Anomaly detection in the context of Reinforcement Learning

MICHALSKI, PATRIK THOMAS, University Kiel - Department of Computer Science, Germany

Anomaly detection is applied in several critical areas, such as early disease detection in healthcare, fraud detection in finance, and intrusion detection in cybersecurity. Anomalies often have different causes, resulting in different classes of anomalies having distinctly heterogeneous characteristics. In addition, they also occur infrequently and unpredictably in datasets. It is, therefore, difficult to obtain training data that covers all possible classes of anomalies. Using machine learning these data can be transformed into insights that enable data-driven decisions. In particular, Reinforcement Learning methods have attracted significant interest due to their ability to learn complex behavior in the high-dimensional data space. Thereby, the anomaly detector makes no assumptions about the concept of anomalies. Instead, it identifies new anomalies by continuously expanding its knowledge through collected reward signals.

This paper attempts to show potentialities to overcome the difficulties of traditional anomaly detection methods with Reinforcement Learning. Th, possible methodologies are explained and substantiated with current research approaches (if available) that attempt to solve these problems. Thus, their requirements, strengths, and weaknesses are analyzed.

CCS Concepts: • **Computing methodologies** → **Reinforcement learning**; • **Information systems** → *Data mining*.

Additional Key Words and Phrases: Anomaly detection, Machine Learning and Reinforcement Learning.

1 INTRODUCTION AND MOTIVATION

Accelerated technological advances have increased unstructured data, contributing to rapid growth in data volumes. The amount of unlabeled data available is far more tremendous than practically processed [9]. No traditional technique can analyze and control these vast amounts of data without coping with an increase of several difficulties, e.g., a decrease in computational efficiency [6]. Therefore analytical and predictive tools are needed. Unsupervised can often detect various anomalies because labeled data do not constrain them. Hence such approaches have dominated this area for decades [6]. However, they can produce many false positives due to the lack of prior knowledge about true anomalies [5, 12]. Using machine learning techniques can transform this data into knowledge that enables data-driven decisions that increase the efficiency, robustness, and scalability of anomaly detection approaches [6, 11].

This paper is divided into three sections. The first section of the paper starts with the theoretical background and covers all the areas that should better understand. It should be noted that some basic knowledge about data collection, preprocessing, and machine learning, in general, is assumed. The second part then deals with the actual objective. The respective methodologies, algorithmic approaches, and their advantages and disadvantages are discussed. Finally, the last section then summarizes all these findings and presents possible further research ideas.

2 THEORETICAL BACKGROUND

This section of the paper explains the theoretical background and covers all the crucial areas for a better understanding of the topic. It should be noted that some basic knowledge about data acquisition, preprocessing, and machine learning, in general, is assumed.

2.1 Anomaly detection

Anomaly detection, also known as novelty or outlier detection, refers to detecting data instances that deviate significantly from most data instances. Due to increasing demand and applications in broad areas such as financial monitoring, health, and security, anomaly detection plays an increasingly important role in various fields [6]. Moreover, anomaly detection has been one of the critical research areas in data science for several decades due to its ubiquity [7]. Anomaly detection refers to the techniques for finding specific data points or patterns that do not fit the normal distribution of the dataset. It deals with rare events, minority, uncertainty, and unpredictability leading to unique problem complexities. In manufacturing, for example, it can be used to identify parts that are likely to fail. In security, it can be applied to detect potentially threatening users, and in social networks, it can identify people with unusual characteristics.

