Explainable Contextual Anomaly Detection: a focus on QCAD

Gauderic Gumbs, Maurras Togbe, Yousra Chabchoub, Patrick Perrot

Explainable Contextual Anomaly Detection: a focus on QCAD

Gauderic Gumbs[1,2], Maurras Togbe[1], Yousra Chabchoub[1], and Patrick Perrot[2]

[1] ISEP -Institut Supérieur d'Electronique de Paris. 10 rue de Vanves, Issy-les-Moulineaux, 92130-France.
firstname.lastname@isep.fr,
[2] Commandement du ministère de l'Intérieur dans le cyberespace, Paris
firstname.lastname@gendarmerie.interieur.gouv.fr

Abstract. The integration of explainability techniques in anomaly detection systems is crucial for improving transparency and trust in decision-making processes across various application domains. Indeed, in a world increasingly reliant on data and algorithms, it is imperative for users to understand how and why a decision was made by an automated system. This is particularly true in sensitive sectors such as cybersecurity as well as finance, healthcare, where misinterpretation of results could have serious consequences. Traditional anomaly detection methods often lack explanations, making it difficult for users to understand the decisions made by these systems. To address this issue, the explainability of models is becoming a priority. In this paper, we conduct a detailed review of methods designed to enhance the explainability of anomaly detection models based on machine learning and artificial intelligence. We particularly study the various comparison metrics for all these methods: the positioning of explainability in relation to the model, the genericity of the explainability, the type of model, etc. We then focus on a recent explainable contextual anomaly detection method using quantile regression forest: QCAD developed by Li et al. in 2023. This method uses explainability while taking into account the context, which is crucial for a more efficient anomalies detection in many fields where the context has an important impact on the normal behavior. Li et al. showed that QCAD often yields excellent results in terms of PRC-AUC. We tested this method on a real-world dataset called "Bodyfat" to evaluate the impact of context size on its performance. Moreover, we exploited the explainability layer of QCAD to investigate the reasons behind its mitigated results on the Bodyfat dataset.

Keywords: artificial intelligence, anomaly detection, contextual anomaly detection, explainablity

1 Introduction

Anomalies are patterns whose behavior varies significantly from the most of the data [7]. In the field of anomaly detection, it is crucial to establish the expected behavior, known as normal behavior. Usually, the normal behavior is modeled according to the majority of the data. Once this profile is established, new incoming data are compared against this model. Data items are evaluated to determine how closely they match the expected characteristics. Any data item that deviates significantly from this expected behavior is then flagged as an anomaly. These anomalies can be identified using a variety of statistical techniques [7], classical machine learning algorithms [14] and deep learning algorithms [17].

Thus, anomaly detection has been widely studied and there is a plethora of comprehensive reviews and articles in different application domains such as cybersecurity [18] or finance [10]. However, less attention has been devoted to the explainability of anomaly detection methods. Explainability of models gives the decision makers a more complete view of the obtained results and enables them to better justify their decision. It is one of the key issues for the adoption of algorithms in society and is becoming an ethical and regulatory requirement in safety-critical areas.

Moreover, the majority of existing research studies on anomalies detection, largely neglect the detection of contextual anomalies. Contextual anomaly detection is particularly relevant in domains where the context has a notable impact on the behavior. In such a situation, the normal behavior is not universal but closely related to the considered context. Therefore, several normal behaviors should be defined in this case and the anomaly, in a given context, is defined as a large deviation from the normal behavior of the addressed context. In many fields the context refers to space-time features, but it can also for example concern the age and sex of patients for the prediction of certain diseases [16]. Hence data features can be classified into two categories: contextual and behavioral, with the help of experts or using automatic classification methods. Taking into account the contextual characteristics enables a more relevant anomalies detection. The paper is structured as follows, first we highlight the state of the art of 3 contextual anomaly detection methods, then we address the notion of explicability and its importance, following that we experiment on one of the 3 methods, and finally we end with a conclusion and future works.

