Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 04| Issue- 09| 2025

Research Article **Received:** 15-08-2025| **Accepted:** 05-09-2025 | **Published:** 30-09-2025

Intelligent Anomaly Detection in Real-Time Big Data Engineering

Maheshkumar Mayilsamy

Zillow Group Inc., USA

Abstract: Intelligent anomaly detection in real-time big data engineering represents a critical advancement in how organizations identify and respond to unexpected patterns across high-velocity data streams. This article explores the evolution from retrospective analysis to proactive defense mechanisms enabled by the convergence of stream processing technologies and advanced machine learning techniques. The article encompasses the fundamental categorization of anomalies in streaming environments—point, contextual, and collective—alongside complementary detection approaches that range from rule-based systems to sophisticated deep learning architectures. Key challenges, including data velocity management, concept drift adaptation, alert fatigue mitigation, and system scalability, are addressed within the context of production implementations. The article extends to the evolving ecosystem of stream processing engines, specialized detection libraries, and monitoring infrastructure supporting these capabilities. Through sector-specific case studies spanning financial services, manufacturing, cybersecurity, and healthcare, the article demonstrates the transformative impact of real-time anomaly detection across industries. Finally, emerging trajectories, including the integration of explainable AI, federated learning approaches, edge computing deployment, and generative AI applications, are explored as future directions that will reshape the field.

Keywords: Real-time anomaly detection, Stream processing, Machine learning, Predictive maintenance, Explainable AI.

INTRODUCTION

In today's digital economy, detecting anomalous trends within data streams in real-time has become essential. The combination of high-velocity stream processing with state-of-the-art machine learning has transformed organizational responses to anomalies. Retrospective analysis approaches have been superseded by advanced systems capable of detecting behavioral deviations within milliseconds, transitioning detection from a retrospective process to a proactive defense mechanism.

The transportation sector provides substantial evidence of these benefits. Traffic management systems enhanced with real-time anomaly detection have significantly reduced congestion incidents while expediting emergency response throughout metropolitan areas. These systems continuously analyze inputs from roadside sensors, surveillance cameras, and connected vehicle networks to identify unusual traffic patterns that may indicate accidents or infrastructure failures (Samaila, Y. A. *et al.*, 2025). The benefits extend beyond basic incident detection—municipal authorities throughout Europe and Asia report significant improvements in resource deployment efficiency and public satisfaction following implementation of AI-enhanced monitoring across smart city initiatives.

Performance requirements for streaming architectures continue to increase exponentially, particularly within financial services, telecommunications, and digital commerce, where substantial transaction volumes require analysis with minimal latency. Recent academic research has established the fundamental requirements for scalable distributed stream processing systems capable of handling these workloads. Cherniack et al. highlight the architectural challenges of developing distributed data stream management systems that can process continuous queries over multiple high-volume, time-varying data streams with stringent performance constraints (Cherniack, M. *et al.*, 2003). Their work identifies core design patterns necessary for stream-oriented systems to maintain performance at scale, while enabling the complex analytical capabilities required for effective anomaly detection.

This capability is particularly valuable in scenarios with subtle indicators of anomalous behavior that may signal potential fraud, service quality degradation, or emergent opportunities requiring immediate attention. The integration of sophisticated learning models with streaming architectures has facilitated the transition from elementary rule-based detection processes to more nuanced context-aware systems where models adapt to changing data patterns while reducing false alerts.

Implementation challenges remain significant despite these advancements. Organizations must address sophisticated architectural requirements, including robust reliability, fault tolerance, and integration with diverse data sources without compromising performance (Cherniack, M. *et al.*,

*Corresponding Author: Maheshkumar Mayilsamy

DOI- https://doi.org/10.5281/zenodo.17238490

2003). Nevertheless, the operational utility of real-time anomaly detection provides compelling justification for investments across healthcare, manufacturing, and other sectors where implementation continues to refine not only the methodological and architectural approaches involved in intelligent real-time responses, but also the strategic application of these techniques for identifying outliers in operational data streams.

WHY ANOMALY DETECTION MATTERS IN REAL-TIME SYSTEMS

The operational context determines the relevance of real-time anomaly detection. Financial institutions implement these systems to detect fraudulent transactions before fund transfers are completed. Security teams rely on them to identify potential network breaches prior to data exfiltration. Manufacturing operations employ detection techniques to anticipate equipment failures before costly production interruptions occur.

