Demystifying Cyberattacks: Potential for Securing Energy Systems With Explainable AI

Shuva Paul, Sanjana Vijayshankar, Richard Macwan
Energy Security and Resilience
National Renewable Energy Laboratory, Golden, Colorado, USA
Email: {shuva.paul, svijaysh, richard.macwan}@nrel.gov

***Abstract*—Modernization of energy systems has led to increased interactions among multiple critical infrastructures and diverse stakeholders making the challenge of operational decision making more complex and at times beyond cognitive capabilities of human operators. The state-of-the-art machine learning and deep learning approaches show promise of supporting users with complex decision-making challenges, such as those occurring in our rapidly transforming cyber-physical energy systems. However, successful adoption of data-driven decision support technology for critical infrastructure will be dependent on the ability of these technologies to be trustworthy and contextually interpretable. In this paper, we investigate the feasibility of implementing explainable artificial intelligence (XAI) for interpretable detection of cyberattacks in the energy system. Leveraging a proof-of-concept simulation use case of detection of a data falsification attack on a photovoltaic system using XGBoost algorithm, we demonstrate how Local Interpretable Model-Agnostic Explanations (LIME), a flavor XAI approach, can help provide contextual and actionable interpretation of cyberattack detection.**

***Index Terms*—Artificial Intelligence, Cybersecurity, Energy System, Explainable Artificial Intelligence, Events and Anomaly Detection, Energy System Security, etc.**

I. Introduction

As energy systems evolve by integrating large-scale clean energy generation, flexible energy demand, emerging energy storage technologies, advanced telecommunications, and software–new cybersecurity risks emerge. There are ample opportunities for adoption of AI technology to resolve cybersecurity challenges. However, the opacity of most AI solutions can challenge users to understand and trust the model's decisions [1]. Lack of transparency and interpretability in AI-based decision-making systems are critical hurdles in the adoption of cybersecurity applications including risks of missing out crucial operational contexts and safety implications. For instance, if an automated cybersecurity intrusion/threat detection system blocks an IP address, flags a particular communication stream as compromised, or detects an instance of a data-stream as anomalous/cyberattack without a clear rationale, system administrators or operators will not be able to make an informed decision whether the red flags/detected anomalies are false positives, a minor threat, or a severe security breach. This ambiguity or lack of interpretability can lead to improper action execution, ranging from ignoring real threats to overreacting to benign activities. Therefore, trusting an opaque and unexplainable decision-making system not only increases the risk of operational difficulties, but also compromises the overall security of the system.

For AI technologies to be widely adoptable and impactful in managing these cyber risks will require trustworthy and easily interpretable explanations for system operators. Adding a layer of explainability will help to improve operators understanding of the cyberthreat detection mechanism, bolstering trust and transparency of the model. The U.S. Department of Energy initiated a number of programs and initiatives to enhance cybersecurity and resilience of future autonomous energy systems, as well as on the nation's existing critical energy infrastructure [2]. The Defense Advanced Research Projects Agency (DARPA) initiated the explainable artificial intelligence (XAI) program in 2017 to enable a deployed machine learning or deep learning model to explain its decisions to its user [3]. In parallel to the Department of Energy and DARPA, a myriad of both federal and private sector initiatives have emerged aiming to adopt artificial intelligence and machine learning to address the challenges in cybersecurity domain, such as the CISA Artificial Intelligence and Machine Learning for Cyber curriculum, and NIST special publications, technical notes, and interpretations [4], [5].

II. Preliminaries

XAI is a type of AI technique that helps an AI model to be transparent and interpretable. In simpler words, the reasoning behind the model's decision can be easily understood by general audiences without expert knowledge. According to [6], *"XAI encompasses Machine Learning (ML) or AI systems for demystifying black models internals and/or for explaining individual predictions."* XAI is a growing research domain among AI researchers and engineers and is being deployed in multidimensional research areas [7]–[10]. For instance, in healthcare XAI is being adopted in many areas including medical diagnosis, automate medical coding, monitoring patients using wearable, risk assessment, personalized medicine and decision making [11]. In aerospace, XAI is used for improving decision making processes in critical situations [12].