2.2 Reinforcement Learning

Reinforcement Learning (RL) is a fundamental paradigm of machine learning along with supervised and unsupervised learning [10]. It differs from supervised learning in that it does not learn from a training set of labeled examples provided by a knowledgeable external supervisor. RL also differs from unsupervised learning, which is typically about finding structures hidden in collections of unlabeled data. Instead, it is about how intelligent agents should perform actions to maximize cumulative rewards. RL problems are closed-loop problems since the actions of the learning system affect its subsequent inputs. The outcomes of each action, including reward signals that affect over time, are the main distinguishing features of RL. One of the challenges in RL is the tradeoff between exploration and exploitation [3]. To obtain different rewards, an RL agent must favor actions that it has tried in the past and have proven effective in generating rewards. However, to discover such actions, it must try actions that it has not previously selected. The agent must exploit what it already knows to obtain a reward, but it must also explore to make better future choices. So, generally speaking, RL problems are about learning what to do and mapping situations to actions. It involves capturing the most important aspects of the environment through interactions and then learning from them.

3 ANOMALY DETECTION IN THE CONTEXT OF REINFORCEMENT LEARNING

In this section, the requirements for selecting suitable RL-based algorithms for anomaly detection are discussed. Based on this, the difficulties of traditional anomaly detection methods are explained,

Author's address: Michalski, Patrik Thomas, stu207680@mail.uni-kiel.de, University Kiel - Department of Computer Science, Germany, Schleswig-Holstein, Christian-Albrechts-Platz 4, Kiel, 24118.

and possibilities that can be overcome with the help of RL are identified. Then, approaches from current research will be presented that attempt to solve these problems. Thus, their requirements, strengths, and weaknesses are analyzed.

The choice of an RL algorithm for an anomaly detection method depends primarily on the characteristics of the input data [7]. Input data can be categorized as sequential, e.g., music, speech, text, or non-sequential, e.g., graph, image, and table data (which will focus on this work). In addition, the input data can be divided into low- or high-dimensional data depending on the number of features. With the increase of data, two main problems arise. First, the performance of conventional algorithms in anomaly detection is suboptimal because they cannot capture complex structures as the data increases, and it becomes nearly impossible for conventional methods to scale to such large datasets to find anomalies [7]. Secondly, there are too few anomaly datasets with labels. For conventional data, many datasets are labeled, which can then be easily used for supervised learning [7]. This advantage cannot be exploited in unsupervised learning for anomaly detection since anomalies can occur in different variations and strongly depend on the underlying datasets.

3.1 Feature selection

Since the cost of data collection decreases, the space of features used to characterize a particular predictor of interest grows exponentially [8]. It becomes nearly impossible for conventional methods to scale to such large datasets to find anomalies [8]. Moreover, the performance of conventional algorithms are suboptimal because they cannot capture complex structures as the data increases [8]. Therefore, identifying the most characterizing features that minimize the variance without tampering with the models' bias is critical to successfully training a machine learning model [8]. In feature selection approaches, the purposes of the features are maintained while the feature space is optimally reduced according to a particular evaluation criterion. There are many potential benefits of using feature selection, including increasing accuracy, finding optimal computation costs, and overcoming the curse of dimensionality to improve predictive performance, and removing the irrelevant features according to a given criterion [4, 8].

Mehdi and Rasoul et al. [4, 8] propose two similar approaches how to improve feature selection for large datasets. Thus, they use a temporal difference algorithm where the state space comprises all possible subsets of the features. The action space for every state is defined by the number of features not already included in the model. An action represents a feature included in the model. The reward function determines the scored accuracy, which is used to evaluate the current set's predictive ability. Thereby, the reward is the difference in accuracy value for the current state and the next state after including an additional feature in the model. The average of all collected rewards for that feature in several iterations is considered as its final score. This difference between the values of two consecutive states shows the effectiveness of the corresponding feature that causes the transition allowing to (I) handle high dimensional features space and (II) being robust enough for any non-linear relationship between the predictors and the response feature [8].

3.2 Labeling with few labels

The second major problem that anomaly detection has to tackle is that there are too few anomaly datasets with labels [6]. For conventional data, many datasets are labeled, which can then be easily used for supervised learning, allowing for adequate and fast training on the one hand and subsequent modeling on the other. This advantage cannot be exploited in unsupervised learning for anomaly detection since anomalies can occur in different variations and strongly depend on the underlying datasets.