The business impact becomes evident when examining quantifiable results across sectors. Within industrial environments, implementing real-time anomaly detection for equipment monitoring has fundamentally transformed maintenance approaches from reactive to predictive strategies. Manufacturing facilities utilizing machine learning detection methodologies demonstrate significantly better outcomes compared to conventional threshold-based systems. Research conducted in semiconductor production environments demonstrates how multiple classifier methodologies substantially enhance maintenance decision accuracy by combining varied algorithmic strengths to detect subtle equipment degradation indicators before failures manifest (Susto, G. A. *et al.*, 2014). These platforms continuously analyze complex sensor data feeds from production machinery, enabling maintenance teams to transition toward condition-based maintenance routines rather than calendar-based maintenance intervals, thereby maximizing equipment uptime and optimizing maintenance resource utilization.

The migration of analytical systems to streaming models represents a fundamental shift in operational methodology within critical analytics. In their seminal work, Stonebraker et al. established the definitive requirements for real-time stream processing systems, identifying eight essential criteria that any stream processing application must satisfy to deliver genuine real-time capabilities (Susto, G. A. *et al.*, 2014). These requirements include processing data elements individually using constrained computing resources, delivering guaranteed results within time constraints, handling stream inconsistencies, generating predictable outcomes, and seamlessly integrating historical data with current streams. Their research provides the theoretical foundation underlying modern stream processing architectures, guiding implementation decisions for systems operating under strict latency constraints.

Organizations implementing systems meeting these specifications report substantial improvements in operational responsiveness, with capabilities to detect and address anomalies within milliseconds rather than hours, becoming significant competitive advantages. The convergence of appropriate stream processing architectures with increasingly sophisticated learning models has created detection capabilities identifying complex anomaly patterns across diverse data sources while maintaining sub-second response times essential for meaningful intervention.

Publisher: SARC Publisher

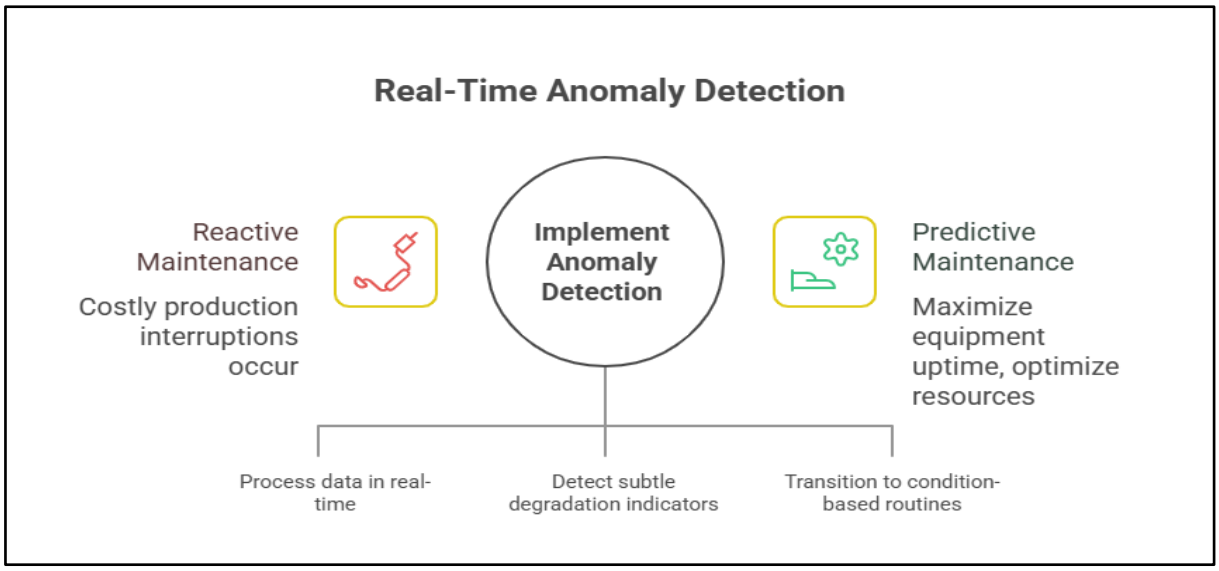

Fig 1: Real-Time Anomaly Detection (Susto, G. A. *et al.*, 2014; Stonebraker, M. *et al.*, 2005)

TYPES OF ANOMALIES IN STREAMING DATA

The velocity, volume, and variability of streaming data present distinct detection challenges. Analysts typically categorize streaming anomalies into three primary classifications.

