Decoding Anomalies! Unraveling Operational Challenges in Human-in-the-Loop Anomaly Validation

Dong Jae Kim
Steven Locke
Tse-Hsun (Peter) Chen
{k_dongja,s_loc,peterc}@encs.concordia.ca
Concordia University
Montreal, Canada

Andrei Toma
Sarah Sajedi
Steve Sporea
Laura Weinkam
{andrei.toma,steve.sporea,laura.weinkam,sarah.sajedi}@era-ehs.com
ERA environmental management solutions
Montreal, Canada

ABSTRACT

Artificial intelligence has been driving new industrial solutions for challenging problems in recent years, with many companies leveraging AI to enhance business processes and products. Automated anomaly detection emerges as one of the top priorities in AI adoption, sought after by numerous small to large-scale enterprises. Extending beyond domain-specific applications like software log analytics, where anomaly detection has perhaps garnered the most interest in software engineering, we find that very little research effort has been devoted to post-anomaly detection, such as validating anomalies. For example, validating anomalies requires human-in-the-loop interaction, though working with human experts is challenging due to uncertain requirements on how to elicit valuable feedback from them, posing formidable operationalizing challenges. In this study, we provide an experience report delving into a more holistic view of the complexities of adopting effective anomaly detection models from a requirement engineering perspective. We address challenges and provide solutions to mitigate challenges associated with operationalizing anomaly detection from diverse perspectives: inherent issues in dynamic datasets, diverse business contexts, and the dynamic interplay between human expertise and AI guidance in the decision-making process. We believe our experience report will provide insights for other companies looking to adopt anomaly detection in their own business settings.

CCS CONCEPTS

• **Software and its engineering → Requirements analysis**.

KEYWORDS

Anomaly Validation, Requirement Engineering

FSE Companion '24, July 15–19, 2024, Porto de Galinhas, Brazil

ACM ISBN 979-8-4007-0658-5/24/07
https://doi.org/10.1145/3663529.3663857

ACM Reference Format:
Dong Jae Kim, Steven Locke, Tse-Hsun (Peter) Chen, Andrei Toma, Sarah Sajedi, Steve Sporea, and Laura Weinkam. 2024. Decoding Anomalies! Unraveling Operational Challenges in Human-in-the-Loop Anomaly Validation. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering (FSE Companion '24), July 15–19, 2024, Porto de Galinhas, Brazil.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3663529.3663857

1 INTRODUCTION

Numerous traditional software applications, including compliance reports, bookkeeping, and management tools, heavily depend on user-provided data for functionality. These systems store and process user-provided data to assist users in achieving their business goals, such as filing tax returns or generating reports. Therefore, ensuring the quality of user input is imperative from a software engineering perspective for product success.

To identify potential errors in the data, anomaly detection algorithms can be applied, revealing data points that may deviate from the normal distribution in the rest of the datasets. While extensive algorithms exist for anomaly detection, the most challenging, yet rarely discussed step in the machine learning pipeline, is the validation of anomalies [1, 6, 8, 10]. For example, most anomaly detection in software engineering, such as log analytics, takes advantage of pre-labeled data, such as loghub [22], HDFS (Amazon EC2) [21] or BGL [17]. The majority of efforts focus on enhancing the performance of evaluation metrics for anomaly detection mechanisms, yet there is a lack of emphasis on post-detection validation.

However, it is not enough to say that some data points are anomalous; there is a *"cause and an effect"* relationship. Most anomaly detection recommender systems, even in a supervised setting, where the labels exist, only predict a data entry as a binary classification of anomaly vs non-anomaly [7, 22]. However, there are always contexts involved with anomalies, i.e., for user-provided datasets, seasonality may influence inflation, and then we can no longer say some data points are anomalous. To incorporate such contextual understanding, feedback from domain experts is essential. Therefore, a critical software requirement is to facilitate the fast delivery of anomalies to domain experts from data collection sensors. However, challenges persist in presenting these anomalies effectively, as they may not be easily comprehensible to domain experts. Furthermore, the anomalous behavior is often dynamic in nature, e.g., new types of anomalies might arise, for which there is no labeled training