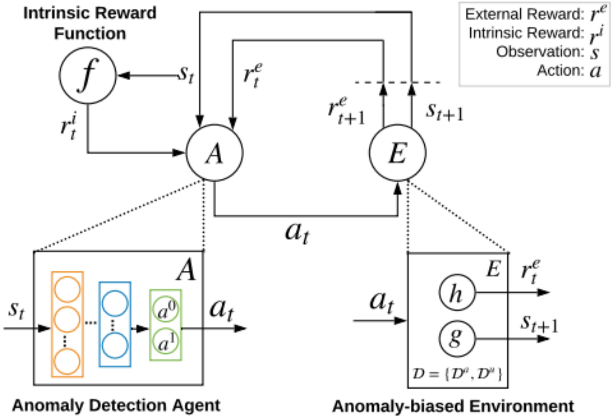

Fig. 1. Pang et al.'s anomaly detection-oriented deep RL approach [6].

The following approach by Pang et al. [6] addresses the problem described above by using learning knowledge recognition models from a small set of partially labeled anomalies and a large unlabeled dataset. Pang et al. propose a deep RL-based approach that actively searches for new classes of anomalies that lie outside the domain of the labeled training data. This approach learns to achieve a balance between exploiting the existing data model and searching for new classes of anomalies. It can then use the labeled anomaly data to improve detection accuracy without limiting anomalies searched to the given anomaly examples. Therefore, an anomalous environment is defined (denoted with E) by a mixture of an external reward function and an observation generator to train the agent (denoted with A). The deep RL method tries to learns and select an optimal action (denoted with a_t) from two possible actions. This action corresponds to labeling a given observation as normal or anomalous. At each time step, the agent receives an observation generated by the observation generator and performs an action. An intrinsic reward function (denoted with f) is defined to provide a second reward (denoted with r_t^i) based on the abnormality of observation to encourage unsupervised active exploration to detect possible unlabeled anomalies. The anomaly can be inferred from the agent's estimated value, i.e., the expectation of the future reward if a given action is performed during a given observation. Figure 1 illustrates the above-described procedure with a extension of the anomaly detection agent and the anomaly-biased environment for better understanding. Thereby, the agent is realized by a typical RL structure, e.g., Q-Learning.

3.2.1 Advantages and disadvantages. Analysis of the approach has shown that (p.I) it can use the limited labeled anomaly data and actively explore the heterogeneous and sparse anomaly data in the large unlabeled datasets. (p.II) It allows to detect significantly more anomalies than existing methods [6]. (p.III) This technique can support traditional unsupervised anomaly detection approaches by raising the founded/generated labels to the anomaly detection method to improve the model's accuracy. In doing so, (p.IV) it achieves a 23% - 98% increase in relative AUC-PR by only increasing the number of training steps. Moreover, (p.V) it outperforms all competing methods, e.g., state-of-the-art deep semi-supervised detectors DevNet, DevNet^{+}, and Deep SAD by 1% - 10% in AUC-ROC [6]. (p.VI) Increasing the number of known anomaly classes provides more monitoring information to achieve significant improvement, especially for datasets where the first known anomaly class cannot provide much generalizable information. The main problem of this approach is that (c.I) the overall performance becomes worse when the dataset has too low dimensionality since the model is not expressive enough to capture complex relationships in most datasets [6]. In addition, (c.II) the time complexity of the training is slower than for the competing methods due to the high degree of parameters.

3.3 Hyperparameter optimization

The increasing complexity and amount of datasets also increased training time by requiring more resources, making hyperparameter optimization a general problem in machine learning and playing a significant role, limiting potentially the underlying anomaly detection approaches as earlier described [2]. It is an integral aspect of achieving the best performance for each model. It decides whether a trained model turns out to be state-of-the-art or simply moderate [2]. Thus, hyperparameter tuning must keep a low computational budget performance, be robust and scalable. Hyperparameters are often optimized by training a model on a grid of possible hyperparameter values. The set of parameters is then taken that performs best on a validation sample.