2 Contextual Anomaly Detection Methods

Where traditional anomaly detection methods treat each feature uniformly, contextual approaches categorize attributes into two groups: contextual (or environmental) attributes and behavioural (or indicator) attributes. Contextual attributes define the environment or the conditions under which an object exists, whereas behavioural attributes are used to measure the normality of that object. For example, a heart rate of 100 bpm is abnormal for an adult at rest, but normal for a child or adult in physical activity [5]. To emphasize the importance

of separating and adequately modelling the relationships between contextual and behavioural attributes for effective anomaly detection, three main methods for contextual anomaly detection, are presented: CAD, QCAD and ConQuest. These methods provide different ways of separating contextual and behavioural attributes to take correct action in the relevant context for accurate anomaly detection

2.1 Conditional Anomaly Detection (CAD)

CAD [26] is developed by in Song, Xiuyao, et al. in 2007. Like most of the contextual anomaly detection methods, CAD separates attributes into environmental attributes, which shape the context of the data without directly signaling anomalies, and indicator attributes, which are useful to identify anomalies. CAD, firstly, utilizes a statistical model (Gaussian Mixture Model [21], GMM), in most cases for capturing the trends and relationships present in historical data; these enable greater accuracy in the identification of true anomalies by modeling the inter-relationship between environmental and indicator attributes in order to reduce false positives. Subsequently, it calibrates those parameter values by adjusting its model parameters through the use of Expectation-Maximization (EM) algorithms to learn the relationships among environmental and indicator attributes. The Conditional Probability Density Function (Conditional PDF) is a major part of CAD, which calculates the distribution of indicator attributes based on values of environmental attributes. In this way, CAD detects new observations where the values of the indicator attributes, when conditioned by the values of the environmental attributes, deviate significantly from the expected mean values. This technique of context-related anomaly detection makes the identification of the anomalies more robust and useful. Subsequently, a probability threshold is used to identify whether an observed point is in fact an anomaly. Learning in CAD as the same principal as other contextual anomaly models (detect anomalies by identifying data points that deviate from expected patterns) and can be expressed using different algorithms, like Direct-CAD algorithm, GMM-CAD-Full algorithm, and GMM-CAD-Split algorithm, all adopting EM for parameter optimization of the model to properly represent the dependencies in the data.

2.2 Context discovery for anomaly detection (ConQuest)

ConQuest [5] developed by Calikus et al in 2024 is an innovative approach to contextual anomaly detection, characterized by the ability to automatically discover and integrate multiple relevant contexts for anomaly detection and interpretation. The process begins with the extraction of context-behaviour pairs from a database using a sliding window; this allows data to be analyzed in continuous segments in order to capture local variations. Reference groups are defined based

on the similarity of contexts. ConQuest formulates objective functions. These are mathematical criteria that evaluate and steer the optimization of solutions. Three goals can be identified: maximizing dependency of context and behaviour (objective function 1), contextual redundancy minimized (objective function 2), and anomaly discrimination maximized (objective function 3). Multi-objective optimization is based on a genetic algorithm called NSGA-II. This methodology has a scope to explore the context space. The algorithm mentioned produces a Pareto front of non-dominat ed solutions [20]. A solution is said to be Pareto optimal when no further improvement is possible in any objective function without degrading at least one of the other objective functions [6]. The choice of the top solution with m contexts from among the Pareto front is done by ConQuest using a selection method named TOPSIS, which stands for Technique for Order Preference by Similarity to Ideal Solution. A simple yet effective method that chooses the best alternative based on the shortest and farthest Euclidean distances from the positive ideal solution (PIS) and negative ideal solution (NIS). Lastly the algorithm MCAF (Multi-Context Anomaly Factor) tries to combine the contextual and behavioural distances to produce an anomaly score by aggregating scores from multiple contexts. It also has the capability to explain and visualize the anomalies so that they can be compared with their reference groups.