Point Anomalies manifest as individual data elements that significantly deviate from established norms. Examples include credit card purchases substantially exceeding typical spending behaviors or sudden traffic spikes from individual network devices. While often the simplest to detect, these anomalies frequently generate excessive false positives without proper contextual frameworks. According to comprehensive research examining detection methodologies, point anomalies can be addressed through various techniques, including classification-based, nearest neighbor-based, and statistical methods, each with distinct computational requirements affecting their suitability for streaming environments (Chandola, V. *et al.*, 2009). The primary challenge involves balancing detection accuracy against processing efficiency, as methods effective for static datasets typically require substantial adaptation for continuous data streams. Contemporary detection platforms implement optimized algorithmic variants capable of processing individual data points upon arrival while maintaining sufficient statistical context for accurate outlier identification.

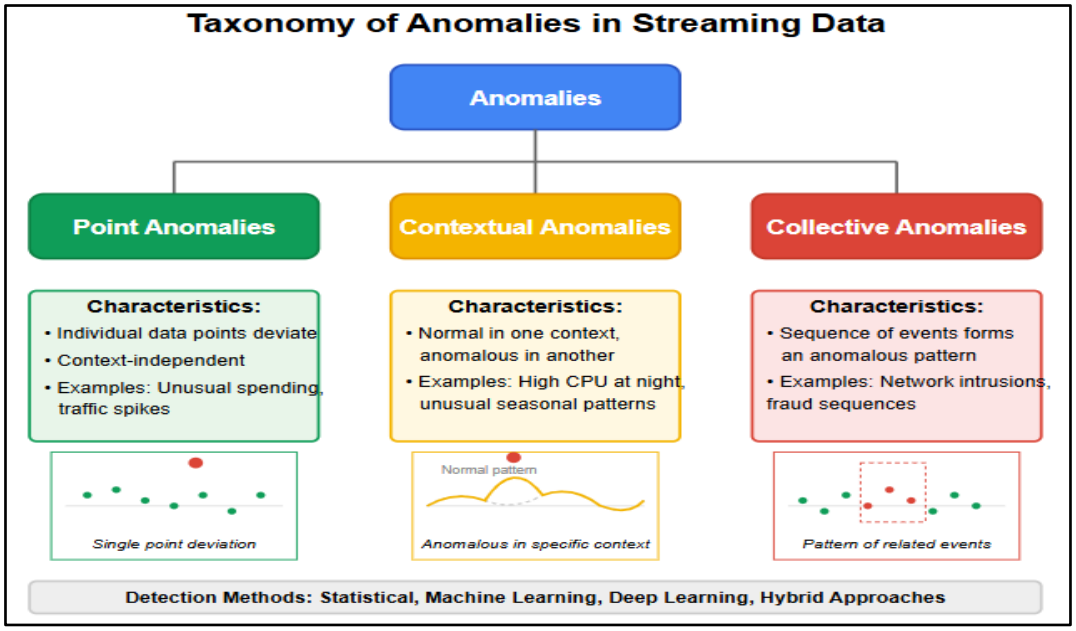

Fig 2: Taxonomy of Anomalies in Streaming Data (Chandola, V. *et al.*, 2009; Gupta, M. *et al.*, 2013)

Publisher: SARC Publisher

Contextual Anomalies emerge when data appears irregular only under specific circumstances. For instance, elevated CPU usage during scheduled maintenance represents normal operations, while identical metrics during off-hours may indicate security compromises. Contextual variables, including time periods, geographic locations, or user segments, become essential factors for accurate identification. Research focused specifically on temporal data anomaly detection establishes that contextual anomalies require algorithms capable of distinguishing two key components: contextual attributes defining situational parameters and behavioral attributes defining anomalous actions within those parameters (Gupta, M. *et al.*, 2013). Streaming environments necessitate specialized techniques for efficiently maintaining and updating contextual profiles while processing high-velocity data streams. Systems incorporating these approaches typically achieve substantially improved precision compared to context-agnostic methods, particularly within domains characterized by strong seasonal or cyclical patterns.

Collective Anomalies occur when sequences of events, each potentially normal independently, form irregular patterns when examined together. Sophisticated cyber attacks frequently manifest as collective anomalies, with attackers orchestrating multiple actions that individually remain below detection thresholds. Identifying these patterns requires analyzing temporal relationships and sequential patterns throughout data streams. Detecting collective anomalies within temporal data represents one of the most challenging aspects of stream processing, requiring specialized techniques to identify irregular patterns across observation sequences (Gupta, M. *et al.*, 2013). These approaches encompass point-based techniques adapted for temporal windows, model-based methods capturing expected sequential behaviors, and pattern-based strategies identifying unusual event combinations. Effectiveness varies significantly based on application domains and data characteristics, with hybrid approaches frequently delivering optimal performance across diverse streaming scenarios.

