, anomaly points may be covered under the similarity of the other unrelated and unimportant dimensions.
- Decreases the precision and increases false alert frequency. As the dimensionality of the data increases, the domain space of data also growth. So, there are chances that normal points fall away from the others and be labeled as anomaly. From statistical point of view, the possibility of having more “tails” increases, which consequently may push the normal points fall in the tails of some feature distributions [21], while that feature is not important.

Therefore, **Feature Ranking** is a measure, which is able to evaluate the reliability of model in ignoring irrelevant features or paying attention to them based on their importance. This evaluation metric is just applicable on supervised data and also, is highly correlated with feature selection and explainability context, which is highly method dependent [17,6] and one of the active research area in AD context.

Generally speaking, feature engineering as the process of transforming raw data into features, which act as higher quality inputs for machine learning models is essential and proven through empirical result [55]. Generating “good” representation of data is the one that will eventually help the method achieves higher performance on the task of interest than the situation that it does not have access to this representation [86]. This task is a combination of art and science and “so important, difficult, time consuming, and domain expert dependent, that *Applied machine learning* is basically feature engineering.” as Prof. Andrew Ng stated.

#### 4.5 Lag of emergence of anomalies

In stream analysis of the complex systems, the source of anomaly may trigger system features in various ways; such that based on the other correlated features, potential influential hidden factors and time context, some of the features may react to the source of anomaly, sooner than the others. Thus, they represent the anomaly with different lags in the span of time. This phenomena may happen because of the various reasons say causal relations, which generally leads into situations that dependent feature displays anomaly with more delay after causal feature. For instance, response time is highly correlated to memory leak, through causality relation with process footprints. So, if the memory leakage happens, probably a spike in process footprint may show it up sooner, then frequent process swapping will cause increased response times. Another potential reason of various lags is having a very irregular rate of observation in different dimensions; therefore, in some cases, for very extended periods, we do not have access to the observations, while extreme events are decipherable from the other available dimensions [83].

#### 4.6 Domain specific challenges

In addition to the aforementioned criteria, defining domain specific measures, may help to evaluate capabilities of methods and their scoring in target domain with particular settings. For example, [46] applies *Burst detection rate (bdr)* to capture the potential bursts, which indicate attacks including many network connection, in network IDS context. This measure represents the ratio between the total number of intrusive network connections that have the score higher than threshold within the bursty attack and the total number of intrusive network connections within attack interval.

### 5 Automatic false alarm scaling

In the critical applications like embedded systems, heterogeneous networks, or system calls, automated monitoring of thousands of incoming signals whose expected values may vary according to the small changes of the hidden influential factors is very challenging. In such domains a very fast and unsupervised approach is required to detect all potential anomalies and rank them, then filter non-interesting cases and raise alert for more certain ones.

Let assume that the prediction method is tuned and performs very well; now ADS should assign anomaly score to the new observations and find their rank versus previously observed data and consequently should decide whether to raise the alarm.

As aforementioned, the combination of finalizing the scoring and raising alarm are the most challenging steps in anomaly detection pipeline, in view of the fact that a small bug in scoring or ranking process can lead into huge amount of false positive and false negative.

In this section, we deliberate the approaches which have been taken by different statistical, data mining or machine learning methods to manipulate anomaly scores toward decreasing false alarm rate. As figure 4 represents a structured hierarchy of these methods, in some of these techniques the scoring and ordering are not separate tasks and have been



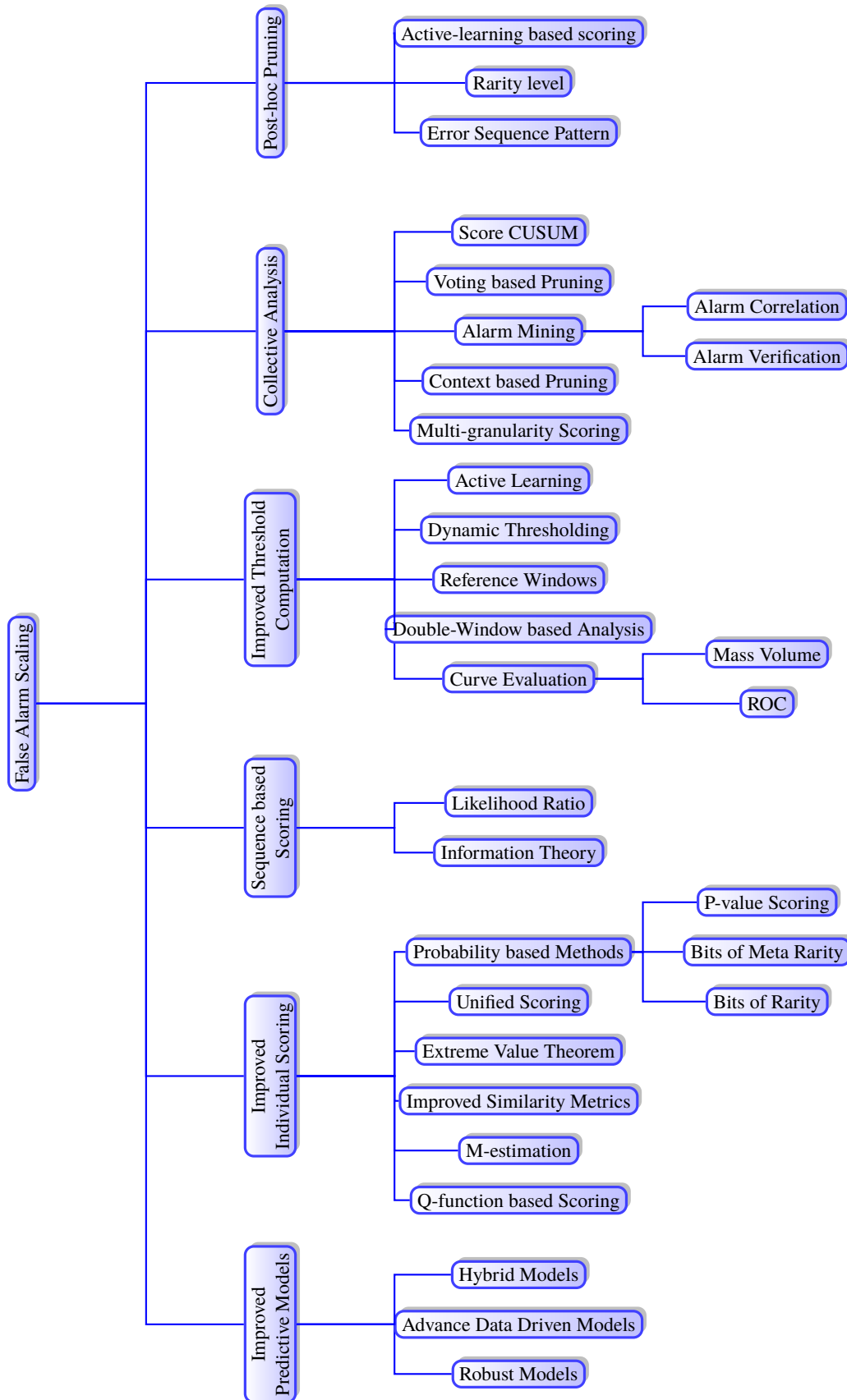


Fig. 4: Hierarchy of anomaly score mitigation methods

performed simultaneously, while some others assign initial scores and then reorder the anomalies based on the obtained score and other potential available sources of information.

We expect the scores, which under investigation ADSs assign to the observations, represent at least a strict weak ordering of the events to be classified as anomaly or normal, so that all the events be sortable based on their deviation from expectation as Hawkins' definition states. In other words, ranking the data points should be performed by a measurable function which maps each data point to real values, denoting the non-negative real line with Borel algebra.

Another point which should be considered is that some of the presented techniques score the data points based on their frequency or similarity to the presumed statistical distribution, so they may seem very simplistic and outdated methods; but, in a high level point of view, each statistical distribution can be defined as a predictor. For example, the density curve of normal distribution, is a representation for the probability of divergence of values from mean  $\mu$  considering the  $\sigma$  value; in other words, prediction is the mean value and other values are the probability of deviation. So, with the same standpoint, in prediction-based ADSs, the same scenario is happening and the obtained error represents the divergence of observations from the predicted value, which can be considered as the momentary mean obtained from complex distribution. Thus, the normal distribution scoring can be applied on error values instead of the original observation values.

Therefore, all of the presented statistical methods are customizable for performing anomaly scoring and ranking on the obtained remainders from the state of the art prediction models. Further, some of these methods have been applied on batch data, but with small changes they are applicable on stream data, too.

### 5.1 Improved individual scoring

Improved scoring includes any scoring technique, which may help to have a better scoring than simply evaluating the pure measure of differences between observations and expectation obtained from behaviour predictor.

**Probability-based scoring** By assuming that data follows a specific distribution, the anomaly score can be computed based on their probabilities. The underlying idea in this technique is that the score assigned by ADS to data points (anomaly score) should be correspond to the statistic of data; so, anomaly scoring has a one-to-one relation with distribution function. Many studies have applied more or less the same approach and have obtained acceptable result. Intuitively, if  $x$  is less likely then it is more anomalous. In another words, an anomaly score  $A(x)$  respects the distribution ( $f$ ) if  $A(x) \leq A(y)$  if and only if  $f(x) \leq f(y)$ .

Even though, this approach seems very reasonable in under control conditions, but it is not as powerful in real-world data [11,25], because:

- Data dispersion is usually far from the known statistical distributions.
- Data is full of noises which may lead ADS into extreme false alert rate.
- It may ignore rarity aspect; with the same threshold very long tail distributions provoke many red flags versus shorter tails
- Obtained scores are not comparable and agreeable in complex configurations like having multiple cooperative detectors, like detectors for network traffic velocity and IP-distribution, with different models.
- Not able to address stream data, which need a dynamic model.
- The most efficient threshold of probability needs to be found

In this section, we mention some of related techniques, which have tried to improve the pure probability based scoring method.

**Bits of Rarity.** This approach has been proposed by [78] and the authors assume the obtained result from model provides a probability density distribution of values or errors, thus define the anomaly score of an event  $x$  with the probability density or mass function  $f$  as

$$R_f(x) = -\log_2(P_f(x))$$

The reason of using the negative sign is to give the highest anomaly score to the most different events. Moreover they apply log of the scores to stabilize their computations and distribute probabilities in larger span of values, because the original probabilities are between  $[0, 1]$ . But, we can argue that

- This technique does not improve the scoring method, just applies one-to-one transformation on the obtained result from predictor model, to present a more explainable ranking.