data [4]. Hence, the most challenging aspect of anomaly detection lies in requirement engineering for defining and validating anomalies. At times, the reason behind an anomaly may be unclear, and even when anomalies are identified, domain experts may not consider them significant due to their contextual nature. Therefore, the question arises: *"How can we effectively collaborate with domain experts and contextualize their feedback to ensure actionable insights, while aligning with constantly changing business values?"* Addressing such validation challenge involves collaboration among many stakeholders, such as data scientists, software engineers, domain experts, and end-users, to establish comprehensive validation criteria. However, incorporating user feedback and domain-specific knowledge to refine anomaly detection models and enhance their precision presents significant challenges.

In this work, we present an experience report detailing the integration of human-in-the-loop validation for anomalies detected by AI models within the industrial partners's software. The subject system of our industrial partners is a large-scale business-to-business enterprise software specializing in environmental and manufacturing domains. Unlike many prior studies that solely recommend anomalies [1, 6, 8, 10], or provide recommendations with embedded explanations [18, 19], none actually go beyond incorporating user feedback to actively improve the machine learning evaluation. Our focus is on elucidating the derivation process of these anomalies during the post-mortem presentation to domain experts. Thus, our experience report discusses the complexities inherent in anomaly validation and proposes strategies to enhance anomaly detection and validation processes. The main contribution of our work are: (1) We report the challenges encountered by domain experts during the anomaly validation process, involving a 12-month-long collaboration and interviews, and (2) we report several valuable insights into a more successful post-anomaly validation processes.

Paper Organization. Section 2 discusses our background. Section 3 discusses the challenges encountered in the integration and our solutions to these challenges. Section 4 discusses implications. Finally, Section 5 concludes the paper.

2 BACKGROUND

In this section, we delve into the background regarding what are anomalies in data and their significance within the context of our subject systems.

2.1 What is an Anomaly?

Data anomalies are deviations or irregularities in a dataset diverging significantly from expected patterns. They are categorized into three types: point anomalies, contextual anomalies, and collective anomalies. Point anomalies are individual data points substantially different from the majority of the dataset, like an unusually high or low temperature spike in a weather dataset. Contextual anomalies occur when data points exhibit abnormal behavior within specific contexts, such as a spike in consumer transactions during holidays. Collective anomalies involve groups of data points displaying unusual characteristics when considered together, like identical amounts and frequencies in a series of transactions within a bank dataset. While these anomaly detection techniques analyze data distributions, they may not fully capture the system's requirements aligning with user values. Therefore, specifying requirements on what constitutes an anomaly can tailor the detection process to the unique needs and objectives of the domain experts.

2.2 Common Approaches to Anomaly Detection

In general, domain experts employ diverse anomaly detection methods to identify anomalies. Often, the processes generating these anomalies are unknown, leading to anomaly detection operating in an unsupervised manner. Unsupervised learning is a method in machine learning where, in contrast to supervised learning, algorithms learn patterns exclusively from unlabeled data [2]. There are many unsupervised-learning algorithm in the wild, such as *clustering-based* [12], *density-based* [5], *kernel-based* [20], *tree-based* [15] and *deep-learning based* [13]. These methods typically search for outliers in datasets, with the assumption that they represent anomalies in real-world applications.

2.3 Anomalies within the context of our subject system

The subject system of our industrial partners is a large-scale business-to-business enterprise software specializing in environmental and manufacturing domains. This software empowers manufacturers to manage and regulate chemical usage, ensuring compliance with government health and safety regulations. The software handles thousands of user inputs with many configurations, values, and units daily. Recognizing the direct impact of data quality on manufacturing companies, our industrial partners have deployed dedicated environmental analysts to verify the quality of inputs. Maintaining high-quality data is of utmost interest to stakeholders (i.e., manufacturers) since even slight deviations may cause negative consequences for stakeholders against government regulations. To reduce the manual intensive stage of anomaly detection, we implemented and integrated AI to automatically detect anomalies and recommend them to domain experts. While many algorithms can be used to detect anomalies in the data, the most challenging stage is the validation of anomalies. Challenges in validation involves determining how anomalies should be defined, considering that there are edge cases where anomalies may not be perceived as problematic by domain experts. Below, we will further discuss the operationalization challenges of anomaly requirement and validation.