A. XAI Taxonomy

It is important to learn about the XAI taxonomy to properly deploy it in the domain-specific application. There are a few ways to categorize the use of XAI, including based on the type of explanation, application domain, and the levels of transparency [13]. XAI has previously been classified according to scope, methodology, usage, complexity, and models [14], [15].

979-8-3503-7099-7/24/$31.00 ©2024 IEEE

Based on the type of explanation, XAI can be categorized into four categories: inherently interpretable models, blended models, self-explaining models, and post hoc explanations [13]. Based on the level of transparency, XAI can be categorized into white-box, gray-box, and black-box model [16]. Similarly, based on application domain XAI categories include medicine, recommendation systems, and natural language processing [13]. It is important to understand the XAI taxonomy to enable effective cybersecurity solutions because of the diverse and evolving threat landscape, and the variety of machine learning models being deployed in different cybersecurity solutions. For instance, as the threats evolve, significant features or attributes indicative of the cyberattack might evolve as well. Feature attribution techniques, such as SHAP or LIME can highlight these concept drifts, hence assuring timely detection of the novel threats. This knowledge of different categories of XAI will help address model diversity and domain specific challenges in cybersecurity research by selecting right approaches of explanations. For instance, feature visualization or saliency mapping can help interpreting complex models adopted in critical cybersecurity applications by highlighting most important features or sequences.

B. Advantages of XAI in Energy Systems Cybersecurity

XAI can bring transparency and a human-level understanding about the inner workings of complex machine/deep learning models deployed in the decision-making and safety monitoring of energy systems. XAI can facilitate fast and accurate threat/attack/anomaly detection and enhance incident response processes because it can provide contextual explanations in a multidomain environment such as cyber physical energy systems. Another advantage is that XAI generates a human-understandable explanation, allowing users, stakeholders, and asset owners of AI-based solution providers to adopt these solution products with greater trust and reliability. In addition, AI models can be protected against attacks, errors, biases, and many other unforeseen threats arising from the black-box nature of models by adopting XAI techniques [17], [18].

III. Example Deployment of XAI for Energy System Attack Detection

In this section, we will present an example use case of leveraging XAI to cybersecure an energy system with distributed energy resources (DERs)–specifically, photovoltaic systems.

A. Experimental Setup

Our study leverages data from a prior research [19]. The foundation of this experiment was a hardware-in-the-loop (HIL) co-simulation environment, designed to develop and demonstrate software called DER Management System (DERMS) that can effectively manage and optimize the integration of DERs into the power grid. The experiment assumes that next generation smart meters are hosting a DERMS algorithm. The smart meter manages and communicates with the site's hardware inverters to meet system level goals, including voltage regulation and a virtual power plant. The HIL setup emulates a node controlled by the smart meter and a co-simulation representing the wider distribution network, including communication and control elements [20]. This experiment yielded a comprehensive dataset encompassing system, sensor and operational parameters.

B. Dataset

The data set consists of 35 features with $7,260$ data points for each of the features. This data set contains emulated data for electrical measurements (real and reactive control setpoints, forecast and measured power), co-simulation parameters (time delay, latency, and stepsize), and timestamps. For ease of experimentation, we selected 6 relevant features–namely, forecast and measured power as well as control setpoints for two unique and functionally different inverters (labeled ”sma” and ”fron,” respectively).

C. Threat Scenario

We assume a hypothetical threat scenario in which an attacker targets the photovoltaic plant by injecting anomaly attacks into the selected data streams. Two types of anomalies are injected: point anomalies and missing values, as shown in Algorithm 1. *Point anomaly*, also referred as global anomaly occurs when a portion of the dataset is considered anomalous with respect to or significantly different from the rest of the dataset [21]. On the other hand, a *missing value attack* pertains to the deliberate act of inserting or substituting particular data points with null or undefined values, mimicking the lack of data in situations where values are usually present. This type of attack has the potential to distort analyses, leave records unfinished, and potentially undermine the credibility of decisions based on data. In the context of adversarial attacks on datasets, a missing value attack can be compared to a variant of data vandalism, intentionally producing lack of information to impede its effectiveness and trustworthiness.