- It causes a huge difference in the obtained anomaly score from various detectors by exposing them to the range of natural number, which makes the compatibility of scoring more problematic.

**P-value scoring.** Traditionally, p-value of test statistic has been used to identify a point as an outlier, which is equal to rejecting the null hypothesis [73]. In fact, applying p-value makes more sense to find anomalies, because it is independent of Probabilistic Distribution Function (PDF). The other advantage of p-value is that it captures both rarity and dissimilarity of the distributions. In other words, it represents a ranking of all the observations by their dissimilarity and rarity, which are presumption of many anomaly detection definitions.

P-value based methods apply probabilistic scoring of anomalous degree, so can provide a sharp bounds on the *alert rate*, in terms of the threshold which only depends on the probabilistic description, not the data distribution. So they narrow down the frequency of alarms for any random distribution.

But, some essential limitations of p-value are:

- It has been customized to focus on extreme values, which are a subset of target anomalies. Strictly speaking, traced data points by p-value are just the cases that lie outside some convex hull of the most of the distribution mass [83,11].
- The choice of significance level at which null-hypothesis should be rejected matters; especially in the context of anomaly detection.
- This technique is just applicable in the settings that assume no other alternative hypothesis is available, but this is not the case in the real-world data.

**Bits of Meta-rarity.** Ferraught and et al. in [25] have tried to improve *bits of rarity* by addressing the non-comparability in different distributions or between different variables issue. So, they propose **Bits of Meta-Rarity** as a modified version, which provides the possibility of direct comparison by considering not the rarity of the event itself, but how rare the rarity of an event is. The formal definition of this measure is as the following:

$$A_f(x) = -\log_2(P_f(f(X) \leq f(x)))$$

By considering the random (discrete or continuous) variable  $X$  with the probability density or mass function  $f$  defined on domain  $D$ , the  $A_f : D \rightarrow R \geq 0$  and  $A_f(x)$  returns the anomaly score of  $x$ . Hence, scoring provides a strict weak ordering of observations by their anomalousness, so that  $x >_a y$  if and only if  $A_f(x) > A_f(y)$ . They discuss that when  $f$  displays the probability distribution of data, the probability that the anomalous degree exceeds the given rate of  $\alpha$ , would not be greater than  $2 - \alpha$ .  $P_f(A_f(x) > \alpha) \leq 2^{-\alpha}$  Therefore, the number of false alarms generated in this approach is regulatable because it only depends on the specified threshold not the distribution of data ( $f$ ). Moreover, based on the same theory, their approach provides the chance of comparing  $X$  and  $Y$ , generated by  $f$  and  $g$ , based on  $A_f(X)$  and  $A_g(Y)$ , respectively.

More or less the same technique has been applied in [83] to perform anomaly detection in application performance monitoring context, but the authors call it the generalized p-value (q-value). The main advantage of considering the probability of all values with less density function than the under investigation value ( $x$ ) is providing the ability of addressing anomalous area, which are not necessarily extreme values. For example, as figure 5 illustrates in a mixture of two well-separated univariate Gaussians, some points between the modes of these two distributions are expected to be very low probability; consequently they should be classified as anomalous, although they are certainly not extreme values.

But, the main drawback of these kinds of techniques is:

- Being highly dependent on  $\alpha$  as the false positive threshold. Moreover, they may confront a huge false negative rate in case of having clustered anomalies.

**Q-function based scoring** By presuming that the error distribution of the applied prediction method for normal observations should be normal, some studies fit a normal distribution on the obtained error and compute the anomalous scores based on Q-function as the tail distribution function [29,75]. In other words,  $Q(x)$  is the probability that a normal (Gaussian) random variable will obtain a value larger than  $x$  standard deviations. As an example, we introduce the approach taken in [49] to analyze error values based on Q-function. At first they train a stacked LSTM model to predict the next  $l$  values for  $d$  of the input variables so that  $d$  is less or equal to the number of dimensions of input TS dimensions. This method detects anomalies using the prediction error distribution. It predicts each of the selected

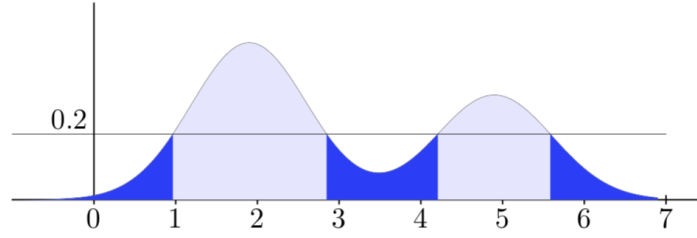


Fig. 5: Bits of meta-rarity of 0.2 in mixture of normal distribution [83]

$d$  dimensions of  $x(t) \in X$  for  $l < t \leq n - l$ ,  $l$  times, so that  $l$  is the prediction length. Then, computes an error vector  $e(t)$  for point  $x(t)$  as  $e(t) = [e(t)_{1l}, \dots, e(t)_{1l}, \dots, e(t)_{dl}, \dots, e(t)_{dl}]$ , where  $e(t)$  is the difference between  $x(t)$  and its value as predicted at time  $t - j$ . The prediction model trained on  $s_N$  is used to compute error vectors of each point in the validation and test sequences. In the next step, the error vectors will be modelled to fit a multivariate Gaussian distribution  $N = N(\mu, \Sigma)$ . The likelihood  $p(t)$  of observing an error vector  $e(t)$  is given by the value of  $N$  at  $e(t)$  (similar to normalized innovations squared (NIS) used for novelty detection using Kalman filter based dynamic prediction model [32]). The error vectors for the points from  $v_{N1}$  are used to estimate the parameters  $\mu$  and  $\Sigma$  using Maximum Likelihood Estimation. An observation  $x(t)$  is classified as ‘anomalous’ if  $p(t) < \tau$ , else the observation is classified as ‘normal’. The sets  $v_{N2}$  and  $v_A$  are used to learn  $\tau$  by maximizing  $F\beta - score$  (where anomalous points belong to positive class and normal points belong to negative class).

The main advantage of this technique is allowing for fast comparisons between new errors and compact representations of the prior ones [75,3]. But, the main limitation is that:

- This assumption is the foundation of many statistical regression methods like OLS, however, it is not very trustable, when parametric assumptions are violated and error values are not random, which is very likely in the data-driven methods [36].

**Similarity based scoring** The methods in this group assign anomaly score to the new observations based on their distance to the other groups or like KNN to a set of  $k$  neighbors. In addition to be originally supervised, they confront threshold setting problem during anomaly detection phase. How much similarity or distance is enough to be considered as close or far? Thus, the challenge of finding a good enough threshold is ongoing, but in a different step.

Therefore, some of the methods have tried to address the drawbacks of simple standard Euclidean distance metric by substituting it with more meaningful distance metrics, in one of the following categories, depends on data type [88]:

- Power distances. Distance measures which use a formula mathematically equivalent to the power ( $p, r$ ) as:

$$\left( \sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{r}}$$

Some of well-known measures in this category are Manhattan and Euclidean distance [18].

- Distances on distribution laws. Describes those measures based on the probability distribution of the dataset, like Bhattacharya coefficient [59] or  $\chi^2$  distance [18]
- Correlation similarities. Characterize the correlation between two datasets as a measure of similarity or distance, such as: Kendall  $\tau$  rank correlation, Learning Vector Quantization [41].

The influence of second and third groups is almost similar to probability-based and information theory based scoring, which have been discussed in their related contexts. So, some methods have aimed to improve similarity evaluation by focusing on power distances like Manhattan, Minkowski, or Hamming distance. For example, if data features have different distributions, Euclidean distance is not able to capture real distance of points from the mean of normal data. But, Mahalanobis distance [50] is one of the metrics that addresses this problem and computes the distance between the particular point  $x = (x_1, x_2, \dots, x_N)^T$  and the distribution with mean  $\mu = (\mu_1, \mu_2, \dots, \mu_N)^T$  as:

$$D_M(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}$$

It is a multi-dimensional generalization of the idea of measuring how many standard deviations away  $x$  is from the mean of  $D$ ; So, it is able to take into account variance and the covariance of the variables in addition to the average value.

[87] is an example of studies that has taken advantage from this measure during the detection phase to calculate the similarity of new data points against the normal profiles and assign anomaly score to them. While, some other studies have established threshold-based ADS by setting threshold and distance computation based on Mahalanobis formula [46].

In sum, this group of methods are valuable and regarding their improvements in similarity evaluation techniques and seem promising in measuring the error value of new observation versus normal expected error vectors, too.

**Extreme value theorem** Stiffer in [76], exploits the extreme value theorem to find distribution-independent bounds on the rate of extreme (large) values for univariate numerical time series. Their method does not require any manual threshold setting, but gets one parameter as the *risk factor* which controls the number of false positives.

The main goal of the extreme value theory is to find the law of extreme events. This theory has been initiated by [26] through stating that the extreme events have similar kinds of distribution, regardless of the main data distribution as long as it would be standard. The definition of Extreme Value Distributions (EVD) or extreme laws is:

$$G_\gamma = x \rightarrow \exp(-(1 + \gamma x)^{\frac{-1}{\gamma}}), \gamma \in \mathbb{R}, 1 + \gamma x > 0$$

$\gamma$  is called *extreme value index* and depends on the original distribution, for example it is zero for Gaussian ( $N(0, 1)$ ). Indeed, when events are extreme (i.e.  $P(X > x) \rightarrow 0$ , represents the tail of the distribution of  $X$ ), there are not very different possibilities for the shape of distributions tail, so  $G_\gamma$  can be fitted on them.

According to this theory, [76] has improvised an algorithm to compute the threshold based on the extreme values. In the first step, they compute a threshold  $z_q$  from  $X_1, \dots, X_n$  and risk  $q$ . To this aim, they, also, set a high empirical threshold  $t$  to retrieve the peaks over  $t$  and fit a Generalized Pareto Distribution to them. They have not provided any systematic algorithm to compute  $t$ , except one condition that it should be lower than  $z_q$ .

In the next step, the distribution of extreme values is inferable to obtain  $z_q$  as hard and  $t$  as soft threshold. During this process, it generates a set ( $Y_t$ ) including all the peaks observed bigger than  $t$ .

Then, as figure 6 illustrates the streaming anomaly detector (SPOT) updates  $z_q$  with the incoming data, while applying it as a decision bound. If a value exceeds  $z_q$ , flags it as abnormal; the anomalies are not taken into account for the model update. Otherwise, two scenarios may happen:

- Peak case:  $X_i$  is greater than the initial threshold, so adds the excess to the peaks set and updates  $z_q$
- Normal case:  $X_i$  is a common value

They have a second theory to update parameters, in case of having drift or non-stationary data, which has been explained in section 7.