2.3 Quantile Contextual Anomaly Detection (QCAD)

QCAD[12] developed by Li, Zhong, and Matthijs Van Leeuwen in 2023 uses quantile regression forests to model relationships between features. Here, reference groups are constructed with two dimensions distance matrix calculated using the Gower's distance. There is a reference group for each observation, containing the set of the k nearest neighbors of the considered observation. A learning process then occurs by constructing quantile regression forests (QRF) for each reference group based on their behavioral attributes. Anomalies are then identified based on the entire probability density function and the local density of each observation in each QRF.

Anomaly scores take into account the width of the quantile interval. The key idea is that the wider the quantile interval, the lower the density around a considered value, which is very likely to indicate an anomaly. The scores are then aggregated over all the behavioural features to obtain the overall anomaly score for a data item. More precisely, a Min-Max normalization is applied on all behavioral features for a better performance.

Once an anomaly is detected, QCAD applies an explainability layer to provide more information and details about the anomaly. It consists of decomposing the overall anomaly score into the contributions of the individual features. This enables the identification of the features that mainly contribute into the anomaly. For this purpose, QCAD provides visual explanations with the use of beanplots that illustrate how an observation's behavioral features deviate from

its reference group. The visualization of the conditional distributions helps understanding why a specific data item is identified as anomalous, ensuring a better transparency and interpretability. A complete architecture of QCAD is provided in Figure 1.

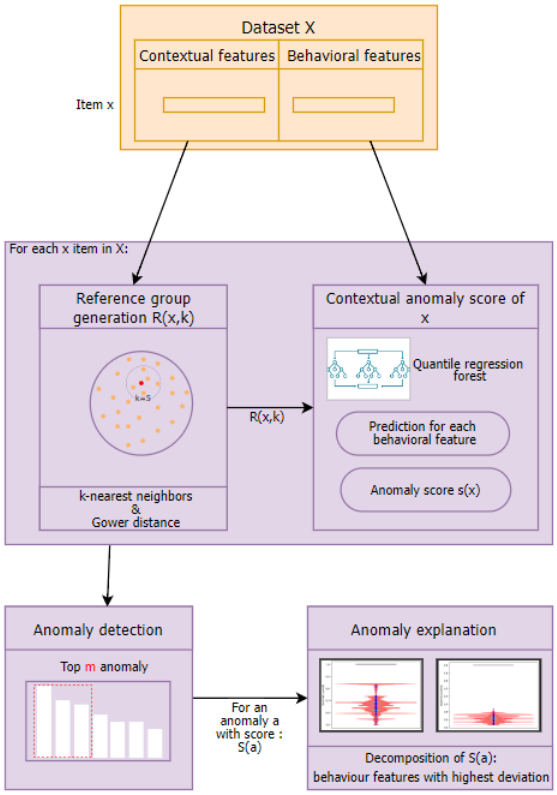

Fig. 1. QCAD architecture

3 Explicability in AI: application in cybersecurity

Explainable Artificial Intelligence (XAI) is an approach in artificial intelligence that aims to improve human understanding of the decisions made by AI models. The use of AI, particularly machine learning and deep learning, enhances detection and defense against cyberthreats, outperforming traditional signature and rule-based methods. These AI-based techniques are frequently applied in an opaque "black box" [2] way, in which both security experts and end users are hard-pressed to understand and rationalize the decisions made by these models.

This often results in a lack of transparency and interpretability, leading to a reduction in user confidence with respect to AI-based cyber defense models, especially with the emergence of more and more complex attacks. Hence, there is an XAI critical need to apply in developing more explicable cyber security models that allow users to understand, trust, and manage next-generation defense mechanisms. This section investigates the various existing XAI techniques.

3.1 Motivation

XAI makes AI model-based decisions less opaque, which increases confidence in the system by end users. With regulations tightening, such as the GDPR in Europe, which mandates reasons for decisions taken by automated systems, XAI becomes essential. Explainable models can hence guide organizations to be in compliance with these rules and provide clear and understandable justification for every decision taken by an AI system. Such regulatory compliance is critical to protect an organization from legal sanctions and maintain a good reputation.