We replace the selected data point with manipulated values, or erase the value by replacing it with *NaN*. The modified datastream is received by the control center, and may lead the system operator to an inaccurate decision which can result in adverse impacts on system stability, equipment damage, or risk to personnel safety. Algorithm 1 helps injecting point anomalies and missing value attacks into the dataset and mimics the real-world data imperfections. This method tests the robustness and adaptability of the proposed solution model in handling incomplete or irregular data, which is common in practical scenarios.

D. Proposed Solution Approach

Figure 1 represents the proposed workflow for adopting XAI to detect cyberattacks. The solution approach takes historical measurements and manipulated data streams to train and test an anomaly detection algorithm. In this experiment, we train an XGBoost classifier to detect different types of anomalies and classify them. We then use Local Interpretable Model-agnostic Explanations (LIME) to explain different decisions of the model from the testing data, as discussed in the experimental results section. *LIME is a method of explanation with the aim of identifying an interpretable model over an interpretable representation that is locally faithful to the predictions of any classifier models* [22]. This technique is leveraged to explain

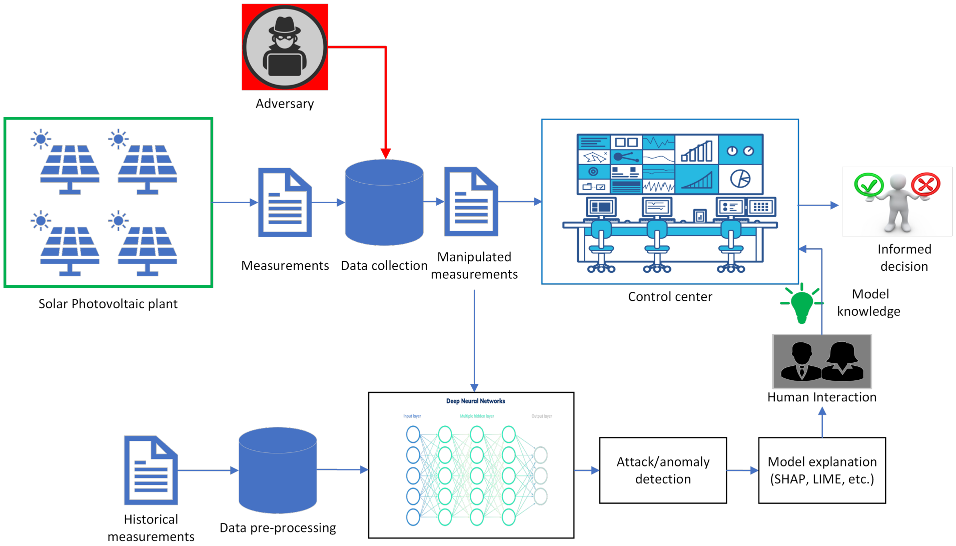

Fig. 1: A comprehensive workflow that delineates the progression from data generation and collection to attacks, their subsequent detection, and the explanatory procedures that elucidate the entire process.

Algorithm 1 Injecting point anomalies and missing value attacks

```
Require: Original Data
Ensure: Anomaly Data
Assign Total Anomaly Count
Point Anomaly Indices $\leftarrow$ empty list
Missing Value Indices $\leftarrow$ empty list
{Inject Point Anomaly attack.}
for $i \leftarrow 1$ to Total Anomaly Count do
    Select random row, col
    Modify value at row, col (e.g., value $\times$ 5)
    Append (row, col) to Point Anomaly Indices
end for
{Inject Missing Value attack.}
for $i \leftarrow 1$ to Total Anomaly Count do
    Select random row, col
    Set value at row, col to NaN
    Append (row, col) to Missing Value Indices
end for
Add label column to Anomaly Data indicating anomaly type
Plot Anomaly Data
return Anomaly Data =0
```

the predictive model (XGBoost, in this study) since LIME can analyze specific instances and explains how a specific instance can contribute to the model's prediction, rather than giving a generic explanation as to why this model is behaving in a particular way [23].