3 CHALLENGES IN REQUIREMENT ENGINEERING FOR ANOMALY DETECTION

In this section, we report our insights into challenges and solutions stemming from a 12-month collaboration and interviews with 8 stakeholders from diverse backgrounds, including project managers, software managers, and environmental scientists.

3.1 Navigating Dynamic Data Environments and Varied Expert Perspectives

Validating anomalies is like finding "*needles in a haystack*", especially when dealing with constantly changing data in different business contexts. Imagine this challenge magnified by the diverse opinions of experts, each suggesting unique ways to make sense of

Table 1: Different Hierarchies in Data Aggregation

Time	Value	New Date	New Value
1/1/2024	1.5	Day 1	1.5
1/2/2024	1.5	Day 2	1.5
1/3/2024	1.5	Day 3	1.5
1/8/2024	1.5	Day 4	0.3
1/15/2024	1.5	Day 5	0.3

(a) Roll-down approach

Time	Value	New Time	New Value
1/1/2024	1.5		
1/2/2024	1.5	Week 1	4.5
1/3/2024	1.5		
1/8/2024	1.5	Week 2	1.5
1/15/2024	1.5	Week 3	1.5

(b) Roll-up approach

the anomalies. Effective validation is about striking the balance between everyday data fluctuations and the intricate decisions shaped by different domain experts.

3.1.1 Challenges in Verifying Anomalies within Hierarchical Dataset. There are significant challenges in managing data points at different temporal granularities. In a simple case of time-series dataset, data points are consistently recorded at frequent intervals from monitoring sensors, simplifying the process of feature engineering. These data points can either be used as-is (daily-basis) or undergo feature engineering, aggregating them into much coarser granularity like monthly or yearly intervals. Subsequently, an anomaly detection algorithm is applied to detect anomalies. This represents the most ideal scenario of feature engineering and model construction in the machine learning pipeline. However, in a business context, data points are dynamic and may lack temporal consistency. As illustrated in Table 1 (a), incoming data may transition from a frequent (daily) interval to a infrequent (weekly) intervals starting on 1/8/2024, introducing complexity to data aggregation. To conduct anomaly detection on a hierarchical dataset, applying feature engineering is essential to transform the dataset into consistent time-scale. For example, we could apply (a) the roll-down strategy or (b) the roll-up strategy. In the roll-down approach, the coarser granularity (weekly) is transformed into a much finer time granularity (daily), while in the roll-up approach, the finer granularity (daily) is transformed into a much coarser time granularity (weekly).

Despite the technical soundness of these strategies, considerations arise regarding end-user comprehension. The roll-down strategy may potentially mislead end-users by altering the original input data. Hence, domain experts need to map the anomalies back to the original input for better anomaly comprehension. Conversely, the roll-up strategy risks losing detailed information necessary for detecting certain anomalies, as some anomalies are only detectable at finer data points (day-to-day) rather than coarser ones (week-to-week).

Moreover, the hierarchical nature of datasets across user behavior anomalies exacerbates these challenges, affecting both anomaly detection and validation processes. In Table 1 (a), 0.3 may be flagged as anomalous if it is low compared to other values, whereas in Table 1 (b), 4.5 may be flagged as anomalous for being too large. The use of two feature engineering techniques provides different perspectives on what should be considered anomalous. However, this perspective is purely algorithmic, assuming that data should be evenly distributed across time duration. Domain experts may offer alternative viewpoints, such as *"It may not be anomalous since we do not care about when it is recorded, but that whenever it is recorded, it should be consistently 1.5."* This difference in perspectives highlights how challenging it is to confirm anomalies. Depending on how the data is grouped, models might identify different anomalies. Yet, many studies do not delve into these specifics of data grouped. Nonetheless, these details are vital for pinpointing anomalies accurately, emphasizing the importance of considering data aggregation in hierarchical data in anomaly detection methods. Therefore, ***feature engineering and anomaly explanation are closely intertwined, as they both rely on domain expertise to capture crucial elements of business value.***

Table 2: Unified View of Hierarchical Dataset.