Explainability enables cyber security experts to identify and correct systems' errors or biases for more effectiveness. In fact, an idea of why a model has wrongly classified an attack may help in the optimization of the algorithms to improve the detection process. XAI offers fine-grained explanation capabilities that guarantee continuous optimization of cybersecurity systems to ensure a better protection from threats. As an example, adversarial attacks are a vulnerability in AI models, being designed for fooling the model through mischievous input. XAI can help point out such an attack through an explanation revealing an anomaly in the model's decision, hence reacting in time and appropriately.

In this sense, many cybersecurity professionals need to have insight into decisions made by AI systems to carry out thorough investigations and incident responses. XAI provides the ability to understand the decisions taken by models, hence making such models interpretable by security professionals in their tasks. This is important for effective post-incident analysis as well as building stronger defense strategies.

3.2 XAI methods

SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are among the first designed explainability algorithms. Their emergence respectively in 2017 and 2016 marked a major turning point in the field of XAI, especially for complex, black-box models. SHAP[13] uses "Shapley values" with an approach based on game theory. It calculates the importance of each feature as an additive contribution to the prediction of a specific instance.

LIME[22], developed by Ribeiro et al., is a "model agnostic" method that gives an explanation of the behavior of a model for a given prediction or for a subset of predictions with the SP-LIME method. The principle of this method is called "Kernel based": to give an explanation of a model, the LIME creates an explanatory model based on the original model.

Many other explainability algorithms and libraries have been recently developed[28]. They are based on different concepts and approaches [27] [11] and generate different output formats (numerical, textual, visual). Among the commonly used explainability algorithms, we can cite the Activation Maps known as heatmaps which are based on visualization and particularly used for image processing. Many extensions of Activation Maps have been developed like Class Activation Map (CAM) and Grad-CAM. Decision Trees (DT) are also used to explain some particular models. We propose in Tables 1, 2 and 3 a complete classification of the most used explainability algorithms. We considered three criteria : 1) Using all observations to provide an explainability or not, 2) Dependency between the explainability and the ML model[1], and 3) Positioning of the explainability: it begins when building the model or after. We define in these tables each type of algorithms and give the most pertinent examples of algorithms for each category.

Explanation type	**Description**	**Example**
Local	Focuses on explaining a specific decision or instance within a model's inference, providing explanations limited to individual cases.	Activation Maps [29], CAM [30], Combinatorial Methods [15]
Global	Aims to make the entire inferential process of a model transparent and comprehensible as a whole, explaining overall observations across the dataset.	BCM [4]
Hybrid Approach	Combines both local and global explanations, providing both specific instance-level explanations and insights into the overall behavior of the model.	Activation maximization [9], Cell Activation Values [30], Grad-CAM [24], LIME [22]

Table 1: Global/local explanation with respect to the observations

Explanation type	Description	Example
Model-agnostic	Techniques independent of the type of machine learning model used, capable of being applied across different learning approaches.	Ada-WHIPS, CAM [30], LIME [22]
Model-specific	Applicable only to particular kinds of models, such as interpreting the weights or activation values specific to neural network models.	DT [19], Rule lists [23]
Hybrid Approach	Utilizes methods that cater to specific model types while also being adaptable to various model architectures.	Grad-CAM[24], SHAP[13]

Table 2: Dependency of the explanation on the model

Explanation type	Description	Example
Intrinsic	Achieves interpretability by building inherently interpretable models with simplicity and transparency embedded from the model's creation and training phase.	Activation Values of Hidden Neurons [29], Ant Colony Optimization (ACO) [8]
Post-hoc	Explains or mimics the behavior of a trained model using external explainers during testing or after model training without altering the model itself.	Attention Alignment [3], Average Activation Values [25]

Table 3: Positioning of the explanation

4 Experimental study of QCAD

We focus in this section on QCAD as the authors showed in [12] that QCAD often outperforms many other well-known anomalies detection algorithms: LoPAD, ROCO, CAD, IForest, LOF, K-NN, SOD and HBOS. They compared the performance of these algorithms over 20 real-world datasets. More precisely, they evaluated the PRC AUC (Precision-Recall Curve Area Under the Curve) metric, which measures the balance between precision and recall. This metric is particularly useful for imbalanced datasets and so it is well adapted for anomalies detection.