E. Experimental Results

In this subsection, we discuss the experimental results for anomaly detection, classification, and explanation of attacks on the photovoltaic system data set.

TABLE I: Classification report for XGBoost classifying attacks and anomalies in photovoltaic system dataset.

	Precision	Recall	F1-score	Support
0 (Normal)	1.00	0.98	0.99	84
1 (Point Anomaly Attack)	1.00	1.00	1.00	1284
2 (Missing Value Attack)	0.99	1.00	0.99	84
Accuracy			1.00	1452
Macro Avg	1.00	0.99	0.99	1452
Weighted Avg	1.00	1.00	1.00	1452

Table I represents the performance of the XGBoost method to classify the attacks to the photovoltaic system data. We can see that the adopted method successfully detects and classifies the point anomalies, missing values, and normal datapoints with high accuracy. Within our analysis, the classification report generated by the XGBoost model demonstrates the performance metrics across three distinct classes: normal, point anomaly attacks, and missing value attacks. *Precision* indicates the accuracy of positive predictions, showing the majority of the instances labeled within the classes are correctly predicted. *Recall* represents the model's adeptness at classifying the majority of genuine instances within a class. This minimizes the risk of overlooking crucial datapoints. The *F1-score* acts as a balanced representation of both precision and recall. This metric indicates an equilibrium between classifying true positives and avoiding false alarms. Uniformly high metrics across the classes add confidence that the model is capable of distinguishing between normal data and the two attack types. However, these metrics, while representing performance of the XGBoost model, do not clarify the reasoning behind the model's decisions.

Figure 2 represents the confusion matrix for the attack on the photovoltaic system dataset with proper representation of true and predicted labeling. The presented confusion matrix

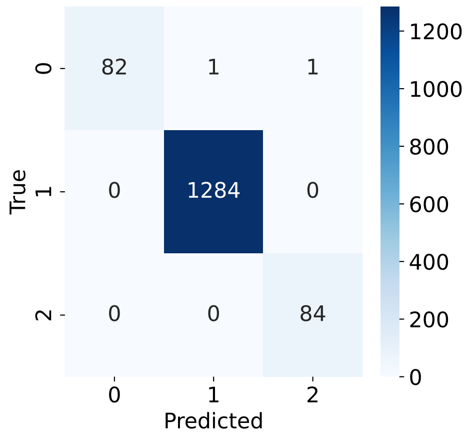

Fig. 2: Confusion matrix of the XGBoost algorithm in classifying true and predicted classes.

details the performance of our XGBoost classifier across three distinct classes: 0, 1, and 2, representing normal, point anomaly attacks, and missing value attacks, respectively. Each row of the matrix represents the true class instances, in contrast to the predicted classes in the columns. Although off-diagonal cells, with various color intensities, demonstrate misclassifications, the diagonal cells' strong color gradient represents true predictions for each class. A higher number of misclassified instances is indicated by a more intense color. From the dataset, it is evident that there is an imbalance. Although this type of imbalance is common in scenarios like the one under consideration, where the number of abnormal instances is likely to be higher than the normal instances, particularly during an attack, it can still be effectively addressed. Various techniques can be employed to mitigate this issue, including resampling methods (such as undersampling the majority class and oversampling the minority class), generating synthetic data, and implementing data augmentation strategies. Beyond immediate numerical insights into true and predicted classes, the matrix raises a deeper question: Why do specific misclassifications manifest and what intricacies drive them?

As mentioned earlier, we used LIME to interpret the XGBoost classifier as a predictive model to differentiate between normal and anomalous data points. LIME generates a surrogate model around a particular instance to understand the complex behavior of the black-box model. A surrogate model is an approximation method that mimics the behavior of a more complex model or system. It is often used when the original model is too computationally expensive or time-consuming to evaluate directly, or when the original model is a black-box and difficult to interpret. To assess a specific instance, LIME perturbs the instance's feature space and looks for the variations in the predictions of the model concerned (here, XGBoost is the model). In this experiment, we applied LIME to six selected features, which can help us visually understand and identify a particular feature, which is crucial in classifying a particular instance as attack (anomalous) or normal.