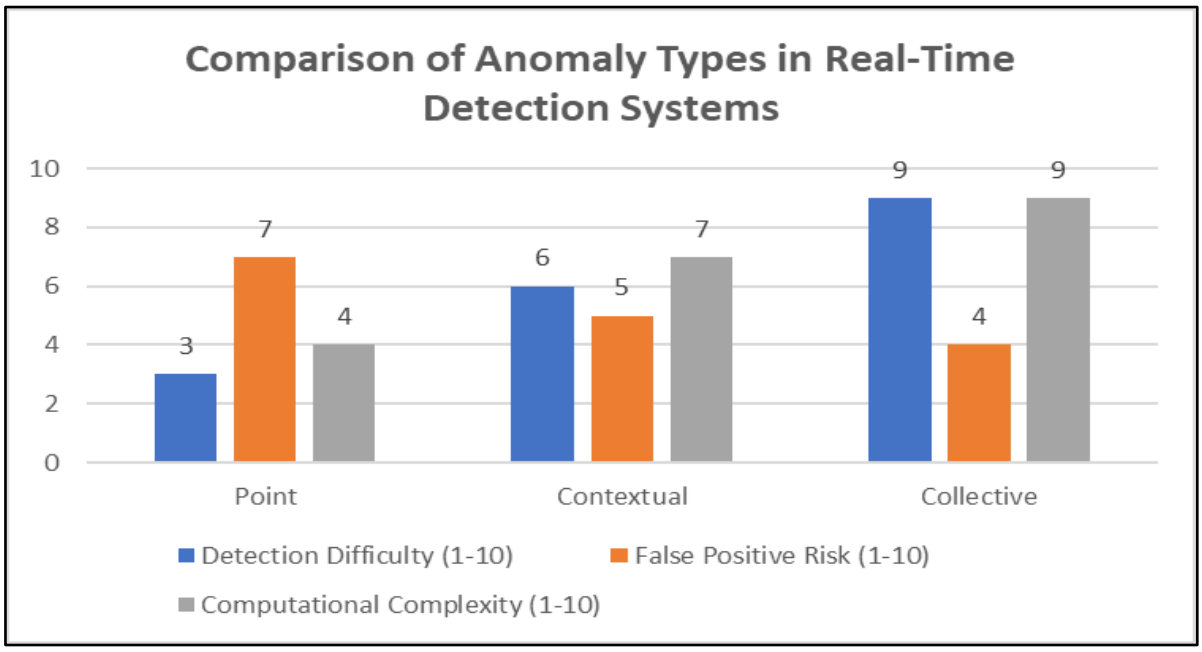

Fig 3: Characteristics of Streaming Data Anomaly Categories (Chandola, V. *et al.*, 2009; Gupta, M. *et al.*, 2013)

TECHNIQUES FOR INTELLIGENT ANOMALY DETECTION

Modern anomaly detection systems employ multiple complementary methodologies to achieve optimal detection rates while minimizing false alerts:

Rule-Based Detection forms the initial defense layer, applying predefined thresholds and business logic to identify obvious violations. Though limited in detecting novel anomalies, these systems deliver immediate responses with transparent decision rationales. Financial systems typically implement rules such as "flag transactions exceeding specific thresholds" or "block purchases from restricted regions" as foundational protections. Despite their apparent simplicity, rule-

Publisher: SARC Publisher

based approaches remain prevalent within production environments due to their interpretability, computational efficiency, and capacity for directly encoding domain expertise within detection frameworks. Research examining real-world implementations across various sectors reveals that organizations typically maintain extensive rule libraries developed through iterative refinement processes, with mature systems containing hundreds or thousands of specialized detection rules (Ahmad, S. *et al.,* 2017). System effectiveness depends heavily upon regular maintenance and calibration, as static rules quickly become outdated as normal patterns evolve.

Statistical Methods expand beyond basic rules by establishing dynamic behavioral baselines. Techniques including moving averages, exponential smoothing, and statistical process control adapt to gradual pattern shifts. These methods excel at identifying subtle deviations potentially missed by rule-based detection. Comprehensive evaluations of statistical approaches for streaming data demonstrate that properly calibrated statistical models achieve detection rates comparable with more complex methodologies while maintaining the computational efficiency required for high-throughput environments (Ahmad, S. *et al.,* 2017). Recent innovations include adaptive statistical models that automatically adjust sensitivity based on observed false positive rates and multivariate statistical techniques that consider correlations between various metrics when establishing normal behavior baselines.