QCAD gives an excellent PRC AUC for all the datasets except a particular dataset called "Bodyfat" where the PRC AUC equals only 0.6. Bodyfat contains 252 samples with 13 contextual and 2 behavioral features. Our objective in this section is to investigate the reasons behind the relatively low PRC AUC. We also discuss in the context size which is an important input parameter of QCAD.

4.1 Impact of the context size

One of the key parameters that directly influences QCAD performance is the number k of nearest neighbors used to generate the reference group. For this purpose, we considered several values of k and we evaluated the PRC AUC based on Bodyfat dataset. Figure 2 shows that when k increases, QCAD is more efficient. However, PRC AUC becomes constant beyond a given threshold. It means that the gain become negligible while the computational cost continues to rise, hence the importance of limiting the value of k for an optimal balance. In [12], the authors recommend to set k using the following formula $k = min(N/2, 500)$ where N is the size of the dataset.

This means that the context size is a constant parameter and is common for all dataset items. We believe that the context size can be different from an item to another. It depends on the distribution of the contextual behavioral features in the dataset. A preliminary step should be added to QCAD to identify the different contexts using an unsupervised clustering algorithm. This enables to detects the natural contexts instead of identifying the k nearest neighbors for each data item. In fact, using k nearest neighbors implies a strong assumption of equal context sizes which is not the case of most of real-world datasets.

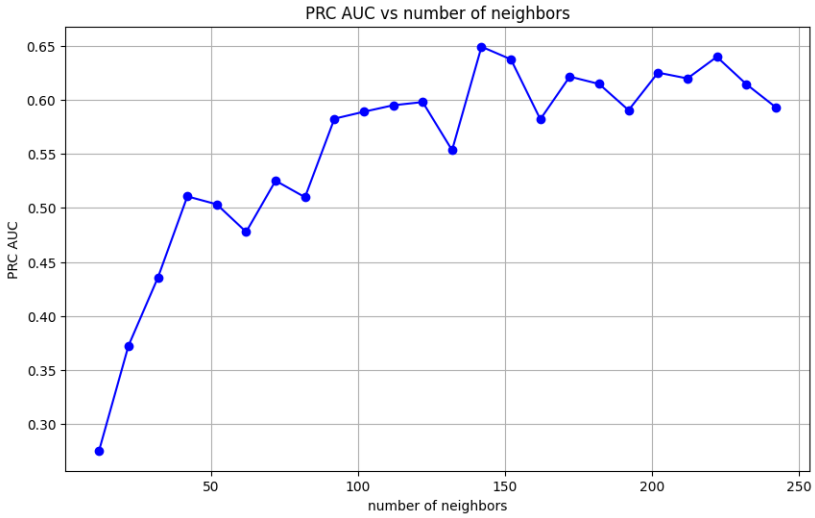

Fig. 2. impact of the number of neighbors

4.2 Impact of the dataset distribution, and anomalies injection

The 20 real-world datasets considered in [12] do not naturally contain anomalies. So, the authors injected, as an example in Bodyfat dataset 20 anomalies in the preprocessing step. Then they applied the quantile random forest to identify top 20 anomaly scores. They use the same method to inject anomalies in all the considered datasets.

We conducted the same experiment on Bodyfat and we obtained a PRC AUC of 0.6. The confusion matrix is given in Table 4 for more precise details. Positive refers to anomalies.

Within the top 20 anomaly scores, we observe that 9 are False Positives.

This means that, in the original dataset, 9 of the 232 data items that were not injected received higher anomaly scores than the 9 injected ones. This suggests that within their context, these items have behavioral features more indicative of anomalies than the injected ones. The beanplots given in Figures 3, 4 and 5 further clarify this by showing the position of the considered item compared to the estimated conditional distribution of each behavioral feature. We recall that Bodyfat dataset contains 2 behavioral features. The red area represents the occurrence probability. The horizontal black lines indicate the values reported for the data item feature under investigation.

