EUROPEAN JOURNAL OF EMERGING CYBERSECURITY AND INFORMATION PROTECTION

Toward Unified Interpretability and Robustness in Machine Learning–Based Anomaly Detection Across Industrial, Network, Financial, and Cyber-Physical Domains

Johannes K. Albrecht
Technical University of Munich, Germany

Lukas M. Reinhardt
ETH Zurich, Switzerland

V0LUME03 ISSUE01 (2026)

Published Date: 08 January 2026 // Page no.: - 01-05

ABSTRACT

Anomaly detection has emerged as one of the most intellectually complex and practically consequential subfields of machine learning, driven by the accelerating digitization of industrial processes, networked infrastructures, financial systems, and cyber-physical environments. Across these domains, anomalies often represent rare, evolving, and context-dependent deviations whose significance is not merely statistical but operational, economic, and ethical. This research article develops a comprehensive, theory-driven, and empirically grounded synthesis of machine learning–based anomaly detection by integrating insights from industrial process monitoring, network intrusion detection, financial fraud analysis, healthcare Internet of Things security, and large-scale data systems. Drawing on a broad and deliberately heterogeneous body of literature, the article advances a unified interpretive framework that explains why anomaly detection remains resistant to universal solutions despite decades of algorithmic innovation. Particular emphasis is placed on comparative methodological perspectives, including classical statistical approaches, shallow machine learning methods, kernel-based novelty detection, clustering paradigms, and deep learning architectures. The analysis is anchored by contemporary empirical findings in industrial screw driving data, which illustrate how algorithmic performance is inseparable from domain semantics, feature engineering choices, and evaluation protocols (West and Deuse, 2024). Rather than summarizing prior work, the article expands each conceptual strand through historical development, theoretical debate, and critical comparison, exposing persistent tensions between accuracy, interpretability, adaptability, and computational feasibility. The methodology section articulates a text-based comparative research design that synthesizes cross-domain findings without relying on mathematical formalism or visual artifacts, thereby foregrounding epistemological assumptions and methodological limitations. Results are presented as interpretive patterns grounded in literature-based evidence, highlighting recurring phenomena such as sensitivity to hyperparameter tuning, dataset bias, and the contextual ambiguity of ground truth labels. The discussion extends these findings into a broader theoretical discourse on the future of anomaly detection research, arguing that progress depends less on novel architectures than on integrative evaluation philosophies and domain-aware learning paradigms. The article concludes by outlining a research agenda that prioritizes interpretability, cross-domain generalization, and ethical accountability as central criteria for next-generation anomaly detection systems.

Keywords: Anomaly detection, Machine learning, Industrial analytics, Network security, Fraud detection, Interpretability, Unsupervised learning.

INTRODUCTION

Anomaly detection occupies a paradoxical position within machine learning research, simultaneously perceived as a mature field with a long statistical heritage and as an open-ended frontier continually reshaped by new application domains and data modalities. At its core, anomaly detection concerns the identification of patterns, observations, or behaviors that deviate meaningfully from an established notion of normality, yet this deceptively simple definition conceals deep conceptual and practical challenges that have persisted since the earliest statistical studies of outliers (Simpson, 1951). The contemporary resurgence of interest in anomaly detection is inseparable from the proliferation of complex, high-dimensional, and continuously generated data across industrial automation, networked systems, financial platforms, and healthcare infrastructures, where anomalies often correspond to safety-critical faults, malicious attacks, or fraudulent behaviors (Darsh, 2021; Khan, 2024).

Historically, anomaly detection emerged from classical statistics, where deviations were interpreted primarily as noise or measurement error rather than as signals of intrinsic interest. Over time, this perspective shifted as

researchers recognized that rare events could carry disproportionate informational value, particularly in domains such as quality control, intrusion detection, and risk management (Rieck, 2009). This shift was accompanied by the gradual incorporation of machine learning techniques, which promised to model complex distributions and nonlinear relationships beyond the reach of traditional parametric methods (Schölkopf and Smola, 2002). However, the adoption of machine learning did not resolve the fundamental ambiguity of what constitutes an anomaly, as definitions of normality remain inherently domain-dependent and socially constructed (Palakurti, 2024).