Monthly	Value	Weekly	Value	daily	Value
1/2024	7.5	1/1/2024	4.5	1/1/2024	1.5
				1/2/2024	1.5
				1/3/2024	1.5
		1/8/2024	1.5		
		1/15/2024	1.5		

3.1.2 Challenges in Incorporating Diverse Feedback and Varied Perspectives. Domain experts often differ in their preferences regarding what recommendations they find valuable; what holds importance for one may not necessarily be prioritized by another within the development team. For example, due to difference in their responsibilities and roles within the team, it is possible that managers may want a more holistic view of the anomalies. In contrast, more technical experts may require more detailed information. Due to the challenge posed by hierarchical data, some domain experts (not all) prioritize specific perspectives within the dataset. One group of domain experts expressed that while the daily perspective offers detailed insights, it tends to exhibit significant variance, making summarization more challenging and harder to make actionable decisions. Hence, they expressed that a high-level view of the summarization is easier to understand for novice domain experts new to the validation task. However, other group domain experts said, a detailed view poses no issue and they prefer a detailed view as it facilitates root cause analysis. While both presentations are equally valuable, it is challenging to know which view we should adopt to improve the decision-making skills of all of the domain experts.

When domain experts have differing views, incorporating diverse perspectives becomes challenging. We attempted to mitigate this issue by showing anomalies in the data in a way that everyone could understand. Table 2 shows the detected anomalies concurrently across all time granularities. We prioritize displaying coarser

granularity first, meeting the preference of some experts for higher-level summarization that is more understandable and actionable. Subsequently, as indicated by other domain exerts, we integrated the finer granularity in parallel to the rest which needs detailed information to drill down to the root causes of anomalies. When presented with a unified view, both opposing sides agreed to the approach, highlighting the effectiveness of finding a middle ground.

This experience highlights the need for engineers to find a balance between making data-driven actions more general and accessible to domain experts while considering the integration of sophisticated analytical methods that may introduce complexity and pose challenges for interpretation. Therefore, ***quick feedback of our anomalies is important for end-to-end optimization of the anomaly validation. In our case, we find that summarizing anomalies into a higher granularity allows ease of interpretation and actionability.*** In hindsight, while achieving high accuracy is an important goal of AI, depending on the business domain, where AI assists the decision-making process of the domain experts, perhaps sacrificing high performance for better explainability may be beneficial for obtaining better feedback.

3.2 Balancing Model Recommendations with Human Expertise in Decision-Making

The decision-making process involves a nuanced interplay between AI and human involvement. While AI interaction should be conversational: AI serves as an assistant, and humans ultimately make decisions. However, in our experience, particularly with unexplainable anomalies, humans may find themselves compelled to follow AI recommendations. Consequently, the challenge lies in ensuring that AI enhances, rather than undermines, the expertise of human validators. Likewise, domain experts must leverage their knowledge without blindly trusting AI results. Navigating this delicate equilibrium is crucial to avoiding pitfalls in the anomaly validation process.

3.2.1 Challenges in Distinguishing True Anomalies from Context-Driven Anomalies. Finding anomalies is interesting when there is no clear explanation. Domain experts, in uncovering anomalies, conduct root cause analysis to understand unique circumstances. If such circumstances explain the reason for the anomalies, then they become the new context. This prompts adjustments to feature engineering or model, removing anomalies from future analysis, as they are special circumstances, not driven by data distribution. Such workflow is an example of *human-in-the-loop*, where AI applications rely on numerous iterations of trial-and-error processes of experimentation and feedback from domain experts. Specifically, as discussed from Table1, due to hierarchical nature of the dataset, where the anomalies depend on the type of feature engineering and context, one domain expert remarked, "*Time variation can be normal. We probably should not care about the roll-up approach for anomaly detection. In the future we should remove it to reduce efforts.*" However, this may also depend on the case-by-case scenario, since the software is in a business-to-business domain, and different businesses may have different business logic.