Figure 3 presents an injected anomaly correctly labeled as anomalous, while Figure 4 shows an injected anomaly misclassified as normal. Figure 5 displays an original, non-injected data point that was labeled as anomalous. The global anomaly score of the item considered in Figure 5 is higher than the score of the item of figure 4. That is why this latter is not part of the top 20 detected anomalies. This is also the case for the other 8 missed injected anomalies. They are not enough anomalous to be ranked in the top 20 scores.

These observations show that anomalies injection should not be performed in the same manner for all the datasets. It must take into account the dataset specificity and anomalies already present in the dataset before injecting new anomalies. As a conclusion, the low PRC AUC of QCAD obtained for the Bodyfat dataset is due to in-adapted injection of anomalies and not to the QCAD algorithm.

		Reality	
		Positive	Negative
Prediction	Positive	11	9
	Negative	9	223

Table 4. Confusion Matrix, 20 anomalies injected in Bodyfat dataset

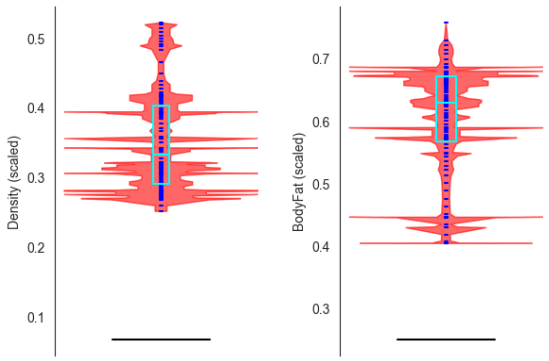

Fig. 3. Injected anomaly labeled as anomalous

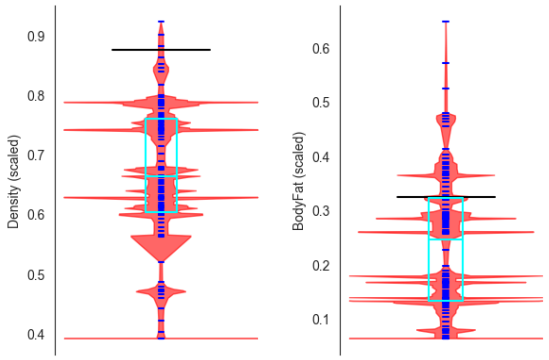

Fig. 4. Injected anomaly labeled as normal

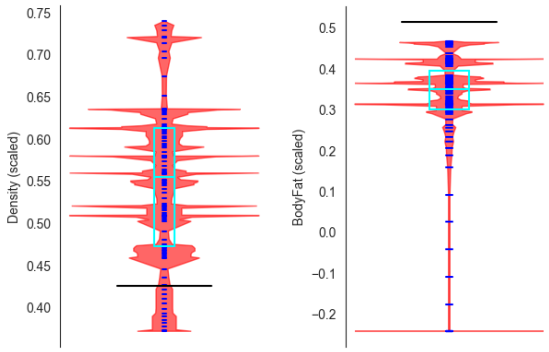

Fig. 5. Non-injected data labeled as anomalous

5 Conclusion

In this paper, we highlighted the challenges and opportunities of explainable contextual anomaly detection. Our review on XAI, emphasizes the importance of selecting the appropriate explanation technique based on the constraints and the context. Moreover, we focused on an efficient explainable contextual algorithm: QCAD. Our experiments show that for more robust results, the specificities of each dataset should be taken into account, namely the already present anomalies. Finally, we suggest to add a preprocessing step to QCAD to dynamically define the different contexts which can be of different sizes within the same dataset. This will be investigated in our future research work where we will explore a more refined approach to the selection of the parameters for QCAD.

6 Acknowlegement

This work is supported by the National Gendarmerie Research Center, under the COFRA program.