Jomaa et al. [2] propose a hyperparameter tuning method. Their agent learns to explore the hyperparameter space of fixed network topology. The agent starts at a random location in the hyperparameter space of the dataset and navigates on the surface of a given model type. Thereby, the agent explores the environment by selecting the following best hyperparameter configuration and receives a reward with every step until a terminal state is reached. The observed reward depends only on the dataset and the selected hyperparameter configuration. Once an action is selected, a new hyperparameter configuration is evaluated.

The proposed approach (I) does not suffer from the cubic dimensionality problem like Gaussian-based approaches [2]. Jomaa et al. show that their method (II) outperforms the state-of-the-art approaches, especially with a smaller budget, where the average distance to the minimum is small from the first selection [2]. Moreover, (III) the agent balances exploration and exploitation by enforcing termination when an episode is repeated.

3.4 False positives

In practice, analysts typically examine the top instances in a ranked list of anomalies identified by an anomaly detection system to identify true anomalies [12]. This analysis process generates labels that can be used to re-rank the anomalies to discover more true anomalies and reduce false positives. Existing strategies have focused on making the top instances more likely to be anomalous based on feedback. As a result, they then greedily select the top instance to query. However, these greedy strategies can be suboptimal, as some low-ranked instances are maybe more helpful in the long run [12].

Zha et al. [12] propose Active Anomaly Detection with Meta-Policy (Meta-AAD), a novel strategy that learns a meta-policy for query selection to solve the problem of false positives. Figure 2 illustrates the above-described procedure. Meta-AAD uses deep RL to train with meta-policy to select the most appropriate instance to optimize the number of anomalies detected during the query process (step 01). In each iteration, the meta-policy chooses one of the instances (step 02) and queries an analyst, i.e., human (step 03a and b), to optimize its knowledge base (step 04a and b). It uses three types of information to decide which instance to query (step 03/04) by starting with the anomaly detector's calculated anomaly scores. These provide information about which instances are further from the majority to help the meta-policy identify more anomalous instances. Using the already labeled anomalous instances can be helpful to identify more anomalous instances by promoting those similar to these known anomalous instances, which improves performance. In addition, using non-anomaly labeled instances is also helpful. Similar instances can then be rejected, which reduces the number of false positives. Finally, it outputs after several training phases possible anomalies (step 05).

3.4.1 Advantages and disadvantages. Zha et al. show that (p.I) Meta-AAD outperforms state-of-the-art ranking strategies [12]. Moreover, the analysis shows that (p.II) the trained meta-policy is intrinsic and transferable, achieving a balance between short-term and long-term rewards. (p.III) It achieves more than 25% improvement in letters and speech compared to the best alternatives [12]. This approach also (p.IV) converges fast, making training the meta-policy computationally efficient and easy to apply. (p.V) A strong meta-policy can be trained even with small datasets since the features are transferable and the proposed training strategy is effective. (c.I) Meta-AAD incorporates human feedback into anomaly detection, which reduces the degree of automation and increases the overall effort. Moreover, (c.II) rewards depend heavily on the dataset. (c.III) Too large negative rewards can lead to a decrease in performance.

3.5 Causal Reinforcement Learning

RL is concerned with maximizing cumulative reward over a period, while causal inference provides techniques to combine structural information about the data generation process and the data itself to make derivations and inferences up to a counterfactual nature. The main difference between causal inference and inference of association is that causal inference analyzes the response of an effect variable when a cause of the effect variable is changed. Adding causal structural information to sample-efficient RL techniques can (I) improve accuracy, learning performance, and optimality [1].