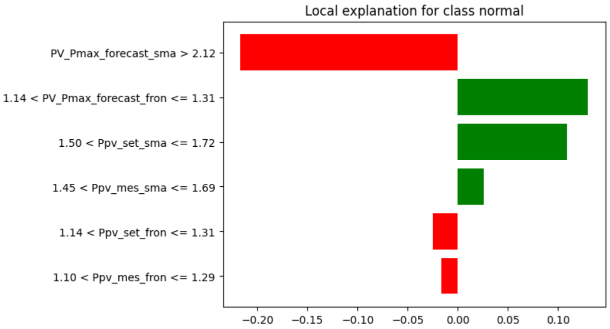

Fig. 3: Local explanations using LIME for photovoltaic dataset for 47th instance.

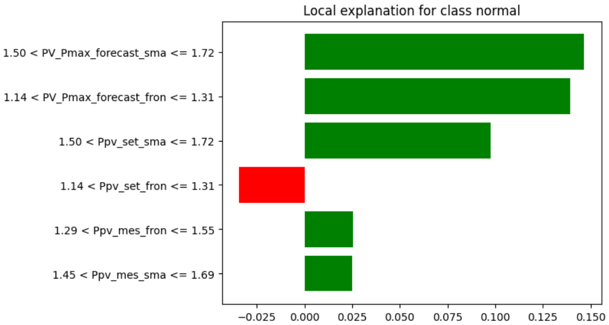

Fig. 4: Local explanations using LIME for photovoltaic dataset for 31st instance.

In Figure 3, we observe that, among the features assessed, one feature (PV_Pmax_forecast_sma) strongly stands out, being represented entirely in red for a particular instance. This results allows us to infer that this particular feature influenced the XGBoost model strongly in classifying this particular instance as anomalous data point because of its high absolute weight in the LIME explanation. On the other hand, the three features followed by PV_Pmax_forecast_sma, are in the green zone contributed toward a prediction of normalcy, which means that, in this particular instance it can be inferred that these three features were not manipulated during the attack. In Figure 4, we see a shift in the influence of the features on the model's predictive decision. Among the features evaluated, five are represented in green (contributing toward deciding normalcy) and one was represented in red (contributing to identification of an attack). Thus, we observe an instance-specific explanation, resulting in enhanced interpretability of the complex XGBoost model, leading toward building trust in the model's decision to detect anomalous datapoints resulting from a cyberattack.

It is crucial to consider the necessary proof of correctness for a machine learning-based system that involves XGBoost and LIME for anomaly detection in a complex dataset that requires a multifaceted approach. Firstly, XGBooptimizes the feature selection and decision making through gradient boosting, helping identify the anomalies. Second, the integration of LIME enhances the interpretability of the model by locally approximating its behavior on individual predictions. Finally, by injecting point anomalies and missing value attacks into the dataset, the effectiveness of the anomaly detection process can be tested and validated in a controlled environment. This ensures that the model is both statistically sound and applicable to real-world applications.

IV. Limitations, Challenges, Vulnerabilities

Even though XAI is increasingly becoming a new frontier of AI systems' adaptability, a wide variety of limitations, challenges, and vulnerabilities pose significant risks, requiring further extensive research on XAI for energy system cybersecurity. The complexity of explanations, the trade-off between accuracy and interpretability, time and resource-intensive computation, data privacy concerns, adversarial attacks, standardization, and regulation, are a few limitations and challenges of XAI's utility for cybersecurity deployments. In addition, the explanation method needs to capture the physics or complexity of the energy system and learn the interdependencies to properly explain the behavior of the models. Additionally, the explainability is vulnerable to cyberattacks by malicious actors, including model manipulation, information leakage, model poisoning, model evasion, model stealing, and many other machine learning attacks.