Machine Learning Approaches transcend rigid statistical models by discovering complex patterns directly from data. Unsupervised techniques, including DBSCAN and isolation forests, identify outliers without requiring labeled anomaly examples. Supervised methods leverage historical incidents to train classifiers recognizing similar future events. Semi-supervised approaches learn representations of normal data, flagging instances that deviate from learned patterns. Comparative analysis examining machine learning algorithms for streaming anomaly detection reveals that ensemble methods typically outperform individual algorithms, with random forest and gradient boosting approaches demonstrating particularly strong results across diverse datasets (Pang, G. *et al.,* 2021). These techniques excel at discovering non-linear relationships and complex feature interactions potentially missed by rule-based or statistical approaches.

Deep Learning for Streams addresses temporal characteristics of streaming data through specialized architectures. Recurrent Neural Networks and Long Short-Term Memory networks maintain internal state representations capturing sequential dependencies. Transformer models with attention mechanisms have achieved significant results in analyzing complex time series, recognizing subtle anomalous patterns across varying time scales. Empirical studies evaluating deep learning architectures for streaming anomaly detection demonstrate that these approaches significantly outperform traditional methods, particularly for high-dimensional data with complex temporal dependencies (Pang, G. *et al.,* 2021). However, performance advantages come with increased computational requirements and reduced interpretability, creating practical deployment challenges for many organizations. Recent research focuses on developing more efficient neural architectures specifically optimized for streaming contexts, enabling deployment on edge devices while reducing latency.

Hybrid Approaches combine multiple techniques, leveraging complementary strengths. Systems might employ rules addressing known threats, statistical methods detecting gradual deviations, and deep learning identifying novel patterns. This multi-layered architecture provides both processing speed and analytical sophistication, with simpler models handling obvious cases while complex models focus on edge scenarios. Analysis examining production anomaly detection systems reveals that successful implementations predominantly adopt hybrid approaches, employing tiered detection strategies that balance computational efficiency against detection accuracy (Ahmad, S. *et al.,* 2017). Contemporary frameworks typically implement "detection pipelines" where incoming data points undergo initial screening through lightweight rule-based and statistical methods, with only ambiguous cases escalated toward sophisticated machine learning or deep learning models. This approach optimizes resource utilization while maintaining robust detection capabilities across diverse anomaly categories.