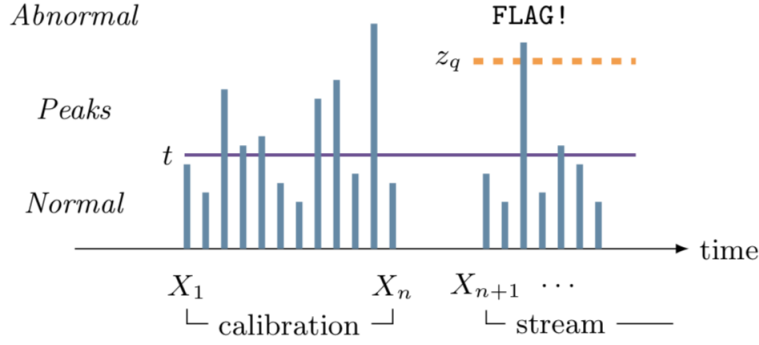
## 5.2 Unified scoring

Kreigel et al. in [43] bring up the issue of non-comparability and non-interpretability of different ADS results, because of scoring the events in various scales. So they have aimed to propose an unification approach to convert any arbitrary "anomaly factor" to the interpretable range of  $[0, 1]$  as an indicator of the anomalous probability. Their definition of anomaly is originated from Hawkins' idea in [31], that states: "a sample containing anomalies would show up such characteristics as large gaps between *outlying* and *inlying* observations and the deviation between outliers and the group of inliers, as measured on some suitably standardized scale". Their unification transformation method includes two main steps, where either step may be optional (depending on the type of score ( $S$ )):

- A regularization basically maps a score  $S$  onto the interval  $[0, \infty)$ , so that,  $Reg_S(o) \approx 0$  represents inliers and  $Reg_S(o) \gg 0$  indicates outliers
- A normalization to transform a score into the interval  $[0, 1]$

The applied transformation methods should be *ranking-stable*, which means it should not change the ordering obtained by the original score.

For any arbitrary  $o_1, o_2$ :  $S(o_1) \leq S(o_2) \rightarrow TS(o_1) \leq TS(o_2)$ . On this basis, they have proposed some manually crafting



**Figure 3: Anomaly detection overview**

Fig. 6: Updating anomaly score in stationary streams [76]

transformation to convert the obtained score from a limited number of methods to a comparable score in the interval  $[0, 1]$ . Except providing some general hints like applying a stretching interesting ranges while shrinking irrelevant regions, for scores with extremely low numeric contrast, they have not proposed any solid algorithm or direction toward applying the aforementioned mapping on any arbitrary AD scoring.

### 5.3 M-estimation scoring

Cl  men  on and et al. in [14] and later in [15] have focused on unsupervised anomaly detection from statistical point of view and addressed scoring and ranking of anomalies in multivariate domain space. He argues that in univariate domain space, means of tail estimation techniques indicates the anomalous level of data points and proposes M-estimator to simulate the same property in higher dimensional settings. In other words, M-estimator captures the extreme behavior of the high-dimensional random vector  $X$  by the univariate variable  $s(X)$ , which can be summarized by its tail behavior near 0; In the way that the smaller the score  $s(x)$ , the more abnormal/rare the observation  $x$  should be considered.

To have an estimation of density function, they apply the Mass Volume ( $MV$ ) curve as a functional performance criterion, which its optimal elements are strictly increasing transformations of the density, almost everywhere on the support of the density.

They, also, have provided a strategy to build a scoring function  $s^{(x)}$  which its  $MV$  curve is asymptotically close to the empirical estimate of optimum Mass Volume ( $MV^*$ ).

They optimize the functional criterion based on a set of piece-wise constant scoring functions. To this aim, their algorithm estimates a sequence of empirical minimum volume sets whose levels are chosen adaptively from the data. At the end, they overlay the feature space with a few well-chosen empirical minimum volume sets as figure 7 illustrates.

### 5.4 Improved Threshold Computation

In this part of our survey, we go over the techniques which have customized the finding threshold value in inference step toward mitigating the false alarm rate.

**ROC curve** ROC, which stands for Receiver Operating Characteristic curve is a graph showing the performance of a classification model at all classification thresholds. Many supervised studies use ROC or Precision-Recall curve (AP) to find the best error threshold for discretizing the ranked observations. Because of the potential data evolution scenarios, application of this technique in ADS which works on stream data is challenging, unless it regularly obtain the threshold based on the updated ROC curve. In addition, it is not suitable for unsupervised ADSs because of the lack of label. But, as [14,15,?] discuss  $MV$  curve which stands for Mass volume curve can be considered as ROC curve in the unsupervised anomaly detection settings and it provides the possibility of performing the same analysis. Finding threshold in unsupervised setting based on  $MV$  curve has been explained in 5.3.

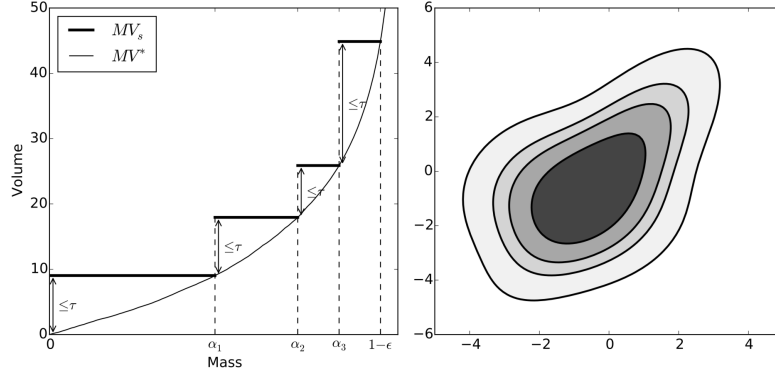


Fig. 7: left: Piece-wise adaptive approximation of  $MV^*$  and right: associated piece-wise scoring function [15]

**Double window scoring** Ahmed and et al. in [3] propose an HTM-based ADS to predict vector value of the next step; then compare the prediction result with the real vector to obtain an anomaly score based on their difference. The score that they obtain directly from model is bounded to  $(0, 1)$ , which is directly applicable as anomaly scores. However, this study aims to obtain anomaly likelihood, as a probabilistic metric, to be more robust in the noisy environment. To calculate the anomaly likelihood, they maintain two following windows:

- $W$ . A window of the last  $W$  error values. They model this distribution as a rolling normal distribution.
- $W'$ . A window of the recent short term history of prediction errors, so that  $W \gg W'$

Moreover, they update the sample mean, and variance in these windows continuously:  $\mu_t = \frac{\sum_{i=0}^{W-1} s_{t-i}}{W}$  and  $\sigma_t^2 = \frac{\sum_{i=0}^{W-1} (s_{t-i} - \mu_t)^2}{W-1}$ . And on this basis, the final anomaly likelihood is:

$$L_t = 1 - Q\left(\frac{\mu_{t'} - \mu_t}{\sigma_t}\right)$$

Where  $Q$  is Q-function. At the end they label a data point  $x_t$  as anomaly if  $L_t \leq 1 - \tau$  where  $\tau$  is a threshold parameter. Because of keeping two windows, this method is also potentially able to address the concept drift problem.

**Dynamic thresholding** In [36], a dynamic thresholding approach has been proposed for evaluating residuals, which is almost non-parametric. This method aims to address diversity, non-stationarity, and noise issues, by automatically setting thresholds for data streams characterized by varying behaviors and value ranges. The main idea is impressed by fitting normal distribution on the remainders, but has been configured toward fixing some of its drawbacks through the following steps:

**Errors and smoothing** After predicting the expected value  $y'(t)$  and error value as  $e(t) = |y(t) - y'(t)|$  for each step  $t$ ,  $e(t)$  is appended to a one-dimensional vector of errors:

$$e = [e(t-h), \dots, e(t-1), e(t)]$$

Which  $h$  determines the number of historical error values used to evaluate the current errors. Then, by applying an exponentially-weighted average, they smooth the set of errors to  $error_s$  and give it to the scoring step.

$$error_s = [e_s(t-h), \dots, e_s(t-1), e_s(t)]$$

**Threshold calculation and anomaly scoring** At this stage, a threshold is found that, if all values above are removed, would cause the greatest percent decrease in the mean and standard deviation of the smoothed errors  $e_s$ . The function also penalizes for having larger numbers of anomalous values ( $|e_a|$ ) and sequences ( $|E_{seq}|$ ) to prevent overly greedy behavior. Then the highest smoothed error in each sequence of anomalous errors is given a normalized score based on its distance from the chosen threshold.

To this aim, starting with a threshold  $\epsilon$  selected from the initial set, the appropriate anomaly threshold will be computed as follow:

$$\epsilon = \mu(e_s) + z(e_s)$$

while  $\epsilon$  is

$$\epsilon = \operatorname{argmax}(\epsilon) = \frac{(\Delta(\mu(e_s))/\mu(e_s)) + (\Delta(\sigma(e_s))/\sigma(e_s))}{e_a + (E_{seq})^2}$$

So that:  $E_{seq}$  = continuous sequence of  $e_a \in e_a$

$$\Delta(\mu(e_s)) = \mu(e_s) - \mu(\{e_s \in e_s | e_s < \epsilon\})$$

Same approach is taken to compute  $\sigma(e_s)$  and  $e_a = \{e_s \in e_s | e_s > \epsilon\}$ .

Values evaluated for  $\epsilon$  are determined using  $z \in Z$  where  $z$  is an ordered set of positive values representing the number of standard deviations above  $\mu(e_s)$ . Once  $\operatorname{argmax}(\epsilon)$  is determined, each resulting anomalous sequence of smoothed errors  $e_{seq} \in E_{seq}$  is given an anomaly score,  $s$ , indicating the severity of the anomaly.

$$s^{(i)} = \frac{\max(e_{seq}^{(i)}) - \operatorname{argmax}(\epsilon)}{\mu(e_s) + \sigma(e_s)}$$

Some of the drawbacks of this method are:

- Finding the best value for  $z$  is highly context dependent. So, the problem of finding the best value for a parameter remains, but in smaller scales.
- Adding gradual anomalies to data can lead the system toward increasing the threshold value and miss the real attack.