True anomalies, by definition, do not follow data distribution, lacks easily understood contexts, making them more challenging to identify than anomalies driven by specific contextual factors. In the literature, there is limited discussion on what happens in the initial iteration of anomaly detection, particularly when anomalies cannot be explained. For instance, while software log anomaly detection often provides recommendations, the explanation of anomalies may be absent, as most models strive to predict that some event is anomalous or non-anomalous [7, 9, 10, 14, 22]. This raises challenges when deriving labels in the first iteration, as users may struggle to explain unexplained anomalies. Consequently, ***true anomalies are more difficult to explain, as domain experts may not understand why the model has flagged certain results as anomalies and may struggle to perform root cause analysis.***

3.2.2 Challenges in Balancing Human Decision or AI decision. The increasing reliance on AI models in decision-making poses significant challenges. One major concern is that this dependence can lead domain experts to unquestioningly trust the black-box decisions of AI systems, altering their perception of what constitutes an anomaly. Previous research by Bogert et al. [3] supports this, highlighting domain experts' tendency to favor AI-generated results over other sources.

In our experience, when faced with unexplained anomalies, domain experts often turn to AI for explanations, i.e., seeking *algorithms transparency*, asking questions like, "*Why has the model decided that such pieces of values are anomalous?*". Given the algorithmic assumption that anomalies deviate from the normal distribution, an initial explanation might be, "*Although data fluctuates, it is consistent between 1 and 4, whereas 3 is rare, hence anomalous.*" While sound, this illustrates that if domain experts struggle to interpret results, they may overly rely on AI-generated explanations. One side effect illustrated here is that there is a risk that the model may start to take a more dominant role in guiding the validation process, rather than serving as an assistant for user decisions.

Moreover, this reliance on AI explanations introduces the risk of confirmation bias, where users favor explanations that align with their preconceived beliefs, even if those beliefs are influenced by the AI itself [11]. Consequently, ***when anomalies cannot be explained, there is a danger that AI may increasingly drive decision-making processes, potentially compromising the effectiveness of anomaly validation.***

3.2.3 Challenges in Transparent Model Explanations in Anomaly Detection. While a seemingly naive solution to mitigate the aforementioned issue of human validators conforming to AI is to "*make human-in-the-loop give us the context in the anomalies*," in practice, domain experts find it challenging to reason about these anomalies. Providing clear and understandable justifications for why certain values are flagged as anomalies can empower domain experts and alleviate concerns about blindly conforming to a blackbox model.

To explain the black-box model, we could adopt feature importance strategy, which shows that in high-dimensional data anomaly detection, such feature influenced anomalies. However, improving anomaly explanations for domain experts is tricky. They usually need the model to explain results in simpler language. For example, explaining model anomalies, especially those related to data distribution, is still a challenge for domain experts. Interestingly, making these explanations better relies on the expertise of domain professionals to reason about anomalies to integrate into the model, creating a bit of a paradox. Therefore, ***achieving explainability***

in anomaly detection is challenging. On one hand, domain experts demand explainability in the results, yet obtaining explainability in the model necessitates domain knowledge.

4 DISCUSSION

4.1 Need for a better holistic approach

A holistic approach to anomaly detection entails a comprehensive consideration of both the technical components and the human-centric aspects involved in the validation process. This implies that effective anomaly detection strategies go beyond purely algorithmic considerations and extend to understanding the nuances of human decision-making, domain expertise, and contextual understanding. Namely, there is a need for a synergy between data-driven insights provided by algorithms and the qualitative insights contributed by human experts. Therefore, ***balancing technical sophistication with an understanding of the broader operational landscape becomes pivotal for successful anomaly validation.***