Publisher: SARC Publisher

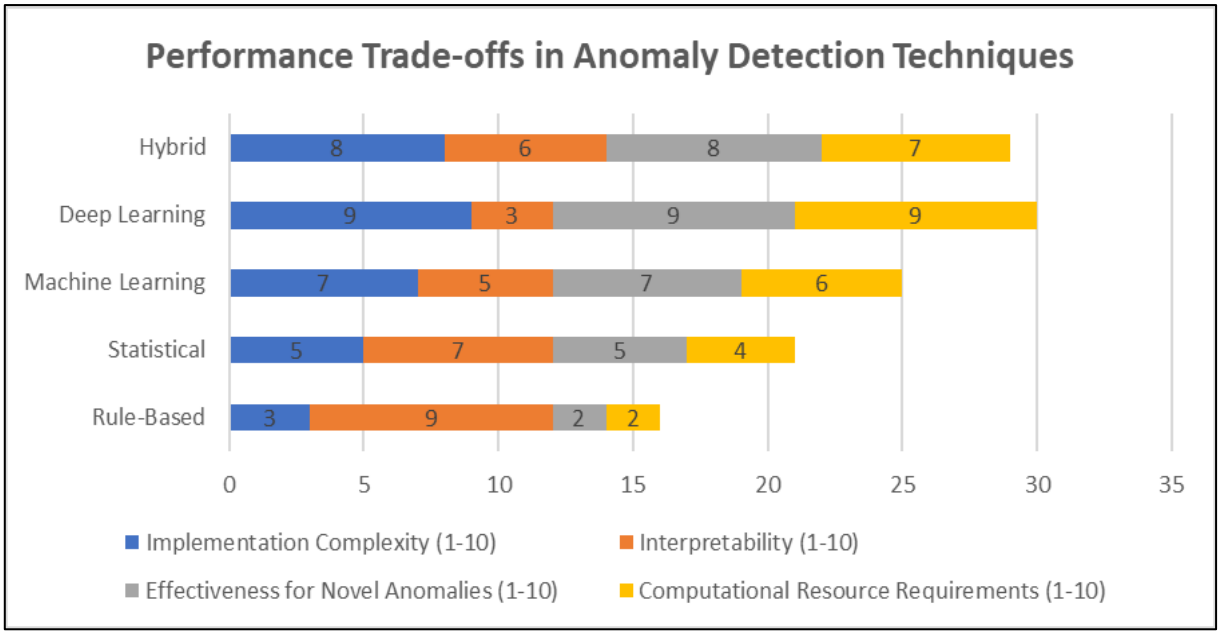

Fig 4: Comparative Analysis of Real-Time Anomaly Detection Methods (Ahmad, S. *et al.*, 2017; Pang, G. *et al.*, 2021)

CHALLENGES IN REAL-TIME ANOMALY DETECTION

Despite significant progress, several formidable obstacles persist when implementing effective real-time anomaly detection:

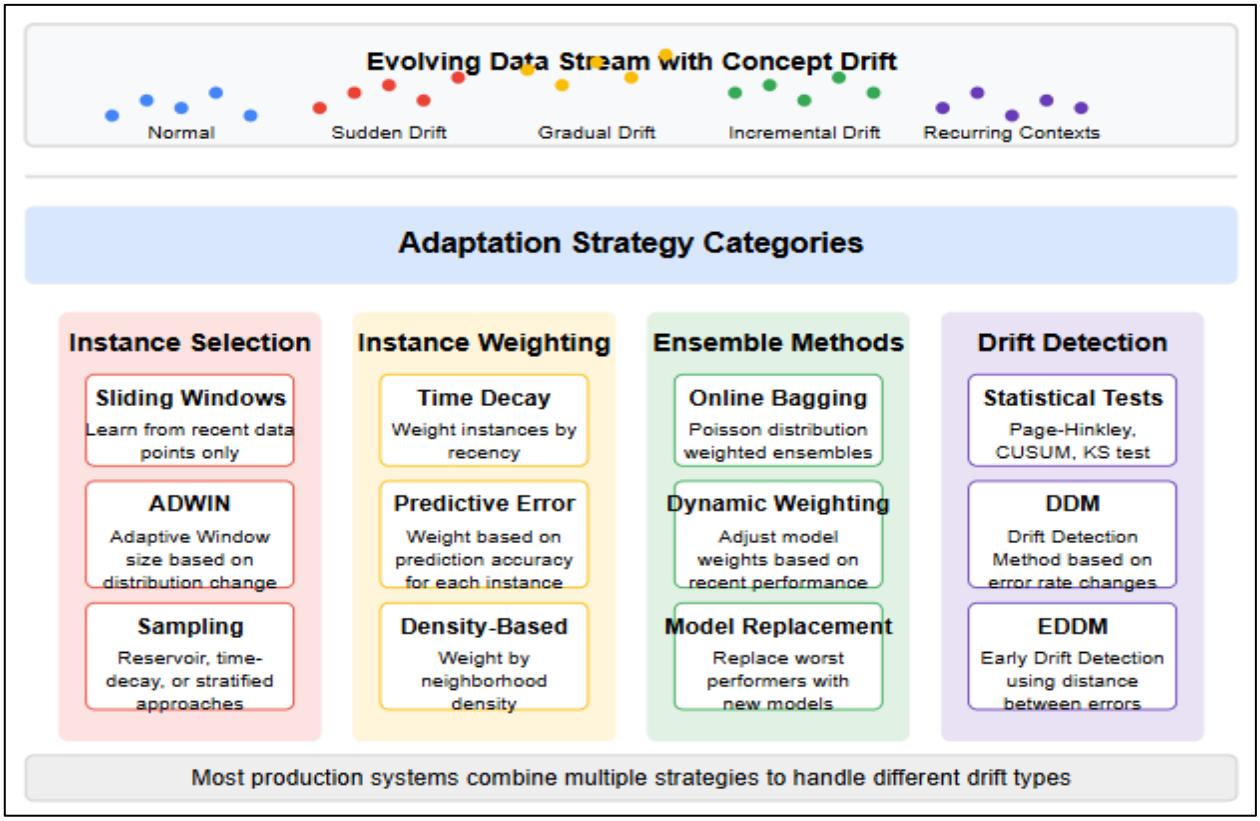

Fig 5: Concept Drift Adaptation Strategies (Flouris, I. *et al.*, 2017; Gama, J. *et al.*, 2014)