The industrial domain offers a compelling illustration of this ambiguity. In modern manufacturing environments, sensor-rich systems generate vast streams of process data, within which anomalies may reflect equipment wear, calibration drift, operator intervention, or genuinely novel failure modes. The comparative study of machine learning approaches applied to industrial screw driving data demonstrates that algorithmic performance is tightly coupled to the semantics of the production process and the granularity of feature representation, challenging the assumption that superior models can be identified independently of context (West and Deuse, 2024). This finding resonates with earlier work in network intrusion detection, where the same algorithm may exhibit dramatically different behavior depending on traffic composition, attack prevalence, and labeling practices (Pranto et al., 2022; Rieck and Laskov, 2007).

Within the broader literature, anomaly detection methods are often categorized according to their learning paradigm, such as supervised, semi-supervised, or unsupervised. Supervised approaches rely on labeled examples of both normal and anomalous behavior, offering strong discriminative power but limited applicability due to the scarcity and evolving nature of anomalies (Shawe-Taylor and Cristianini, 2004). Semi-supervised methods typically model normal behavior exclusively, treating deviations as potential anomalies, a strategy that aligns well with novelty detection frameworks such as support vector data description (Schölkopf et al., 2001). Unsupervised techniques, including clustering and density estimation, dispense with labels altogether, yet face interpretive challenges when distinguishing rare but benign patterns from genuinely problematic events (Habeeb et al., 2019).

The rise of deep learning has further complicated this landscape. Deep architectures promise automatic feature learning and scalability to high-dimensional data, but their opacity raises concerns about interpretability and trust, particularly in safety-critical contexts (Ruff et al., 2021). Comparative analyses of deep learning architectures for anomalous message detection in aviation surveillance systems reveal that performance gains often come at the cost of increased complexity and reduced transparency, complicating deployment decisions (Karam et al., 2020). Similar tensions are evident in financial fraud detection, where regulatory requirements and ethical considerations demand explainable decisions even as fraud patterns grow more sophisticated (Pan, 2024).

Despite the breadth of existing research, a persistent literature gap lies in the lack of integrative frameworks that reconcile methodological diversity with domain-specific constraints. Much of the anomaly detection literature remains fragmented, with studies focused narrowly on particular datasets or algorithms, limiting the transferability of insights across domains (Steinbuss and Böhm, 2021). Moreover, comparative evaluations are often undermined by inconsistent preprocessing, hyperparameter tuning, and performance metrics, making it difficult to draw robust conclusions about algorithmic superiority (Soenen et al., 2021). The industrial case study provided by West and Deuse underscores this issue by showing how evaluation outcomes shift depending on feature selection strategies and operational definitions of anomalies (West and Deuse, 2024).

This article addresses these challenges by advancing a comprehensive, theory-driven synthesis of machine learning–based anomaly detection that foregrounds interpretability, robustness, and contextual awareness as central organizing principles. Rather than proposing a new algorithm, the study interrogates the epistemological assumptions underlying existing methods and examines how these assumptions manifest across industrial, network, financial, and cyber-physical applications. By integrating classical statistical perspectives with contemporary machine learning research, the article seeks to illuminate why anomaly detection resists standardization and how future research might navigate this complexity more effectively (Palakurti, 2024; Ruff et al., 2021).

The contribution of this work is threefold. First, it offers an extensive theoretical elaboration of anomaly detection concepts, situating modern machine learning approaches within a broader historical and philosophical context. Second, it provides a detailed methodological rationale for comparative, literature-based analysis as a means of synthesizing heterogeneous findings without oversimplification. Third, it develops a critical discussion that connects empirical observations, such as those reported in industrial screw driving data, to enduring debates about evaluation, generalization, and ethical responsibility in anomaly detection research (West and Deuse, 2024; Darsh, 2021). Through this integrative approach, the article aims to support both researchers and practitioners in navigating the complex trade-offs that define the field.