4.2 Supporting anomaly validation from novice to expert: dynamic recommendations

Dynamic recommendations play a crucial role in assisting anomaly validation across a spectrum of expertise levels, from novices to experts. In anomaly detection systems, dynamic recommendations must provide tailored suggestions and insights based on the user's proficiency level, domain knowledge, and specific requirements. For novice users who are new to anomaly validation tasks, dynamic recommendations should focus on providing simplified and easy-to-understand insights. These recommendations should aim to guide novices through the validation process by highlighting key anomalies. On the other hand, for expert users who are experienced in anomaly validation, dynamic recommendations should offer more advanced and detailed insights. These recommendations can include in-depth analyses of anomalies. Expert users may also benefit from dynamic recommendations that highlight subtle anomalies or patterns that may not be immediately apparent, allowing them to uncover hidden insights and make informed decisions. Therefore, ***dynamic recommendations in anomaly detection systems should adapt to the user's expertise level and provide personalized support throughout the validation process.***

4.3 Support for better model explanation to human-in-the-loop anomaly validation

Validating anomalies can be challenging for domain experts when anomalies lack clear explanations. If explaining anomalies detected by the AI model proves challenging, one can utilize simpler *proxy explanation*. For example, instead of directly explaining the model's decisions, utilize proxy explanations that provide intuitive or analogical reasoning for why certain anomalies were flagged. Analogies can sometimes make complex concepts more accessible to domain experts, even if they do not fully capture the underlying model mechanics. Based on our experience, detailed model mechanics may not necessarily enhance domain experts' intuition in a manner that facilitates anomaly validation. For instance, *algorithm transparency* [16], may not necessarily benefit domain experts' understanding.

More concretely, while AI offers automatic feature extraction and anomaly detection, it is still a black-box and hard to understand. We argue that there may still be value in utilizing heuristics in order to explain complex contextual anomalies. Benefit of heuristics is that we know exactly the anomalies that we are looking for in the data search space, which is easier for domain experts to understand. One opportunity for improving explanation is to use AI to narrow down search space, and upon understanding the contexts behind anomalies, heuristics detection can be formulated on top of the anomalies. Such strategy may be similar to *explain-by-example* methods in anomaly detection [16]. This heuristics are examples that are detected and presented to the domain experts. Moreover, these heuristics are the database of past anomalies encountered by domain experts to establish a knowledge base. With information retrieval techniques, this knowledge base can be leveraged to explain anomalies.

5 CONCLUSION

In this paper, we provide an experience report discussing the challenges in the requirement engineering process for anomaly validation in dynamic data environments with diverse feedback from domain experts, aiming to guide anomaly validation for other practitioners.

Firstly, in dynamic systems with hierarchical datasets, feature engineering significantly influences anomaly detection and interpretation. For example, data aggregation may reveal different anomalies and be valued differently by domain experts with varying expertise levels. To accommodate these diverse opinions, we provide a unified visualization, starting with a more generalized and higher-level summary, then narrowing down to show the complexities of specific anomalies. Dynamic recommendations tailored to users' expertise levels can assist in anomaly validation, guiding users through the process and providing personalized support.

Secondly, we discuss challenges associated with balancing model recommendations with human expertise in decision-making. Domain experts often find it hard to reason about anomalies and need explanations from the AI. However, anomaly detection algorithms based on data distribution cannot provide context, which can only be derived from human intelligence. This often leads to a tendency for humans to blindly accept the results returned by the AI. Consequently, there is an ongoing cycle of *"AI as a snake that bites its own tail"*, where AI learns from the data that is blindly accepted by humans as anomalous, potentially compromising the quality of anomaly validation.

Hence, enhancing AI explainability with simplified proxy explanations can provide intuitive reasoning for why certain anomalies were flagged, even if they do not fully capture the underlying model mechanics. A holistic approach to anomaly detection, considering both technical sophistication and an understanding of the broader operational landscape, as well as the human aspect, is necessary. Addressing these challenges in anomaly validation can improve the operationalization of anomaly detection systems to meet the complex requirements of real-world applications.

REFERENCES

[1] Crispin Almodovar, Fariza Sabrina, Sarvnaz Karimi, and Salahuddin Azad. 2024. LogFiT: Log Anomaly Detection using Fine-Tuned Language Models. *IEEE Transactions on Network and Service Management* (2024).

[2] Horace B Barlow. 1989. Unsupervised learning. *Neural computation* 1, 3 (1989), 295–311.