## 5.5 Sequence based scoring

Collective anomaly detection is extremely essential in a variety of application domains, such as in image processing, astronomy, and IDS. Even in some scopes like network IDS, collective anomaly detection is more meaningful, because attacks mainly happen in a sequence of operations, so each single command is not meaningful representation of any attack. The ADS obtain a birdview from the sequence to increase recall, while keeps low false alert rate. But, this opportunity comes in the cost of loosing realtime response time, unless ADS gradually performs anomaly detection (i.e. attaches each new observation to the so-far investigated sequence and regularly repeat AD process) [95]. There are many different methods that have applied collective anomaly detection like [4,95]. Two main strategies to compute the sequence anomaly score in prediction based ADSs are:

- Predicting the whole target expectations at once and consequently computing the anomaly score
- Computing the anomaly score of each observation in test sequence and summarizing their combinations using any arbitrary aggregation function on behalf of the whole target sequence

In this section, we review two well-known techniques which apply the collective relation of datapoints in scoring the whole sequence. To the best of our knowledge, these procedures have been only applied on the original values, but they seem to be applicable on prediction error, if the prediction error is not normal and includes implicit correlations.

**Information theory application in AD** Traditionally, information theoretic measures like *(Conditional) Entropy*, *Relative (Conditional) Entropy* and *Information Gain* were very popular techniques in tracking the likelihood of having anomaly in the data. The overall steps of ADS could be defined as the following:

- Measure regularity of train data and perform appropriate data transformation
- Iterate previous step if necessary till having high regularity dataset
- Train the profiling model
- Use relative entropy to determine the validity of model on the new observations

For example, based on the conditional entropy  $H(X|Y)$  on the obtained subsequences from system calls, the model can determine the  $n^{th}$  system call [47].



**Likelihood Ratio Method** The overall process of methods in this category is based on computing probability of any data point in the series stand on the values at previous few time steps. Therefore, TS anomaly score is a function of its datapoints anomaly score and the sequences with a very low generation probability should be marked as anomaly. Many ADS studies and applications in various areas like intrusion detection and speech recognition, have applied different modifications of this strategy. Three main high-level methods in this category are:

- Finite State Automata (FSA). After generating the model, if FSA ends up to a state which does not have any outgoing edge to the next value in the test sequence, it will be labeled as anomaly [80]
- Markov Models. Obtain the conditional probability of the observed symbols and their transition to each other and detect anomaly series based on their generation probability [77]
- Hidden Markov Models (HMM). An HMM is fitted on the training sequences and verify the likelihood of a test sequence generated by the learned HMM using decoding algorithms like *Viterbi algorithm* [95]

A very special property of HMM families that distinguishes them from the other sequential scoring models is their ability to consider the correlation and order between observations to assign a score to the whole sequence. Therefore, incompatibility of anomalous subsequences with previous and the next observation doubles its anomaly score. Moreover, it decreases the possibility of detecting short-term abrupt noises as anomaly because its influence does not dramatically change the overall score of the whole sequence.

## 5.6 Alarm Verification

This technique is one of the applied methods in the signature-based IDS, which its modified version is applicable in AD context, as well. The whole idea of this post-analysis technique is verifying whether the detected unusual cases will impact on system or there are some chances they be successful, then categorize them based on their seriousness [10]. There are two types of verification mechanisms:

- Active verification. Verifies online the generated alarm.
- Passive verification. Verifies the alarm versus a database including possible success cases.

Active verification is more promising in detecting zero-day attacks and applicable in the context of anomaly detection on stream data. Even though, active learning is more expensive, but based on the target domain, these kinds of methods can be near to real-time and very successful in mitigating the false-alarm rate. On the other hand, attackers can generate some spurious patterns of responses to misguide IDS to believe that attacks will fail. For example, in signature-based methods, there is a class of *Mimicry attacks*, attack sends a fake response as in case of normal scenario on behalf of server. Thus, IDS fails to detect malicious behavior and ignores the alarm [79].

## 5.7 Collective analysis

This section describes the methods that take advantage of the extra information like previous observations, contextual or correlation information to rescore or relabeled the observations.

**Voting-based methods** Applying a hybrid of techniques based on different detectors to perform anomaly detection is very promising and popular. Ranking alarms based on voting decreases the chance of raising false alarm. Even, some studies apply voting based on combination of the current generated alarms and obtained historical feedbacks from system administrator to rank alerts. Zohrevand and et al. in [95], to address over-fitting and decrease the false alarm rate, have applied a voting technique based on a hierarchy of HSMMs to assign anomaly score to the new observations. The models learned from shorter time intervals include more detailed description of data behavior but are very context sensitive. On the contrary, the models learned from longer period located in upper levels have a more global view and do not lose the information in transition between shorter time intervals. Confirmation procedure through this hierarchy works as the following: It considers a group of Reference Windows (RW), which their context are very similar to the current Test Window(TW). Thenceforth, checks the similarity of sequence transition of the extended TW and RWs:

- If both windows have a similar overall transition, the detected anomaly in the lower level is unreasonable, and its anomaly score should be decreased based on the similarity ratio.
- Otherwise, the anomaly score obtained in leaf nodes is increased according to the inverse of similarity ratio.

**Scoring in different levels of granularity** Modeling each detector independently also allows traceability down to the finer granularity level and decreases the chance of losing those low-level patterns. The obtained anomalies in those low-levels can later be grouped to ultimately find subsystem level anomalies. A very logical break-down in network IDS is modeling detectors in the node and network level to be able to trace both focused and distributed attacks.

Another approach, in this context, is applying break-down and aggregation on the time span to be able to trace and balance the scores based on their short and long-term stability. As we described in the previous section, [95] applies a hierarchical confirmation procedure to improve the accuracy. This method applies more general Markovian models in higher levels of hierarchy to verify the result of leaf nodes. The overall approach in the higher levels is very similar to the lowest level, except that the aim of these comparisons focuses on the transition of states instead of the real values of observations. Thenceforth, the method checks the similarity of sequence transition of the extended Test Window (*TW*) and Reference Windows (*RW*):

- If both windows have a similar overall transition, the detected anomaly in the lower level is unreasonable, and its anomaly score should be decreased based on the similarity ratio. It is noteworthy that temporal patterns happening due to any time shift — contraction or expansion of behavior on the time dimension — are the common reasons of such situations.
- Otherwise, the anomaly score obtained in leaf nodes is increased according to the inverse of similarity ratio.

It is worth noting that their framework evaluates the possibility of data drift and applicability of the trained model by sanity checking of the assigned sequence of transitions to *RW*, which will be explained more in Section 7.

Even in some studies, the hierarchy of granularity is based on space dimension. For example, to find the spatio-temporal object whose thematic attributes are significantly different from those of the other objects, [13] proposes a method based on multi-granularity and cluster differentiation, which includes the following four steps:

- Applies classification or clustering to find the regions,
- Reduces the spatial resolution of the data by aggregating them,
- Compares the result obtained in two different granularity levels, which can be done either by exploratory visualization analysis or visual data mining. The objects which are found in step 1 but not in step 2 will be considered as potential anomalies,
- Evaluates the temporal neighbours of the suspected anomalies points to finalize the candidates.

**Alarm correlation** Constructing the potential attack scenarios based on the aggregation of data is one of the other techniques in mitigating false alarm. This aggregation can be done by grouping a bunch of alarms possibly generated by different detectors or in different places of network or sequence and then reconstructing the attack scenarios. Performing correlation analysis and generating the possible scenario helps to extract some concrete and interpretable inferences, which decreases the false-alarm rate. As aforementioned this correlation can be considered in different levels of abstraction, like:

- Correlation of events from same or heterogeneous detectors
- Correlation of events in one detector through time dimension
- Correlation of events through different nodes in network

In the context of signature-based IDSs, a generic view of alarm correlation is as the following [35].

1. Alarm normalization. to bring all of the alarms to the same format and scale them.
2. Alarm clustering. To group the alarms generated by different detectors or sensors [51]. Like clustering alarms with shared causes or alarms belong to the same vulnerability, alarms in the same session, or similar alarms based on their feature values.
3. Alarm correlation. Find the relation of alarms intra and inter clusters.
4. Intention recognition. Identify the attacker's plan and report an interpretable and robust view of attack scenario.

In the following, we introduce some of the existing alarm correlation analysis techniques reviewed by [35].

### Multi-step correlation

By assuming that usually a sequence of actions is required before a malicious event in the system, finding the correlation of the observed anomalies may help to assert a malicious event, before it happens [54,35]. A systematic approach, to perform this step can potentially be very robust and prosperous in finding complicated scenarios. However, if the

correlation analysis algorithm would be very strict about pre and post conditions, missing one event (a few false negatives) can hugely influence on its performance. Some studies consider IDS alarms as transactions and apply frequent pattern mining to track frequent alarm combinations as indicators of malicious sequences before intruder or system fault causes the main and critical event [71].

#### Causal relation based correlation

It verifies the causality correlation between existing variable and detectors. For example by applying Bayesian network of nodes as alarms and edges as relationships obtained from time-based coincidence of alerts and their mutual information, system can generate hyper-alerts [65]. In addition, because of the existing lags between detectors, finding these kinds of causality relations can improve ADS result twofold in 1) Finding malicious events through less dangerous cases and detectors, before critical detector raises the alarm and 2) mitigating false alarm by ignoring the red flags that are happening with expected lag in the same or another detectors, but because of the same reason, which is not desirable in terms of false alarm rate. For instance, Figure 8 illustrates an attack and its influence on a sensor and the time it takes for the system to stabilize, so meantime the detectors may generate many more anomaly indicators because of the same reason.

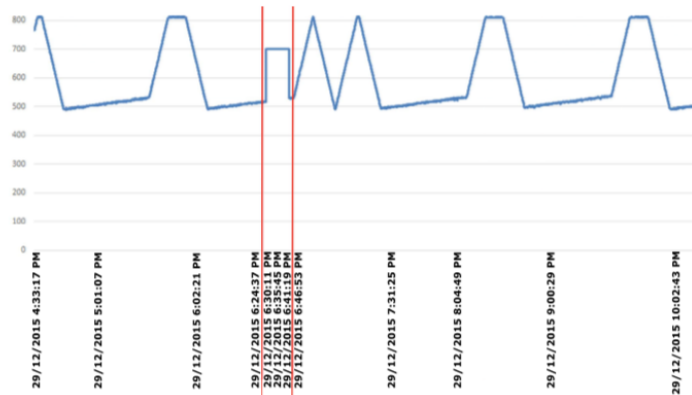


Fig. 8: Influence of attack on system till it stabilize [42]

**Subsystem graph based correlation.** Since a system may include many subsystems, they may have cascading influence on each other. For example, an attacker may find a weak protected subsystem because of being known as low impact vulnerability and use it as footrest to reach most critical systems and servers in the network. In critical infrastructures, this cascading influence is even more highlighted.