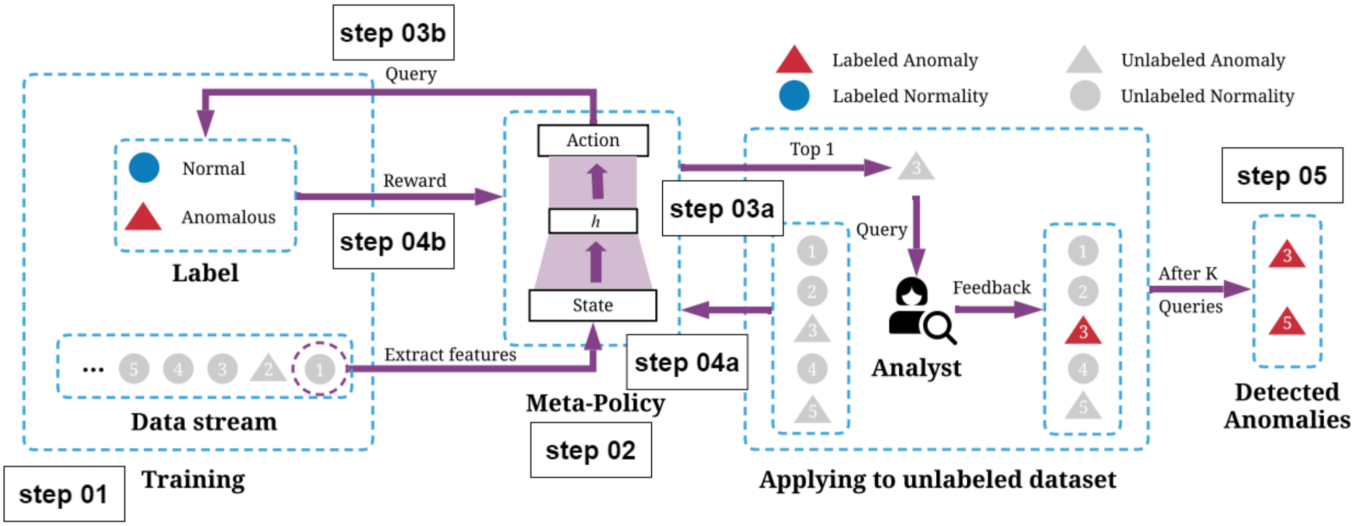

Fig. 2. Zha et al.'s active anomaly detection with deep RL (Meta-AAD) approach [12].

The idea behind causality-based off-policy policy evaluation is that the target policy is treated as a kind of intervention comprising counterfactual actions different from those in the behavioral policy. Therefore, to evaluate the target policy based on observational data generated by the behavioral policy, the focus lies on the difference between the two policies' outcomes. Dealing with off-policy evaluation is nothing more than predicting the decision-making system under counterfactual interventions. From this perspective, causality enhances previous off-policy evaluation methods by transferring the cause-effect relationship from observations to the target policy evaluation process. Causality makes it possible that estimators that do not rest on MDP models can still generalize by dealing with past experiences, which are leveraged to make predictions under interventions. The difficulty of anomaly detection in the context of RL is decision-making in unknown environments without analysis because poor estimation or limited knowledge of the environment can not be further improved by trial and error. Causal inference (II) gives new insights about this problem, equipping the autonomous agent to behave accordingly, with better generalization ability and enabling the agent to estimate the counterfactual outcomes. Furthermore, built on the similarity and close connection between causality and RL, further connections can be made between both approaches for data exploration, error estimation, and lower-bound limits. By using causal inference, (III) new possibilities can be exploited in the context of traditional approaches [1]. One example is the hyperparameter optimization mentioned earlier. If one knew how specific parameters behave when they are changed, one could target the entire search space and reduce it accordingly. It would also reduce the overall run time. Another way to use causal inference would be in the context of detecting anomalies during the decision process. For example, one could look at how the addition or omission of a particular data set affects the entirety of the data set, thus affecting the incoming accuracy rate.

4 RESULTS, CONCLUSION AND DISCUSSION

This section summarizes all presented results from the previous sections and presents some further possible research ideas.