Data Velocity and Volume generate substantial processing constraints. Effective platforms must analyze millions of discrete events every second while maintaining sub-second latency. This necessitates distributed computing architectures and highly optimized algorithms operating within stringent resource limitations. An investigation into complex event processing frameworks reveals

Publisher: SARC Publisher

significant technical challenges in supporting performance at scale, particularly in executing advanced detection logic over high-velocity data streams (Flouris, I. *et al.,* 2017). The transition to processing architectures suitable for supporting real-time analysis requires careful consideration of architectural design, infrastructure requirements, and methodological approaches. Contemporary frameworks must incorporate techniques including operator parallelization, strategic load shedding, and approximate computing to maintain acceptable responsiveness under extreme data burdens. Such systems demand meticulous memory management optimization, as the stateful nature inherent in numerous anomaly detection algorithms frequently conflicts with distributed processing requirements essential for maintaining throughput performance.

Concept Drift occurs when underlying data patterns change naturally, resulting in progressive decline in detection accuracy. This is caused by seasonal variations, changes in user behavior, and business dynamics. This issue can be partially addressed through adaptive detection mechanisms that regularly update parameters using the most recent observations. Detailed examinations of concept drift adaptation techniques have categorized approaches into distinct families encompassing instance selection, instance weighting, ensemble learning, and explicit drift detection methodologies (Gama, J. *et al.,* 2014). Each strategy represents different trade-offs between adaptation responsiveness, computational requirements, and stability under various drift manifestations. Production streaming environments typically encounter multiple drift types simultaneously—including sudden shifts, gradual transitions, incremental modifications, and recurring contextual patterns. This complexity demands hybrid adaptation frameworks responding appropriately across different drift categories while preserving consistent detection performance. Organizations deploying such systems must carefully balance competing demands between rapid adaptation toward legitimate pattern evolution and maintaining stability against transient anomalies.

The False Positive Paradox embodies fundamental tension within anomaly detection system architecture. Highly sensitive detection mechanisms frequently overwhelm operators with excessive alerts, fostering "alert fatigue" and causing critical signals to be overlooked. Conversely, systems optimized for precision often miss subtle anomalous patterns. Determining optimal operational thresholds requires careful examination of specific application domains and associated error costs. Extensive research examining complex event processing platforms identifies alert management among the most significant operational challenges within production environments (Flouris, I. *et al.,* 2017).. The inherently subjective nature of anomaly definitions across domains complicates ground truth establishment, hampering efforts toward optimizing detection thresholds. This dilemma becomes further complicated through asymmetric costs associated with different error categories across application domains—certain contexts prioritize minimizing false negatives (such as missing security breaches), while others emphasize reducing false positives that generate substantial operational overhead undermining system adoption.

Scalability and Fault Tolerance emerge as critical considerations within production deployments. Detection systems must be designed for horizontal scaling to accommodate increasing data volumes while maintaining operational integrity despite individual component failures. The required resilience necessitates advanced stream processing architectures with built-in redundancy and failover capabilities. The distributed nature characterizing modern stream processing frameworks introduces complex state management challenges, particularly affecting algorithms preserving historical context (Flouris, I. *et al.,* 2017).. These platforms must implement specialized techniques maintaining processing guarantees during failure scenarios, including distributed checkpointing, state replication, and recovery protocols. Concept drift adaptation mechanisms essential for maintaining accuracy further complicate these requirements by introducing additional state information requiring consistent maintenance across distributed processing nodes (Gama, J. *et al.,* 2014). Organizations implementing large-scale deployments report that achieving continuous operational reliability frequently demands sophisticated orchestration layers beyond the capabilities provided through standard stream processing frameworks.