Therefore, some IDSs identify the possible penetration to critical systems represented in the form of a graph, hence are called as attack graphs [68]. In other words, these methods focus on existing dependency and interconnections in the network of systems and balances anomaly scores based on the penetration paths which the events may cause. It is worth noting that this method is taking advantage of an extensive knowledge base as meta data which indicates what is the critical rate of the subsystem in the current situation like importance of applications running, or system known vulnerabilities. Finally, the IDS will be able to balance anomaly score based on the critical rank of the target [81].

**Hybrid of correlation.** There is the possibility of applying a combination of correlation-based analysis, through multiple layers. Then, they combine the result of extracted correlations on the span of time and location to decide about risk factor of the under investigation alarms. A preliminary version of this approach in IDS context has been proposed by [24]. These kinds of analysis are more complex, but by providing a bird view can reveal much more information than individual alarms.

**Correlation analysis in ADS.** Even though, active correlation analysis may influence on response time, but by distributing computations and taking advantage of the pre-extracted knowledge, it can be almost performed in real-time. Moreover, the minimum requirement to apply it in ADS scoring is possibility of having access to some predefined correlation between subsystems, systems and detectors. Generally speaking, finding complicated alarm correlation is

very challenging, especially in AD contexts, which is happening in the world of uncertainties. But, taking advantage of time based correlation to perform soft scoring seems very promising in ADSs. For example, let assume that post-filtering phase of ADS adjusts the assigned scores to data points by the core of AD, based on the continuity of the observations. So, if  $x$  and  $y$  be two consequent observations which have obtained the same score from AD algorithm ( $s_A(x) = s_A(y) = \alpha$ ); After passing through post-filtering algorithm  $s_A(y) > s_A(x)$ , because the chance of having two noisy observations in a row is less than one. The aforementioned examples illustrates one of the simplest scenarios that correlation-based filtering can help ADS; beyond question, to improve the accuracy level, the filtering phase should be substituted with more advanced algorithms to capture complex correlations. For instance, sometimes attackers intentionally import a few normal behaviour into their sequence of malicious action to misguide the system. Thus, if ADS relaxes the continuity condition and be able to remember and trace what has been happened in the previous steps will be capable of capturing these kinds of correlations, as well. Applying a similar technique in [82,84] have proved the advantages of this approach in signature based IDSs.

It is worth mentioning that there another line of research on correlated anomalies is Correlated Anomaly Detection (CAD), which has been applied on streaming data as a type of group AD in botnet detection, financial event detection, industrial process monitor, etc. There is a big difference in the approach taken in CAD versus the term AD that we are reviewing; in the correlation analysis, the main aim is finding possible correlation between anomalous events to mitigate false alarm, while the principal presumption in CAD is that the normal data entries are not strongly correlated, so strong correlations can be unlikely and indicator of anomalousness. Finding malicious server visits from distributed denial of service (DDOS) attackers or the correlated price changes in stock markets are some of the specific targets of this category of methods [92,64].

**Alarm mining** Each anomalous event (point or collection of points) has its specific feature values. Based on the available history, ADS can perform mining to summarize them into either TPs or FPs. Characteristics learned during the mining stage are useful to define meta signature for future alarms or adjust their scores. System may have a second level of analysis to cluster [38] or classify [60] the similar types of raised alarms to analyze about the seriousness level of alarms. Even though, these group of techniques are almost automatic and able to correlate zero-day alarms, they are not straightforward because of the dynamics in the underlying network and system context; therefore, the mining-based filtering algorithm should be able to handle the evolution and drift in data.

**Apply contextual information** Some methods profit performing AD in the presence of more contextual information. Applying available information to improve precision may be applicable in the following three ways:

- Multivariate analysis. Consider contextual information as extra features to the existing data and train a multivariate model on data. As [95] considers meaningful contextual features like temperature and time differences, and fit multivariate model, which leads into a higher rate of recall as well as less false alarm.
- Error balancing. These kinds of methods, perform an early prediction and readjust their result by benefitting from contextual information. For example, [96] proposes a deep learning based framework for time series analysis and prediction by ensembling parametric (TBATS) and non-parametric (deep network) methods. This study applies time-based generated features as contextual information to balance the obtained predicted values from two different methods.
- Post verification. Some studies apply the contextual information in post-verification phase. The original idea is to detect the potential anomaly candidates and prune them based on contextual information. For example, in [66] a framework has been proposed to detect anomalous track segments within operator's geographical area of interest from a received AIS data stream with focus on reducing false alarm rate. In this process, contextual verification process is carried out when a potential anomaly is detected. Thus, the end-user will be just notified about the confirmed anomalous behaviour by the contextual verification step; Otherwise, a detected potential anomaly is deemed to be a false alarm.

**CUSUM** CUSUM, which stands for cumulative sum is one of the popular techniques to monitor the changes in continuous processes [56]. The CUSUM method can potentially, be applied to the changes of any property of the process; but, originally it has been applied on the mean to detect changes of a sequential process or time series, like break out, peeks and sudden growth.

The one-sided CUSUM calculation technique is as follows:

$$C(t) = \text{Max}(0, F(t) - (\tau + L) + C(t - 1)) | F(t) = 0$$

When,  $F(t)$  represents the feature to calculate its cumulative sum. And  $\tau$  represents the optimal value of the under investigation factor of process; and  $L$  is known as the slack value or allowance, which determines the relax boundary between normal and anomalous. So, if the  $F(t_i)$  be close enough to the  $\tau$ , so  $C(t)$  will remain small. However, as soon as having a positive shift in the property, the  $C(t)$  value will increase, rapidly. Thus, whenever  $C(t_i)$  exceeds a predefined threshold for interval ( $H$ ), the red flag will be raised and the CUSUM value will be reset to zero [16]. Also, AD can be configured based on two-sided CUSUM (i.e. high and low cumulative sums).

But, generally speaking, presuming that processes are stationary makes this technique incapable in detecting anomalies in the real world stochastic setting. But some studies like [28] have applied it on the obtained error values from their predictive models to find the anomalies. For example, authors of [28] have used LSTM-RNN for learning the temporal behaviour of the data in CPSs. After obtaining the error vector, using tow-sided CUSUM  $SH_i$  and  $SL_i$ , they calculate the potential positive and negative changes.

In this way, ADS do not need to fix a very specific threshold for the prediction error, which may be different from one context to another. But, setting CUMSUM threshold is more general and is highly correlated with expectations about error distribution.

Another benefit of using CUSUM is that it provides the possibility of deciding based on accumulated knowledge to detect the anomalous cases and is able to detect small deviations over time thus reducing the number of false positives. Moreover, it is very quick and about to real-time.

However, at the end of the road, it needs determining upper and lower thresholds as well as slack value. In addition, it is prone to detect many noises as anomalies.

## 6 Post-hoc mitigation of false alarm

The precision of all the ML methods is highly dependent on the rate of comprehensiveness of available observations during the training phase. Especially, for prediction-based anomaly detection approaches, this dependency is more highlighted because they are influenced by training data in two ways:

- To find an accurate prediction model
- To set a precise error threshold

Moreover, in time series and stream data analysis context, observations are more prone to trend and gradual or abrupt evolution, which should be addressed in the applied ADS. At large scales, query and processing historical data in real-time scenarios is expensive, while lack of history can lead to false positives that are only deemed anomalous because of the narrow context in which they have been evaluated. Additionally, when extremely high volumes of data are being processed a low false positive rate can still overwhelm human reviewers charged with evaluating potentially anomalous events [36].

Therefore, the strategies reviewed in this section, utilize user-feedback and history of observations to readjust the assigned score threshold, which may cause reclassification of the data points. These types of techniques are very beneficial in non real-time configurations like fraud detection or other contexts, which ADS should produce suspicious cases as a prioritized flagged cases and there is no need of instant response. But, some ADSs also, take advantage of the obtained knowledge from this step to improve their scoring process.

### 6.1 Maximum error value based pruning

[36] proposes a pruning procedure based on maximum value of all the observed anomalous points to mitigate the false alarm rate, while keeps limited memory and computes cost. This method keeps track of a set,  $e_{max}$ , containing  $\text{max}(e_{seq})$  for all error sequences ( $e_{seq}$ ) sorted in descending order. Moreover, they add the maximum smoothed error that is not anomalous to this vector

$$e_{max} = \text{max}(e_s \in e_s \in E_{seq} | e_s \in e_a)$$

Then, they go through this sequence and compute the percent decrease as

$$d^{(i)} = (e_{max}^{(i-1)} - e_{max}^{(i)}) / e_{max}^{(i-1)}$$

So that at each step  $i \in 1, 2, \dots, (|E_{seq}| + 1)$ ,  $E_{seq}$  is the sequence of the observed errors in anomalous sequences. If at some step  $i$ , a minimum percentage decrease  $p$  is exceeded by  $d^{(i)}$ , all  $e^{(j)} \in e | j < i$  and their corresponding anomaly sequences remain anomalies. If the minimum decrease  $p$  is not met by  $d^{(i)}$  and for all subsequent errors  $d^{(i)}, d^{(i+1)}, \dots, d^{(i+|E_{seq}|+1)}$ , those smoothed error sequences are reclassified as nominal.

## 6.2 Rarity based post pruning

The main idea in this technique is based on the rarity assumption in anomaly detection context. For example in [36] if they frequently observe anomalies in the same magnitude ( $s$ ), consider a minimum score,  $s_{min}$ , such that future anomalies would be re-classified as nominal if  $s < s_{min}$ . Prior anomaly scores for a stream data can be applied to set an appropriate  $s_{min}$ , depending on the desired balance between precision and recall. It is worth mentioning that this pruning step is tricky and like direct rarity-based scoring methods is highly prone to be misled by attackers and give green card to anomalous behaviours.

## 6.3 Active-learning based scoring

If the ADS has a mechanism by which users can provide labels for anomalies, the system can take advantage of the provided labels to set  $s_{min}$  for a given stream. So, the threshold of anomaly pruning can be set based on the obtained lower and upper bound score of the confirmed anomalies. Authors in [17] have aimed to perform AD by designing a hyper-plane that passes through the uncertainty region based on the learned decision boundaries with active learning. To this aim they should find the best thresholds to customize hyper-plane, but they want to minimize the interaction with end-user based on asking for most informative minimal subset. So, by relying on the specific property of ensemble methods, which involve searching for the optimum non-homogeneous decision boundary, they take advantage of ensemble method to perform AD. If the ensemble members would be ideal, the scores of true anomalies will lie in the farthest possible location in the positive direction of the uniform weight vector  $w_{unif}$  by design, so it works very well. However, since this is not the case and some ensemble members have deficiencies in practice, the obtained weight vector ( $W^*$ ) is usually diverged from  $w_{unif}$  by some angle, which may result in many false alarms. Therefore, since the misalignment is usually small, they apply active learning to learn the optimal weights efficiently, based on the top-ranked instances close to the decision boundary.