V. Conclusion And Future Work

Complexities and uncertainties of modern power systems are leading to the deployment of complex machine learning/deep learning models to address challenges in cybersecurity. In this article, we have shown the challenges of leveraging machine learning based cybersecurity approaches and the importance of explainability to address these challenges. Leveraging the usecase of a cyberattack on a PV plant we demonstrated how LIME based explanations can help interpret the results of the XGBoost based attack detection technique. While this paper provides an initial proof of concept use case for XAI in energy system cybersecurity, we aim to expand on this work in the future by investigating different types of explanations including physics based and counterfactual explanations. By providing interpretable results this work aims to accelerate the adoption of machine learning based technologies in the field and help establish and maintain trust in these technologies.

Acknowledgment

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the Laboratory Directed Research Program. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

References

[1] W. J. von Eschenbach, "Transparency and the black box problem: Why we do not trust ai," *Philosophy & Technology*, vol. 34, no. 4, pp. 1607–1622, 2021.

[2] "Doe multiyear plan for energy sector cybersecurity," 2018.

[3] D. Gunning, "Explainable artificial intelligence (xai). darpa," *I20 (DARPA 2017)*, 2017.

[4] N. I. of Standards and Technology, "Security and privacy controls for information systems and organizations," 2023.

[5] Cybersecurity and I. S. Agency, "Cybersecurity infrastructure security agency," August 2023.

[6] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable artificial intelligence approaches: A survey," *arXiv preprint arXiv:2101.09429*, 2021.

[7] M. Nazar, M. M. Alam, E. Yafi, and M. M. Su'ud, "A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques," *IEEE Access*, vol. 9, pp. 153316–153348, 2021.

[8] L. Gaur and B. M. Sahoo, "Introduction to explainable ai and intelligent transportation," in *Explainable Artificial Intelligence for Intelligent Transportation Systems: Ethics and Applications*, pp. 1–25, Springer, 2022.

[9] R. Farrow, "The possibilities and limits of xai in education: a socio-technical perspective," *Learning, Media and Technology*, pp. 1–14, 2023.

[10] R. Machlev, L. Heistrene, M. Perl, K. Levy, J. Belikov, S. Mannor, and Y. Levron, "Explainable artificial intelligence (xai) techniques for energy and power systems: Review, challenges and opportunities," *Energy and AI*, vol. 9, p. 100169, 2022.

[11] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Computer Methods and Programs in Biomedicine*, p. 107161, 2022.

[12] S. Sutthithatip, S. Perinpanayagam, S. Aslam, and A. Wileman, "Explainable ai in aerospace for enhanced system performance," in *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, pp. 1–7, 2021.

[13] G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts," *Data Mining and Knowledge Discovery*, pp. 1–59, 2023.

[14] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.

[15] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[16] T. Speith, "A review of taxonomies of explainable artificial intelligence (xai) methods," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, (New York, NY, USA), p. 2239–2250, Association for Computing Machinery, 2022.

[17] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (xai)," *IEEE Access*, vol. 11, pp. 78994–79015, 2023.

[18] A. Drichel and U. Meyer, "False sense of security: Leveraging xai to analyze the reasoning and true performance of context-less dga classifiers," *arXiv preprint arXiv:2307.04358*, 2023.

[19] J. Comden, J. Wang, and A. Bernstein, "Study of communication boundaries of primal-dual-based distributed energy resource management systems (derms)," in *2023 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pp. 1–5, 2023.

[20] J. Comden, J. Wang, S. Ganguly, S. Forsythy, R. Gomezy, and A. Bernstein, "Hardware-in-the-loop evaluation of grid-edge der chip integration into next-generation smart meters," in *Proceedings of the Smart Grid Comm Conference*, 2023. Accepted for publication.

[21] M. Zhao and J. Chen, "A review of methods for detecting point anomalies on numerical dataset," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1, pp. 559–565, 2020.

[22] M. T. Ribeiro, S. Singh, and C. 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*, pp. 1135–1144, 2016.

[23] J. Dieber and S. Kirrane, "Why model why? assessing the strengths and limitations of lime," *arXiv preprint arXiv:2012.00093*, 2020.