4.1 Results

Many fields of machine learning are still in a state of significant change. New approaches are constantly being published that introduce improvements in specific domains. The amount of unlabeled data available is far more tremendous than practically processed through the rapid growth of unstructured data [9]. No traditional technique can analyze and control these vast amounts of data without coping with an increase of several difficulties, e.g., a decrease in computational efficiency [6]. Therefore analytical and predictive tools are needed. Using machine learning techniques can transform this data into knowledge that enables data-driven decisions that increase the efficiency, robustness, and scalability of anomaly detection approaches [6, 11].

4.2 Conclusion

This paper attempted to show potentialities to overcome the difficulties of traditional anomaly detection methods with Reinforcement Learning. Possible methodologies are explained and substantiated with current research approaches (if available) that attempt to solve these problems. Hence, their requirements, strengths, and weaknesses are analyzed.

In conclusion, there is no perfect approach that can solve all problems at once. It always depends on the current problem focus. If one wants to train a supervised anomaly detection approach with few known labels, the approach of Pang et al. [6] would be the most recommended since it can learn much more knowledge with a small number of labels by training the agent. Any anomaly detection algorithm is only as good as the available data set. Therefore, it is essential to have as much good, i.e., meaningful, data available as possible to achieve the best possible coverage. This type of selection

	Feature selection	Labeling with few labels	Hyperparameter optimization	False positives	Causal Reinforcement Learning
further readings	Mehdi and Rasoul et al. [4, 8]	Pang et al. [6]	Jomaa et al. [2]	Zha et al. [12]	Gua et al. [1]
advantages	- handles high dimensional features space - robust enough for any non-linear relationship	- works on heterogeneous and sparse anomaly data - detects significantly more anomalies than existing methods - 23% - 98% increase in relative AUC-PR - outperforms the state-of-the-art approaches - increasing the number of known anomaly classes achieve significant improvement	- does not suffer from the cubic dimensionality problem - outperforms the state-of-the-art approaches - balances exploration and exploitation	- outperforms the state-of-the-art ranking strategies - meta-policy is intrinsic and transferable, achieving a balance between short-term and long-term rewards - achieves more than 25% improvement - converges fast - works even with small datasets	- improve accuracy, learning performance, and optimality - gives new insights about this problem - new possibilities can be exploited in the context of traditional approaches
disadvantages	none	- performance becomes worse when the dataset has too low dimensionality - time complexity of the training is slow	none	- incorporates human feedback - rewards depend heavily on the dataset - large negative rewards can lead to a decrease in performance	none

Table 1. Summary of all explained approaches.

can be performed using the approaches presented by Mehdi and Rasoul et al. [4, 8]. It is possible to use a high-dimensional feature space through these approaches while being robust enough to a nonlinear relationship between the predictors and the response feature [4, 8]. Another problem in machine learning is the increasing amount and complexity of datasets, which thus increases the training time [2]. It makes hyperparameter optimization a common problem in machine learning and plays an important role, potentially limiting the underlying approaches to anomaly detection as described previously [2]. In order to use as few parameters as possible and then initialize them as well as possible, one can use the approach of Jomaa et al. [2]. The agent learns to explore a fixed network topology's hyperparameter space by selecting the best hyperparameter configuration. Finally, arguably one of the most critical issues is the reduction of false positives, as this reduces the accuracy of the entire system, and potentially correct anomalies could be missed. One approach would be Zha et al.'s [12] Meta-AAD. They use Deep RL to train a meta-policy that selects the most appropriate instance to optimize the number of anomalies detected during the query process by querying an analyst, i.e., a human. A completely different direction would be to use a causal RL approach. One could use causal inference to combine structural information about the data generation process and the data itself to draw inferences and conclusions up to counterfactual nature. This technique provides new insights into the problem and gives the autonomous agent a better generalization capability. The agent can then behave appropriately, estimating counterfactual outcomes. Furthermore, building on the close connection between causality and RL, further connections can be made between casual, and RL approaches to data exploration, error estimation, and lower bounds. Combining this method with Zha et al.'s [12] Meta-AAD approach increases the accuracy and correctness of results by reducing the probability of false positives. Table 1 summarizes possible methodologies that are explained and substantiated with current research approaches (if available) that attempt to solve the mentioned problems. The table also lists additional literature sources intended to serve for further in-depth study of the respective areas.