## 6.4 Improved predictor models

Last but not least, in this section, we briefly go through different predictors that have improved the behaviour model step. Because, applying a predictor model which is able to capture latent complex patterns in the data will unquestionably contribute to higher recall level and lower false alarm rate.

**Robust anomaly detection** Since, the number of data points in an anomalous cluster or their distance to the normal cases should not influence on the decision boundary, in the way that model loses a few anomalies (masking) or labels some normal cases as anomaly. In other words, the applied method should be robust enough to not diverge under the presence of masking and swamping phenomena. Some studies like [91,69] focus on improving the quality of the underlying applied method like applying robust PCA instead of PCA, or robust matrix factorization, which can end up better false alarm and recall rate.

**Advanced data driven models** In the era of big data, the volume and variety of data has been hugely increased, so that the traditional algorithms are not scalable enough to handle them. Moreover, since the data is frequently coming from different generators, they are not optimal enough to capture and model the complex patterns in data behaviour. In addition, high-dimensionality is an integral part of datasets and most of the traditional models are incapable of reducing dimensions or selecting the most important features out of thousands of dimensions. While, deep models are flourishing these days and they are qualified methods to capture complex patterns in large scales and high-dimensional data generated by various sources. There are many studies, which have applied different deep networks as predictor model to perform anomaly detection. These predictors can be categorized into three following groups:

- Discriminative models: RNNs like Long Short Term Memory (LSTMs), Gated Recurrent Units (GRUs) [49,53], Convolutional Neural Networks (CNNs) [85], Deep Neural Networks (DNNs) [5].

- Generative models: Different version of AutoEncoders (AEs) [72,7] and Sum-Product Networks (SPNs) [61].
- Generative Adversarial Networks (GANs) [74]

As aforementioned, most of this techniques have been reviewed in the update surveys and we skip repeating them in this study and refer the reader to the following surveys [44,1,52].

**Hybrid models** These days, ensemble and hybrid methods, as the winners of the most of the machine learning competitions, utilize the information fusion concept in slight different ways. Ensemble models by combining multiple but homogeneous, weak models, and hybrid methods, by combining heterogeneous and usually different machine learning approaches, are able to lead into considerably higher quality and performance solutions [?]. The main reason behind grouping the same or different methods is that none of the existing models can be considered as a perfect approach, because of:

- The lack of sufficient data to represent data distribution
- The locally optimal result caused by dependency on starting points
- Functionality of data cannot be modeled based on a single hypothesis, but better approximated by weighted sum of several ones

Since, AD is an open-ended definition in continuously evolving context with ambiguous definition of the normal regions, (i.e. there is no precise boundary between normal and anomalies), hybrid and ensemble models are promising to be able to find the optimum non-homogeneous decision boundaries. Among the existing studies of hybrid predictor modeling or ADS, some have applied very interesting logic to combine model-driven and data-driven methods to cover the potential weaknesses of each side. Many of them have considered Neural Networks or deep learning as suitable candidates for data-driven model. For example [39,20,96] apply combination of ANN and ARIMA for TS, water quality and photo-voltaic power generators forecasting, time series forecasting, respectively.

Another approach toward taking advantage of combination of models in ADS context is deep hybrid models, which is based on two step learning models. Generally, a traditional algorithms like one-class Radial Basis Function (RBF), Support Vector Machine (SVM) classifiers [23,37] or another complex model [89,63] be fed with the reduced, representative features learned within deep models.

## 7 Concept Drift

As aforementioned, addressing drift of data over time is another important matter in anomaly detection, which influences scoring and consequently the frequency of false alarm. In such cases usually model should be retrained and updated, which causes losing some of the pre-acquired knowledge. Some methods by trusting the accuracy level of the algorithm based on its potency in earlier stages, consider any mismatch in the number of expected anomalous cases, can indicate on the two following matters: 1) There is another source of event in observations or 2) The model is not tuned. In either cases, exploration of the different observations will provide insight into both the state of the system and potential changes within [25]. In this section, we review a few of the other proposed strategies to address this matter.

### 7.1 Ratio of anomalies

Generally, the class of methods, which their anomaly scoring is explicitly or implicitly depend on the rarity and frequency of red flags, can indicate the concept drift by observing any manifest change in the ratio of red flags. For example, [25] states that based on the provided upper bound, any deviations in the number of anomalies will indicate that the selected model does not match the real generating distribution and should be tuned or modified.

### 7.2 Ratio of anomalies in the last data points

This group of methods rely on the ratio of the detected anomalies in the limited range of recent observations ( $W$ ). In the simplest version, this ratio can be defined as

$$Rate_A = \frac{Number\ of\ anomalous\ points\ in\ W}{|W|}$$

If  $Rate_A$  would be bigger than a threshold then the concept drift has been happened and model should be updated [19]. Finding this threshold is very problematic, so [57] suggests that this ratio be substituted with cumulative Distribution Function (CDF) of a normal distribution and if complement of CDF on ratio would be smaller than a small threshold, indicates a concept drift.  $CDF_X(x) = P_X(X > x) < \tau | x = Rate_A$  Finding the optimized value for two main parameters of  $W$  and  $\tau$  is problematic, though.

### 7.3 Distribution of the last data points

This method is not dependent on anomaly detector and basically works based on the data distribution. Model keeps track of statistical distribution of the initial data set ( $S_I$ ) and dynamically obtains the data distribution through the last observed data points ( $S_W$ ). These two distributions should be compared, which is possible through various statistical tests like K-S test, which has been applied in [67]. If the result of test with enough confidence level ( $\alpha$ ) shows that they are not similar then the concept drift has been occurred. Again, same to the previous method, finding the best choice of  $W$  and  $\alpha$  may be very challenging.

### 7.4 Reference window based scoring

As aforementioned in [95] a hierarchical HSMM-based ADS is proposed. This method creates an online normal sequence of states based on the context of test windows in order to verify and identify their anomaly level. But, to address the concept drift and potential divergence of HSMM models from data overtime, they verify the abilities by fitting them on genuine normal Reference Windows ( $RW$ ). They, therefore, reconstruct the observation sequence of each  $RW$ , called decoded sequence, using Viterbi algorithm and the emission matrices, and compare the corresponding observations in real and decoded sequences. Being the assigned sequences of transitions to  $RWs$  very far from their regression models implies that the trained Markovian models are not up-to-date. This matter happens in cases that the frequency of mismatches would be greater than a threshold  $\theta$ , so the ADS marks the sequence as unfitted and if the number of unfitted  $RWs$  would be more than  $\tau$  update model based on the new observations. Meantime, ADS switches to regression-based anomaly scoring till obtain the updated HSMMs. This model depends on the two parameters  $\theta$ ,  $\tau$ , but  $\tau$  can be derived from anomaly scoring threshold. As explained in Subsection 7.4, this study applies data drift evaluation through the whole hierarchical analysis.

### 7.5 Anomaly scores distribution for the last data points

[3] works based on two long ( $W_l$ ) and short ( $W_s$ ) window of the last observations, so that  $W_l \gg W_s$  and compares the distribution of anomaly scores in these windows with each other and if the applied test rejects the null hypothesis, means that concept drift has been happened. The technique proposed in [57] by taking similar strategy like [3], proposes comparing two dynamic empirical distributions of anomaly scores. Let assume  $L$  and  $M$  as distribution of the anomaly scores of all the last  $l$  and  $m$  data points, respectively, so that  $l \gg m$ . Then, they use K-S test as the previous methods to compare these two distributions and detect the concept drift if the null hypothesis can be rejected on the level  $\alpha$ . Using this method is challenging in cases that ADS only produces discretized binary labels of normal and anomaly.

### 7.6 Drift detection based on extreme values

In [76] an algorithm has been proposed to overcome drift and seasonality issues by updating the threshold based on the observed peaks in the moving average and absolute values. It models an average local behavior on the last  $d$  observations and assumes that the sequence of local variations are stationary processes. Then applies SPOT algorithm mentioned in 5.1 on both absolute values  $X_i$  and relative gaps and updates the threshold during performing anomaly detection as illustrated in Figure 9.

### 7.7 Drift detection in tree-based ensemble

[17] proposes an algorithm to detect drift in ensemble ADSs. To this aim, they have applied Kullback–Leibler divergence (KL-divergence -  $D_{KL}$ ), which highlights how one probability distribution is different from a reference probability distribution. To determine the sub-tree, representative of sub-models in this study, which should be replaced they consider each leaf node as histogram bins to estimate the data distribution based on. By assuming the number of total



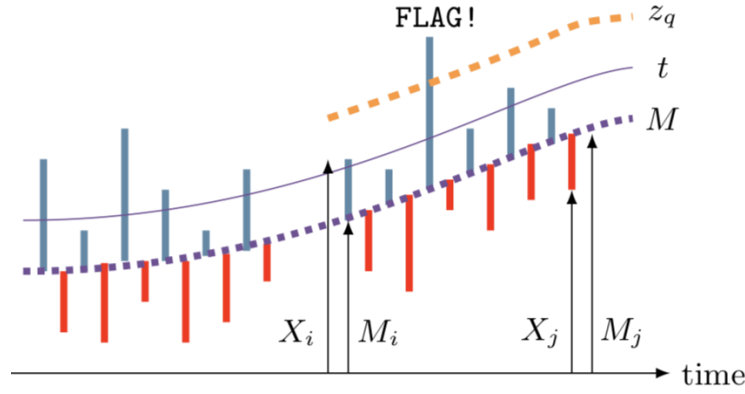


Fig. 9: Updating anomaly score in streams with drift [76]

trees would be  $T$  and  $t_{th}$  tree, be called as  $\tau_t$ , they apply the same window, which  $\tau_t$  and the others are generated based on; to keep track of the baseline distribution for  $\tau_t$  as  $p_t$ . After computing the baseline distributions for each tree, they estimate a threshold for drift  $q_{KL}$  by sub-sampling. When the model observes a new window ( $w$ ), it computes the new distribution  $q_t$  (for  $\tau_t$ ) for this  $w$ . If  $q_t$  differs from  $p_t$ , significantly ( $D_{KL}(p_t||q_t) > q_{KL}$ ) for remarkable number of trees ( $2T\alpha_{KL}$ ), then they replace all of these trees based on the new observed window. Finally, if any tree in the forest is replaced, they update baseline densities for all trees with the data in the new window. The algorithm seems very promising in tree-based methods, but the concept of sub-model is not very clear in the other contexts. Moreover, there are some parameters like size of windows,  $\alpha$ , divergence of distributions threshold, number of sub-trees as threshold, which should be optimized.