METHODOLOGY

The methodological approach adopted in this study is

deliberately interpretive, comparative, and textually grounded, reflecting the recognition that anomaly detection research spans diverse domains, data types, and evaluation cultures that resist reduction to a single experimental protocol. Rather than conducting new empirical experiments, the methodology synthesizes findings from a carefully selected corpus of peer-reviewed studies, doctoral theses, and technical reports, treating them as empirical observations embedded within distinct epistemic contexts (Rieck, 2009; Steinbuss and Böhm, 2021). This approach aligns with prior meta-analytical efforts in anomaly detection, which emphasize the importance of contextualizing algorithmic performance rather than abstracting it from its domain of application (Ruff et al., 2021).

The first methodological pillar involves comparative domain analysis. Studies from industrial manufacturing, network intrusion detection, financial fraud detection, healthcare Internet of Things systems, and large-scale data analytics were examined to identify recurring methodological patterns and divergences. The industrial domain, exemplified by machine learning applied to screw driving process data, serves as a focal point due to its rich sensor data and clear operational consequences of anomalies (West and Deuse, 2024). Network-based studies contribute insights into adversarial behavior and evolving threat models, highlighting the limitations of static anomaly definitions (Rieck and Laskov, 2008; Pranto et al., 2022). Financial and healthcare applications introduce regulatory and ethical dimensions that shape methodological choices, particularly with respect to interpretability and false positive tolerance (Pan, 2024; Khan, 2024).

The second pillar centers on algorithmic taxonomy and theoretical grounding. Methods are grouped according to learning paradigm and representational assumptions, including statistical outlier detection, distance-based and density-based clustering, kernel methods, and deep representation learning. For each category, the analysis traces historical development, from early vector space models and principal component classifiers to contemporary autoencoder variants and deep neural architectures (Salton et al., 1975; Shyu et al., 2003; Shin and Kim, 2020). This historical perspective enables a critical assessment of how foundational assumptions about data structure and similarity continue to influence modern implementations (Schölkopf et al., 1999).

A third methodological component involves evaluation philosophy. Rather than privileging quantitative metrics, the study examines how evaluation criteria are framed and justified within each domain. This includes analysis of ground truth construction, performance metrics, and the role of hyperparameter tuning in comparative studies (Soenen et al., 2021). The industrial study by West and Deuse is particularly instructive, as it demonstrates how evaluation outcomes depend on operational definitions of anomalies and the alignment between model objectives and process semantics (West and Deuse, 2024). Similar concerns are evident in benchmarking efforts using synthetic data, where realism and representativeness remain contested (Steinbuss and Böhm, 2021).

The methodology also explicitly acknowledges its limitations. By relying on published literature, the analysis is constrained by reporting biases and the selective emphasis of individual studies. Furthermore, the absence of new empirical data precludes direct validation of synthesized claims. However, this limitation is offset by the depth of theoretical elaboration and cross-domain comparison, which allows for the identification of structural patterns that may not be visible within isolated empirical studies (Palakurti, 2024). Through this methodological design, the study aims to generate insights that are robust to dataset-specific idiosyncrasies while remaining sensitive to domain-specific realities.

RESULTS

The results of this integrative analysis are presented as interpretive patterns rather than numerical outcomes, reflecting the study's emphasis on conceptual coherence and cross-domain insight. One prominent pattern concerns the persistent dependence of anomaly detection performance on feature representation and preprocessing choices. Across industrial, network, and financial domains, studies consistently report that algorithmic differences often pale in comparison to the impact of feature engineering strategies (West and Deuse, 2024; Pranto et al., 2022). In industrial screw driving data, variations in torque, angle, and temporal aggregation significantly alter model sensitivity, underscoring that anomalies are as much a function of representation as of algorithmic capacity (West and Deuse, 2024).

A second recurring result relates to the instability of comparative rankings among algorithms. Benchmarking studies reveal that no single method consistently outperforms others across datasets and evaluation settings, a finding echoed in both unsupervised outlier detection benchmarks and applied intrusion detection research (Steinbuss and Böhm, 2021; Darsh, 2021). This instability is exacerbated by hyperparameter tuning practices, which can dramatically reshape performance landscapes and complicate fair comparison (Soenen et al., 2021). The implication is that reported superiority of specific methods should be interpreted cautiously and within narrowly defined contexts.