4.3 Discussion

Many RL approaches mentioned in the last section, and other publications have some problems in common. They are very computationally intensive and require much time for training and updating the model [6]. This problem results from the curse of dimensionality of the input data. However, using dimensionality reduction often adds several hyperparameters that are critical for performance [2]. Moreover, most attempts require a static environment and are only usable beforehand to perform anomaly detection. Therefore a possible approach could exploit other aspects of machine learning to adjust these hyperparameters during runtime to make the best possible choice on demand available. In particular, another RL agent could aim to find a suitable parameter set using an appropriate reward function. Using feature selection to find the best features can reduce several independent but highly correlated hyperparameters to improve the actual anomaly detection algorithms. This multi-agent approach would combine the best parts of each algorithm to create a highly adaptable and good-performing anomaly detection approach.

In conclusion, machine learning, especially RL, is a promising research field. Applying it to the domain of anomaly detection will solve many of today's known problems as it already partly does through considering and combining more approaches from different fields of machine learning. With already known methodologies from the traditional anomaly detection domain, new state-of-the-art methods handle and even partially solve the initial problems.

REFERENCES

[1] Ruocheng Guo, Lu Cheng, Jundong Li, Richard P. Hahn, and Huan Liu. 2020. A Survey of Learning Causality with Data. *Comput. Surveys* 53, 4 (2020), 1 – 37. https://doi.org/10.1145/3397269

[2] Jomaa S. Hadi, Josif Grabocka, and Lars Schmidt-Thieme. 2019. Hyp-RL : Hyperparameter Optimization by Reinforcement Learning.

[3] Marlos C. Machado, Sriram Srinivasan, and Michael Bowling. 2014. Domain-Independent Optimistic Initialization for Reinforcement Learning.

[4] Seyed Mehdi Hazrati Fard, Ali Hamzeh, and Sattar Hashemi. 2013. Using reinforcement learning to find an optimal set of features. *Computers & Mathematics with Applications* 66, 10 (2013), 1892 – 1904. https://doi.org/10.1016/j.camwa.2013.06.031

[5] Min-hwan Oh and Garud Iyengar. 2019. Sequential Anomaly Detection Using Inverse Reinforcement Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery (ACM), New York, NY, USA, 1480 – 1490. https://doi.org/10.1145/3292500.3330932

[6] Guansong Pang, Anton Van Den Hengel, Chunhua Shen, and Longbing Cao. 2020. Deep Reinforcement Learning for Unknown Anomaly Detection.

[7] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep Learning for Anomaly Detection. *Comput. Surveys* 54, 2 (2021), 1 – 38. https://doi.org/10.1145/3439950

[8] Sali Rasoul, Sodiq Adewole, and Alphonse Akakpo. 2021. Feature Selection Using Reinforcement Learning.

[9] statista.com. 2021. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025. https://www.statista.com/statistics/871513/worldwide-data-created/. accessed: 2021-07-21.

[10] Richard S. Sutton and Andrew G. Barto. 2014. Reinforcement Learning: An Introduction Second Edition: 2014, 2015.

[11] Tong Wu and Jorge Ortiz. 2021. RLAD: Time Series Anomaly Detection through Reinforcement Learning and Active Learning.

[12] Daochen Zha, Kwei-Herng Lai, Mingyang Wan, and Xia Hu. 2020. Meta-AAD: Active Anomaly Detection with Deep Reinforcement Learning.