## 8 research questions

After studying various techniques for false alarm mitigation we induct the following questions for the research community.

- *Evaluation on a common dataset*: We find most of the works evaluate their techniques on a local custom dataset. The performance of system analyzed in terms of false positives reduction ratio on a common dataset will help understand their usefulness and applicability. working based on benchmarks like NAB [45] as benchmark designed to evaluating real-time anomaly detection algorithms, is promising toward comparing the capacities of different techniques.
- *Evaluation based on a common scoring mechanism*: The reported result for many methods are not clear and do not present method abilities in discovering different challenges like giving credit to the finding anomalies earlier and adjusting to changed patterns. The applied measures in NAB address some of the challenges that we mentioned in Section 4. Thus, collecting measures and labeled datasets which evaluate methods in all respects is a necessity.
- *Uniformity*: An uniform format to show the scores and priority is required as most of the techniques use their own custom range and format. In addition to providing comparability, this will help using the processed information by a combination of ADSs. As aforementioned, some studies have tried to propose a score unification process, but they do not propose a comprehensive algorithm that automatically transfer scores from system  $A$  to the target system, without missing any information.
- *Performance*: Many of the works found in literature, ignore performance aspects of their techniques. How much effort is required for the algorithm to execute is also one of the important characteristics, particularly in real-time contexts like IDS to be able to control damage.
- *Addressing drift in data*: Almost all of the real-world stream datasets include gradual or sharp shift over the course of time. Thus, it is very important for any technique to update itself using incremental learning or other techniques to address drifts and evolutions which may happen in the target domain.

## 9 Conclusion

In this paper, we have presented a survey of false-alarm mitigation techniques found in the literature at the time of writing. We have also evaluated their main advantages and disadvantages in this context. The Figure 4 presented in the Section 5 gives an organized overview of all the techniques. To the best of our knowledge, there is a serious lack of a comprehensive survey, which covers this key step of anomaly detection toward mitigating false alarm rate. Usually, studies focus on improving the main phase of profiling-based ADSs which is behaviour tracker, while final scoring has a very key role in the applicability of the system. We have presented several techniques in this survey. In addition to *Predictive models improvement*, which focus on improvement of predictive models to be able to extract more hidden patterns in data; the *Scoring metrics improvement* group have focused on taking advantage of rarity and probability values. The other essential techniques under the representation of *Improved threshold Computation* overcome the difficulties of finding the best threshold to discretize the assigned scores to observations to raise the alarm. In addition, the techniques under *Post-hoc pruning* are another type of strategies toward automatically updating threshold related parameters, but based on the running system performance. The strategies under *Collective analysis* are aiming to rescale the assigned anomaly scores in a collection of information, like their correlation with each other or their contextual information. Some of the strategies in this category are coming from signature-based IDSs, which seem very promising in the context of ADS, also. We explained about *Sequence based scoring* techniques and their application toward mitigating the false alarm rate by observing as part of a sequence rather than an individual case. Last but not least, a few of the most important methods to indicate data drift and the permanent evolution of system has been reviewed in section 7, which can prevent having a high false alarm rate.

In spite of these all known techniques there are still issues to be addressed. We have also enlisted few of the research questions at the end. In our opinion future research need to address these research questions which will improve usability of the proposed techniques. Further, subject to the criteria defined, one principal key to being successful is the adaptability of techniques in handling all specific features of data.

## References

1. Adewumi, A.O., Akinyelu, A.A.: A survey of machine-learning and nature-inspired based credit card fraud detection techniques. *International Journal of System Assurance Engineering and Management* **8**(2), 937–953 (2017)
2. Agyemang, M., Barker, K., Alhaji, R.: A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis* **10**(6), 521–538 (2006)
3. Ahmad, S., Lavin, A., Purdy, S., Agha, Z.: Unsupervised real-time anomaly detection for streaming data. *Neurocomputing* **262**, 134–147 (2017)
4. Ahmed, M.: Thwarting dos attacks: a framework for detection based on collective anomalies and clustering. *Computer* (9), 76–82 (2017)
5. Akhter, M.I., Ahamad, M.G.: Detecting telecommunication fraud using neural networks through data mining. *Int. J. Sci. Eng. Res* **3**(3), 601–606 (2012)
6. Amarasinghe, K., Kenney, K., Manic, M.: Toward explainable deep neural network based anomaly detection. In: 2018 11th International Conference on Human System Interaction (HSI). pp. 311–317. IEEE (2018)
7. An, J., Cho, S.: Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE* **2**, 1–18 (2015)
8. Bakar, Z.A., Mohamad, R., Ahmad, A., Deris, M.M.: A comparative study for outlier detection techniques in data mining. In: 2006 IEEE conference on cybernetics and intelligent systems. pp. 1–6. IEEE (2006)
9. Ben-Gal, I.: Outlier detection. In: *Data mining and knowledge discovery handbook*, pp. 131–146. Springer (2005)
10. Bolzoni, D., Crispo, B., Etalle, S.: Atlantides: An architecture for alert verification in network intrusion detection systems. In: *LISA*. vol. 7, pp. 1–12 (2007)
11. Bridges, R.A., Jamieson, J.D., Reed, J.W.: Setting the threshold for high throughput detectors: A mathematical approach for ensembles of dynamic, heterogeneous, probabilistic anomaly detectors. *arXiv preprint arXiv:1710.09422* (2017)
12. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. *ACM computing surveys (CSUR)* **41**(3), 15 (2009)
13. Cheng, T., Li, Z.: A multiscale approach for spatio-temporal outlier detection. *Transactions in GIS* **10**(2), 253–263 (2006)
14. Cléménçon, S., Jakubowicz, J.: Scoring anomalies: a m-estimation formulation. In: *Artificial Intelligence and Statistics*. pp. 659–667 (2013)
15. Cléménçon, S., Thomas, A., et al.: Mass volume curves and anomaly ranking. *Electronic Journal of Statistics* **12**(2), 2806–2872 (2018)
16. Das, K., Moore, A.: Searching through composite time series
17. Das, S., Islam, M.R., Jayakodi, N.K., Doppa, J.R.: Active anomaly detection via ensembles. *arXiv:1809.06477* (2018), [Online; accessed 19-Sep-2018]

18. Deza, M.M., Deza, E.: Encyclopedia of distances. In: Encyclopedia of Distances, pp. 1–583. Springer (2009)
19. Ding, Z., Fei, M.: An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. *IFAC Proceedings Volumes* **46**(20), 12–17 (2013)
20. Egrioglu, E., Aladag, C.H., Yolcu, U.: Fuzzy time series forecasting with a novel hybrid approach combining fuzzy c-means and neural networks. *Expert Systems with Applications* **40**(3), 854–857 (2013)
21. Emmott, A., Das, S., Dietterich, T., Fern, A., Wong, W.K.: A meta-analysis of the anomaly detection problem. *arXiv preprint arXiv:1503.01158* (2015)
22. Emmott, A.F., Das, S., Dietterich, T., Fern, A., Wong, W.K.: Systematic construction of anomaly detection benchmarks from real data. In: Proceedings of the ACM SIGKDD workshop on outlier detection and description. pp. 16–21. ACM (2013)
23. Erfani, S.M., Rajasegarar, S., Karunasekera, S., Leckie, C.: High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition* **58**, 121–134 (2016)
24. Feng, C., Peng, J., Qiao, H., Rozenblit, J.W.: Alert fusion for a computer host based intrusion detection system. In: null. pp. 433–440. IEEE (2007)
25. Ferragut, E.M., Laska, J., Bridges, R.A.: A new, principled approach to anomaly detection. In: Machine Learning and Applications (ICMLA), 2012 11th International Conference on. vol. 2, pp. 210–215. IEEE (2012)
26. Fisher, R.A., Tippet, L.H.C.: Limiting forms of the frequency distribution of the largest or smallest member of a sample. In: Mathematical Proceedings of the Cambridge Philosophical Society. vol. 24, pp. 180–190. Cambridge University Press (1928)
27. Gamboa, J.C.B.: Deep learning for time-series analysis. *arXiv preprint arXiv:1701.01887* (2017)
28. Goh, J., Adep, S., Tan, M., Lee, Z.S.: Anomaly detection in cyber physical systems using recurrent neural networks. In: 2017 IEEE 18th International Symposium on High Assurance Systems Engineering (HASE). pp. 140–145. IEEE (2017)
29. Goldstein, M., Uchida, S.: A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one* **11**(4), e0152173 (2016)
30. Gupta, M., Gao, J., Aggarwal, C.C., Han, J.: Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering* **26**(9), 2250–2267 (2014)
31. Hawkins, D.M.: Identification of outliers, vol. 11. Springer (1980)
32. Hayton, P., Utete, S., King, D., King, S., Anuzis, P., Tarassenko, L.: Static and dynamic novelty detection methods for jet engine health monitoring. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **365**(1851), 493–514 (2007)
33. Hill, D.J., Minsker, B.S.: Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software* **25**(9), 1014–1022 (2010)
34. Hodge, V., Austin, J.: A survey of outlier detection methodologies. *Artificial intelligence review* **22**(2), 85–126 (2004)
35. Hubballi, N., Suryanarayanan, V.: False alarm minimization techniques in signature-based intrusion detection systems: A survey. *Computer Communications* **49**, 1–17 (2014)
36. Hundman, K., Constantinou, V., Laporte, C., Colwell, I., Soderstrom, T.: Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. *arXiv preprint arXiv:1802.04431* (2018)
37. Javaid, A., Niyaz, Q., Sun, W., Alam, M.: A deep learning approach for network intrusion detection system. In: Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS). pp. 21–26. ICST (Institute for Computer Sciences, Social-Informatics and . . . (2016)
38. Julisch, K.: Using root cause analysis to handle intrusion detection alarms. Ph.D. thesis, Universität Dortmund (2003)
39. Khashei, M., Bijari, M.: A novel hybridization of artificial neural networks and arima models for time series forecasting. *Applied Soft Computing* **11**(2), 2664–2675 (2011)
40. Kim, J., Scott, C.D.: Robust kernel density estimation. *Journal of Machine Learning Research* **13**(Sep), 2529–2565 (2012)
41. Kohonen, T.: The self-organizing map. *Proceedings of the IEEE* **78**(9), 1464–1480 (1990)
42. Kravchik, M., Shabtai, A.: Detecting cyber attacks in industrial control systems using convolutional neural networks. In: Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and Privacy. pp. 72–83. ACM (2018)
43. Kriegel, H.P., Kroger, P., Schubert, E., Zimek, A.: Interpreting and unifying outlier scores. In: Proceedings of the 2011 SIAM International Conference on Data Mining. pp. 13–24. SIAM (2011)
44. Kwon, D., Kim, H., Kim, J., Suh, S.C., Kim, I., Kim, K.J.: A survey of deep learning-based network anomaly detection. *Cluster Computing* pp. 1–13 (2017)
45. Lavin, A., Ahmad, S.: Evaluating real-time anomaly detection algorithms—the numenta anomaly benchmark. In: Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on. pp. 38–44. IEEE (2015)
46. Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., Srivastava, J.: A comparative study of anomaly detection schemes in network intrusion detection. In: Proceedings of the 2003 SIAM International Conference on Data Mining. pp. 25–36. SIAM (2003)
47. Lee, W., Xiang, D.: Information-theoretic measures for anomaly detection. In: Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on. pp. 130–143. IEEE (2001)
48. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. pp. 413–422. IEEE (2008)
49. Malhotra, P., Vig, L., Shroff, G., Agarwal, P.: Long short term memory networks for anomaly detection in time series. In: Proceedings. p. 89. Presses universitaires de Louvain (2015)