Interpretability emerges as a third key result, particularly in domains where anomalies trigger high-stakes decisions. While deep learning methods often achieve competitive detection rates, their opaque internal representations hinder post hoc explanation and user trust (Ruff et al., 2021). In contrast, kernel-based and statistical approaches, though sometimes less flexible, offer clearer conceptual links between input features and anomaly

scores, facilitating domain expert validation (Schölkopf et al., 2001). The industrial case study illustrates that practitioners frequently favor models whose behavior can be reconciled with process knowledge, even at the expense of marginal performance gains (West and Deuse, 2024).

Another salient result concerns the contextual ambiguity of anomalies. Across domains, anomalies are rarely binary phenomena; instead, they occupy a spectrum of deviation whose significance depends on operational thresholds, cost structures, and risk tolerance (Pan, 2024; Khan, 2024). Network intrusion detection studies highlight how benign but rare traffic patterns can be misclassified as attacks, inflating false positive rates and eroding system credibility (Rieck and Laskov, 2006). Similarly, in industrial settings, process variations that deviate statistically from historical norms may reflect acceptable adaptations rather than faults (West and Deuse, 2024).

Collectively, these results suggest that anomaly detection effectiveness cannot be disentangled from domain knowledge, evaluation philosophy, and human interpretive frameworks. Rather than converging toward a universal solution, the field continues to diversify, with methods evolving in response to domain-specific demands and constraints (Palakurti, 2024).

DISCUSSION

The findings synthesized in this study invite a deeper theoretical reflection on why anomaly detection remains such a contested and dynamic area of machine learning research. At a foundational level, the concept of an anomaly challenges classical statistical assumptions by foregrounding rarity and deviation as objects of interest rather than noise to be eliminated (Simpson, 1951). This inversion complicates learning objectives, as models must balance sensitivity to rare events against robustness to benign variation, a tension that manifests differently across domains (Ruff et al., 2021).

From a historical perspective, the evolution of anomaly detection mirrors broader shifts in machine learning, from rule-based systems and linear models toward kernel methods and deep architectures. Each wave of innovation has expanded representational capacity while introducing new interpretive challenges. The support vector framework for novelty detection exemplifies this trade-off, offering elegant theoretical guarantees at the cost of parameter sensitivity and kernel selection complexity (Schölkopf et al., 1999). Deep learning extends this trajectory, enabling hierarchical feature learning but obscuring the relationship between input data and anomaly decisions (Karam et al., 2020).

The industrial study of screw driving data provides a concrete lens through which to examine these dynamics. The comparative performance of machine learning approaches in this context underscores that anomalies are inseparable from process semantics and operational goals (West and Deuse, 2024). Unlike network intrusion detection, where anomalies often correspond to adversarial intent, industrial anomalies may reflect gradual degradation, operator adaptation, or benign process drift. This multiplicity of meanings complicates labeling and evaluation, reinforcing the need for domain-aware modeling strategies.

A critical debate emerging from the literature concerns the role of benchmarking and standardization. While benchmarks promise comparability and reproducibility, they risk oversimplifying complex phenomena and privileging easily quantifiable metrics over contextual relevance (Steinbuss and Böhm, 2021). The sensitivity of anomaly detection outcomes to hyperparameter tuning further undermines the notion of definitive rankings among methods (Soenen et al., 2021). These issues suggest that future research should prioritize transparent reporting and sensitivity analysis over claims of universal superiority.

Ethical and societal considerations also loom large in the discussion. In financial and healthcare applications, false positives and false negatives carry asymmetric costs that extend beyond technical performance, implicating fairness, accountability, and trust (Pan, 2024; Khan, 2024). The opacity of complex models raises questions about responsibility and oversight, particularly when automated decisions affect individuals or critical infrastructure. Interpretability, therefore, is not merely a technical desideratum but a normative requirement shaped by institutional and cultural contexts.

Looking forward, the discussion points toward a research agenda that emphasizes integrative evaluation, hybrid modeling approaches, and closer collaboration between machine learning researchers and domain experts. Rather than seeking a single best algorithm, progress may depend on developing flexible frameworks that adapt to evolving definitions of normality and anomaly (Palakurti, 2024). The industrial insights articulated by West and Deuse exemplify how such collaboration can ground algorithmic choices in operational reality, enhancing both performance and acceptance (West and Deuse, 2024).