[3] Eric Bogert, Aaron Schecter, and Richard T Watson. 2021. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports* 11, 1 (2021), 8028.

[4] Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Whang, and Martin Zinkevich. 2019. Data Validation for Machine Learning.. In *MLSys*.

[5] Mete Çelik, Filiz Dadaşer-Çelik, and Ahmet Şakir Dokuz. 2011. Anomaly detection in temperature data using DBSCAN algorithm. In *2011 international symposium on innovations in intelligent systems and applications*. IEEE, 91–95.

[6] Zhuangbin Chen, Jinyang Liu, Wenwei Gu, Yuxin Su, and Michael R Lyu. 2021. Experience report: Deep learning-based system log analysis for anomaly detection. *arXiv preprint arXiv:2107.05908* (2021).

[7] Qiang Fu, Jian-Guang Lou, Yi Wang, and Jiang Li. 2009. Execution anomaly detection in distributed systems through unstructured log analysis. In *2009 ninth IEEE international conference on data mining*. IEEE, 149–158.

[8] Hongcheng Guo, Jian Yang, Jiaheng Liu, Jiaqi Bai, Boyang Wang, Zhoujun Li, Tieqiao Zheng, Bo Zhang, Qi Tian, et al. 2024. LogFormer: A Pre-train and Tuning Pipeline for Log Anomaly Detection. *arXiv preprint arXiv:2401.04749* (2024).

[9] Shayan Hashemi and Mika Mäntylä. 2021. OneLog: Towards end-to-end training in software log anomaly detection. *arXiv preprint arXiv:2104.07324* (2021).

[10] Shilin He, Jieming Zhu, Pinjia He, and Michael R Lyu. 2016. Experience report: System log analysis for anomaly detection. In *2016 IEEE 27th international symposium on software reliability engineering (ISSRE)*. IEEE, 207–218.

[11] Joshua Klayman. 1995. Varieties of confirmation bias. *Psychology of learning and motivation* 32 (1995), 385–418.

[12] Rashmi Kumari, MK Singh, R Jha, NK Singh, et al. 2016. Anomaly detection in network traffic using K-mean clustering. In *2016 3rd international conference on recent advances in information technology (RAIT)*. IEEE, 387–393.

[13] Donghwoon Kwon, Hyunjoo Kim, Jinoh Kim, Sang C Suh, Ikkyun Kim, and Kuinam J Kim. 2019. A survey of deep learning-based network anomaly detection. *Cluster Computing* 22 (2019), 949–961.

[14] Van-Hoang Le and Hongyu Zhang. 2021. Log-based anomaly detection without log parsing. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 492–504.

[15] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2012. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, 1 (2012), 1–39.

[16] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review* 56, 4 (2023), 3005–3054.

[17] Adam Oliner and Jon Stearley. 2007. What supercomputers say: A study of five system logs. In *37th annual IEEE/IFIP international conference on dependable systems and networks (DSN'07)*. IEEE, 575–584.

[18] Capital One. 2022. *Operationalizing Machine Learning Achieves Key Business Outcomes.* https://ecm.capitalone.com/WCM/tech/forrester-pdfs/capitaloneforrestersnapshotmloctober2022.pdf

[19] Jiho Shin, Reem Aleithan, Jaechang Nam, Junjie Wang, and Song Wang. 2021. Explainable software defect prediction: Are we there yet? *arXiv preprint arXiv:2111.10901* (2021).

[20] Ashish Sureka. 2015. Kernel based sequential data anomaly detection in business process event logs. *arXiv preprint arXiv:1507.01168* (2015).

[21] Wei Xu, Ling Huang, Armando Fox, David Patterson, and Michael I Jordan. 2009. Detecting large-scale system problems by mining console logs. In *Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles*. 117–132.

[22] Jieming Zhu, Shilin He, Jinyang Liu, Pinjia He, Qi Xie, Zibin Zheng, and Michael R Lyu. 2019. Tools and benchmarks for automated log parsing. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 121–130.

Received 2024-02-08; accepted 2024-04-18