References

[1] Sajid Ali et al. "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence". In: *Information fusion* 99 (2023), p. 101805.

[2] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2015.

[4] Nils Burkart and Marco F Huber. "A Survey on the Explainability of Supervised Machine Learning". In: *Artificial Intelligence Review* 54.5 (2021), pp. 4211–4238.

[5] Ece Calikus, Slawomir Nowaczyk, and Onur Dikmen. "Context discovery for anomaly detection". In: *International Journal of Data Science and Analytics* (2024), pp. 1–15.

[6] Kai Cao, Wenting Zhang, and Tianwei Wang. "Spatio-temporal land use multi-objective optimization: A case study in Central China". In: *Transactions in GIS* 23.4 (2019), pp. 726–744.

[7] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey". In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.

[8] Marco Dorigo and Christian Blum. "Ant colony optimization theory: A survey". In: *Theoretical Computer Science* 344.2-3 (2006), pp. 243–278.

[9] Dumitru Erhan et al. *Visualizing Higher-Layer Features of a Deep Network*. Tech. rep. University of Montreal Technical Report, 2009.

[10] Khaled Gubran Al-Hashedi and Pritheega Magalingam. "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019". In: *Computer Science Review* 40 (2021), p. 100402.

[11] Ibrahim Kök et al. "Explainable artificial intelligence (xai) for internet of things: a survey". In: *IEEE Internet of Things Journal* 10.16 (2023), pp. 14764–14779.

[12] Zhong Li and Matthijs Van Leeuwen. "Explainable contextual anomaly detection using quantile regression forests". In: *Data Mining and Knowledge Discovery* 37.6 (2023), pp. 2517–2563.

[13] Scott Lundberg. "A unified approach to interpreting model predictions". In: *arXiv preprint arXiv:1705.07874* (2017).

[14] Ali Bou Nassif et al. "Machine learning for anomaly detection: A systematic review". In: *Ieee Access* 9 (2021), pp. 78658–78700.

[15] Nagarajan Natarajan et al. "Learning with Noisy Labels". In: *Neural Information Processing Systems (NIPS)*. 2012.

[16] Binh P Nguyen et al. "Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records". In: *Computer methods and programs in biomedicine* 182 (2019), p. 105055.

[17] Guansong Pang et al. "Deep learning for anomaly detection: A review". In: *ACM computing surveys (CSUR)* 54.2 (2021), pp. 1–38.

[18] Marek Pawlicki, Rafał Kozik, and Michał Choraś. "A survey on neural networks for (cyber-) security and (cyber-) security of neural networks". In: *Neurocomputing* 500 (2022), pp. 1075–1087.

[19] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[20] Md Mostafizur Rahman and György Szabó. "Multi-objective urban land use optimization using spatial data: A systematic review". In: *Sustainable Cities and Society* 74 (2021), p. 103214.

[21] Douglas A Reynolds et al. "Gaussian mixture models." In: *Encyclopedia of biometrics* 741.659-663 (2009).

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[23] Ronald L Rivest. "Learning decision lists". In: *Machine learning* 2.3 (1987), pp. 229–246.

[24] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626.

[25] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *arXiv preprint arXiv:1312.6034* (2014).

[26] Xiuyao Song et al. "Conditional anomaly detection". In: *IEEE Transactions on knowledge and Data Engineering* 19.5 (2007), pp. 631–645.
[27] Giulia Vilone and Luca Longo. "Explainable artificial intelligence: a systematic review". In: *arXiv preprint arXiv:2006.00093* (2020).
[28] Véronne Yepmo, Grégory Smits, and Olivier Pivert. "Anomaly explanation: A review". In: *Data & Knowledge Engineering* 137 (2022), p. 101946.
[29] Matthew D. Zeiler and Rob Fergus. "Visualizing and Understanding Convolutional Networks". In: *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 818–833.
[30] Bolei Zhou et al. "Learning Deep Features for Discriminative Localization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2921–2929.