CONCLUSION

This article has presented an extensive, theory-driven examination of machine learning–based anomaly detection across multiple application domains, integrating historical perspectives, methodological analysis, and critical discussion. By synthesizing diverse strands of research, the study demonstrates that anomaly detection is best understood not as a problem with a universal solution but as a family of context-dependent challenges shaped by data characteristics, domain semantics, and human judgment. The industrial case of screw driving data illustrates how comparative algorithmic performance is inseparable from feature representation and evaluation

philosophy, reinforcing the centrality of domain knowledge (West and Deuse, 2024). Ultimately, the future of anomaly detection research lies in embracing this complexity and developing methods and evaluation practices that balance accuracy, interpretability, and ethical responsibility.

REFERENCES

1. Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., and Müller, K.-R. (2021). A unifying review of deep and shallow anomaly detection. Institute of Electrical and Electronics Engineers.
2. Rieck, K., and Laskov, P. (2007). Language models for detection of unknown attacks in network traffic. Journal in Computer Virology, 2(4), 243–256.
3. Pan, E. (2024). Machine learning in financial transaction fraud detection and prevention. ResearchGate.
4. Steinbuss, G., and Böhm, K. (2021). Benchmarking unsupervised outlier detection with realistic synthetic data. ACM Transactions on Knowledge Discovery from Data, 15(4), 1–20.
5. Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural Computation, 13(7), 1443–1471.
6. West, N., and Deuse, J. (2024). A comparative study of machine learning approaches for anomaly detection in industrial screw driving data. Proceedings of the 57th Hawaii International Conference on System Sciences.
7. Darsh, P. (2021). Performance analysis of network anomaly detection systems in consumer networks. IEEE Access.
8. Pranto, M. B., et al. (2022). Performance of machine learning techniques in anomaly detection with basic feature selection strategy: A network intrusion detection system. Journal of Advances in Information Technology, 13(1).
9. Karam, R., et al. (2020). A comparative study of deep learning architectures for detection of anomalous ADS-B messages. IEEE.
10. Khan, M. M. (2024). Anomaly detection in IoT-based healthcare: machine learning for enhanced security. Scientific Reports.
11. Palakurti, N. R. (2024). Challenges and future directions in anomaly detection. ResearchGate.
12. Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., and Platt, J. C. (1999). Support vector method for novelty detection. Neural Information Processing Systems.
13. Soenen, J., Van Wolputte, E., Perini, L., Vercruyssen, V., Meert, W., Davis, J., and Blockeel, H. (2021). The effect of hyperparameter tuning on the comparative evaluation of unsupervised anomaly detection methods. Knowledge Discovery and Data Mining Workshop on Outlier Detection and Description.
14. Habeeb, R. A. A., et al. (2019). Clustering-based real-time anomaly detection—A breakthrough in big data technologies. ResearchGate.
15. Simpson, E. H. (1951). The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society: Series B, 13(2), 238–241.
16. Rieck, K. (2009). Machine learning for application-layer intrusion detection. Ph.D. thesis, Berlin Institute of Technology.
17. Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613–620.
18. Shin, S. Y., and Kim, H.-j. (2020). Extended autoencoder for novelty detection with reconstruction along projection pathway. Applied Sciences, 10(13), 4497.
19. Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., and Chang, L. (2003). A novel anomaly detection scheme based on principal component classifier. Technical report, University of Miami.
20. Rifkin, R. M., and Lippert, R. A. (2007). Value regularization and Fenchel duality. Journal of Machine Learning Research, 8, 441–479.
21. Shawe-Taylor, J., and Cristianini, N. (2004). Kernel methods for pattern analysis. Cambridge University Press.
22. Sonnenburg, S. (2008). Machine learning for genomic sequence analysis. Ph.D. thesis, Fraunhofer Institute FIRST.
23. Schölkopf, B., and Smola, A. (2002). Learning with kernels. MIT Press.
24. Rieck, K., and Laskov, P. (2006). Detecting unknown network attacks using language models. Detection of Intrusions and Malware, and Vulnerability Assessment.
25. Rieck, K., and Laskov, P. (2008). Linear-time computation of similarity measures for sequential data. Journal of Machine Learning Research, 9, 23–48.
26. Elki: A large open-source library for data analysis. Schubert, E., and Zimek, A. (2019). CoRR.