50. McCrae, R.R.: Creativity, divergent thinking, and openness to experience. *Journal of personality and social psychology* **52**(6), 1258 (1987)
51. Mohamed, A.A., Gavrilova, M.L., Yampolskiy, R.V.: Artificial face recognition using wavelet adaptive lbp with directional statistical features. In: *Cyberworlds (CW), 2012 International Conference on*. pp. 23–28. IEEE (2012)
52. Mohammadi, M., Al-Fuqaha, A., Sorour, S., Guizani, M.: Deep learning for iot big data and streaming analytics: A survey. *IEEE Communications Surveys & Tutorials* **20**(4), 2923–2960 (2018)
53. Nanduri, A., Sherry, L.: Anomaly detection in aircraft data using recurrent neural networks (rnn). In: *2016 Integrated Communications Navigation and Surveillance (ICNS)*. pp. 5C2–1. IEEE (2016)
54. Ning, P., Cui, Y., Reeves, D.S.: Analyzing intensive intrusion alerts via correlation. In: *International Workshop on Recent Advances in Intrusion Detection*. pp. 74–94. Springer (2002)
55. Om, H., Kundu, A.: A hybrid system for reducing the false alarm rate of anomaly intrusion detection system. In: *Recent Advances in Information Technology (RAIT), 2012 1st International Conference on*. pp. 131–136. IEEE (2012)
56. Page, E.S.: Continuous inspection schemes. *Biometrika* **41**(1/2), 100–115 (1954)
57. Pajurek, T.: Online anomaly detection in time-series (2018)
58. Patcha, A., Park, J.M.: An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer networks* **51**(12), 3448–3470 (2007)
59. Patra, B.K., Launonen, R., Ollikainen, V., Nandi, S.: A new similarity measure using bhattacharyya coefficient for collaborative filtering in sparse data. *Knowledge-Based Systems* **82**, 163–177 (2015)
60. Pietraszek, T., Tanner, A.: Data mining and machine learning—towards reducing false positives in intrusion detection. *Information security technical report* **10**(3), 169–183 (2005)
61. Poon, H., Domingos, P.: Sum-product networks: A new deep architecture. In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. pp. 689–690. IEEE (2011)
62. Portnoy, L.: Intrusion detection with unlabeled data using clustering. Ph.D. thesis, Columbia University (2000)
63. Poultney, C., Chopra, S., Cun, Y.L., et al.: Efficient learning of sparse representations with an energy-based model. In: *Advances in neural information processing systems*. pp. 1137–1144 (2007)
64. Preis, T., Kenett, D.Y., Stanley, H.E., Helbing, D., Ben-Jacob, E.: Quantifying the behavior of stock correlations under market stress. *Scientific reports* **2**, 752 (2012)
65. Qin, X., Lee, W.: Discovering novel attack strategies from infosec alerts. In: *Data Warehousing and Data Mining Techniques for Cyber Security*, pp. 109–157. Springer (2007)
66. Radon, A.N., Wang, K., Glässer, U., Wehn, H., Westwell-Roper, A.: Contextual verification for false alarm reduction in maritime anomaly detection. In: *Big Data (Big Data), 2015 IEEE International Conference on*. pp. 1123–1133. IEEE (2015)
67. dos Reis, D.M., Flach, P., Matwin, S., Batista, G.: Fast unsupervised online drift detection using incremental kolmogorov-smirnov test. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 1545–1554. ACM (2016)
68. Roschke, S., Cheng, F., Meinel, C.: A new alert correlation algorithm based on attack graph. In: *Computational intelligence in security for information systems*, pp. 58–67. Springer (2011)
69. Rousseeuw, P.J., Hubert, M.: Robust statistics for outlier detection. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1**(1), 73–79 (2011). <https://doi.org/10.1002/widm.2>, <https://doi.org/10.1002/widm.2>
70. Rousseeuw, P.J., Leroy, A.M.: Robust regression and outlier detection, vol. 589. John wiley & sons (2005)
71. Sadoddin, R., Ghorbani, A.A.: An incremental frequent structure mining framework for real-time alert correlation. *computers & security* **28**(3-4), 153–173 (2009)
72. Sakurada, M., Yairi, T.: Anomaly detection using autoencoders with nonlinear dimensionality reduction. In: *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. p. 4. ACM (2014)
73. Schervish, M.J.: P values: what they are and what they are not. *The American Statistician* **50**(3), 203–206 (1996)
74. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: *International Conference on Information Processing in Medical Imaging*. pp. 146–157. Springer (2017)
75. Shipmon, D.T., Gurevitch, J.M., Piselli, P.M., Edwards, S.T.: Time series anomaly detection; detection of anomalous drops with limited features and sparse examples in noisy highly periodic data. *arXiv preprint arXiv:1708.03665* (2017)
76. Siffer, A., Fouque, P.A., Termier, A., Largouet, C.: Anomaly detection in streams with extreme value theory. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 1067–1075. ACM (2017)
77. Sun, P., Chawla, S., Arunasalam, B.: Mining for outliers in sequential databases. In: *Proceedings of the 2006 SIAM International Conference on Data Mining*. pp. 94–105. SIAM (2006)
78. Tandon, G., Chan, P.K.: Tracking user mobility to detect suspicious behavior. In: *Proceedings of the 2009 SIAM International Conference on Data Mining*. pp. 871–882. SIAM (2009)
79. Todd, A.D., Raines, R.A., Baldwin, R.O., Mullins, B.E., Rogers, S.K.: Alert verification evasion through server response forging. In: *International Workshop on Recent Advances in Intrusion Detection*. pp. 256–275. Springer (2007)
80. V. Chandola, V.M., Kumar, V.: A comparative evaluation of anomaly detection techniques for sequence data. In: *IEEE Intl. Conf. on Data Mining (ICDM)*. p. 743–748. IEEE (2008)

81. Valdes, A., Skinner, K.: An approach to sensor correlation. In: Proceedings of RAID 2000 (2000)
82. Vallentin, M., Sommer, R., Lee, J., Leres, C., Paxson, V., Tierney, B.: The nids cluster: Scalable, stateful network intrusion detection on commodity hardware. In: International Workshop on Recent Advances in Intrusion Detection. pp. 107–126. Springer (2007)
83. Veasey, T.J., Dodson, S.J.: Anomaly detection in application performance monitoring data. In: Proceedings of International Conference on Machine Learning and Computing (ICMLC). pp. 120–126 (2014)
84. Vigna, G., Robertson, W., Kher, V., Kemmerer, R.A.: A stateful intrusion detection system for world-wide web servers. In: null. p. 34. IEEE (2003)
85. Vinayakumar, R., Soman, K., Poornachandran, P.: Applying convolutional neural network for network intrusion detection. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). pp. 1222–1228. IEEE (2017)
86. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* **11**(Dec), 3371–3408 (2010)
87. Wang, K., Stolfo, S.J.: Anomalous payload-based network intrusion detection. In: International Workshop on Recent Advances in Intrusion Detection. pp. 203–222. Springer (2004)
88. Weller-Fahy, D.J., Borghetti, B.J., Sodemann, A.A.: A survey of distance and similarity measures used within network intrusion anomaly detection. *IEEE Communications Surveys & Tutorials* **17**(1), 70–91 (2015)
89. Weston, J., Ratle, F., Mobahi, H., Collobert, R.: Deep learning via semi-supervised embedding. In: *Neural Networks: Tricks of the Trade*, pp. 639–655. Springer (2012)
90. Williams, A.W., Pertet, S.M., Narasimhan, P.: Tiresias: Black-box failure prediction in distributed systems. In: *Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International*. pp. 1–8. IEEE (2007)
91. Xiong, L., Chen, X., Schneider, J.: Direct robust matrix factorization for anomaly detection. In: *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. pp. 844–853. IEEE (2011)
92. Yen, T.F., Oprea, A., Onarlioglu, K., Leetham, T., Robertson, W., Juels, A., Kirda, E.: Beehive: Large-scale log analysis for detecting suspicious activity in enterprise networks. In: *Proceedings of the 29th Annual Computer Security Applications Conference*. pp. 199–208. ACM (2013)
93. Zimek, A., Schubert, E., Kriegel, H.P.: A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **5**(5), 363–387 (2012)
94. Žliobaitė, I.: Learning under concept drift: an overview. *arXiv preprint arXiv:1010.4784* (2010)
95. Zohrevand, Z., Glasser, U., Shahir, H.Y., Tayebi, M.A., Costanzo, R.: Hidden markov based anomaly detection for water supply systems. In: *Big Data (Big Data), 2016 IEEE International Conference on*. pp. 1551–1560. IEEE (2016)
96. Zohrevand, Z., Glässer, U., Tayebi, M.A., Shahir, H.Y., Shirmaleki, M., Shahir, A.Y.: Deep learning based forecasting of critical infrastructure data. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. pp. 1129–1138. ACM (2017)
97. Zuo, Y., Serfling, R.: General notions of statistical depth function. *Annals of statistics* pp. 461–482 (2000)