Publisher: SARC Publisher

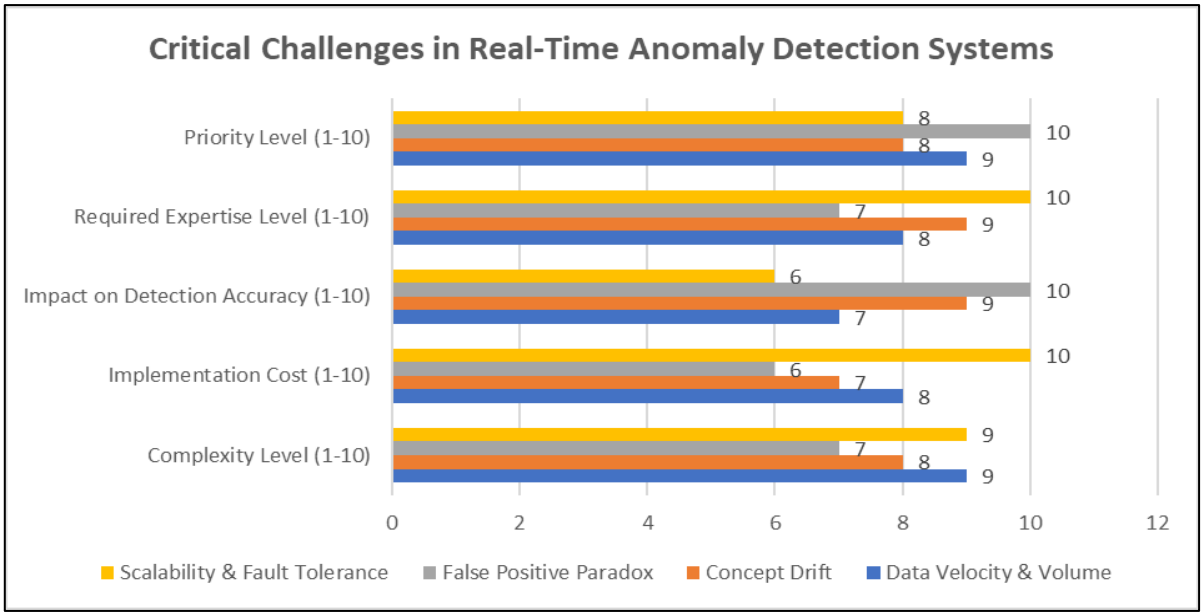

Fig 6: Implementation Barriers for Effective Streaming Anomaly Detection (Flouris, I. *et al.*, 2017; Gama, J. *et al.*, 2014).

REAL-TIME TOOLS AND PLATFORMS

The technology ecosystem supporting real-time anomaly detection continues to undergo rapid evolution:

Stream Processing Engines constitute foundational components within real-time data pipelines. Apache Flink provides scaling with exactly-once semantics and millisecond latency. Apache Kafka Streams delivers lightweight stream processing that integrates seamlessly with Kafka messaging infrastructure. Spark Structured Streaming extends well-known Spark APIs into the streaming context via micro-batch processing. Rigorous benchmarking across these stream processing engines reveals significant performance variations spanning operational metrics, including throughput capacity, processing latency, and resource utilization patterns (Henning, S., & Hasselbring, W. 2021). These evaluations demonstrate that no single platform excels universally, necessitating careful selection based upon specific application requirements. Systems prioritizing minimal detection latency frequently benefit from Flink's sophisticated event-time processing capabilities, while deployments facing extreme scalability demands might favor Kafka Streams' lightweight deployment model. Benchmarking methodologies themselves have evolved substantially, with contemporary approaches emphasizing realistic microservice deployment scenarios rather than isolated performance testing. This methodological shift acknowledges that real-world anomaly detection systems typically function as components within broader architectural ecosystems where integration characteristics significantly impact overall system performance.

Specialized Anomaly Detection Libraries accelerate implementation through optimized algorithmic collections. PyOD delivers over thirty anomaly detection algorithms through consistent API interfaces. River (formerly Creme) concentrates on online learning algorithms that update incrementally with each incoming data point. Numenta's HTM (Hierarchical Temporal Memory) implementations demonstrate particular strength in identifying anomalies within complex time-series data. Detailed analysis examining specialized libraries, including PyOD, confirms their effectiveness in standardizing implementation approaches across diverse detection techniques (Zhao, Y. *et al.*, 2019). PyOD specifically implements numerous outlier detection algorithms spanning classical statistical approaches through recent neural network methodologies, all accessible via consistent scikit-learn compatible interfaces. This standardization substantially reduces implementation complexity while enabling straightforward performance comparisons between alternative approaches. Extensive benchmarking across multiple real-world datasets confirms that no single algorithm consistently outperforms alternatives across all scenarios, reinforcing the value derived from libraries providing multiple detection methodologies with minimal integration overhead. Within streaming contexts, these libraries typically require adaptation for handling

Publisher: SARC Publisher

continuous data flows, with frameworks including River addressing this requirement through native online learning implementations.

Monitoring and Alerting Infrastructure completes feedback loops between detection and response. Prometheus gathers and preserves time-series metrics, providing powerful query language capabilities for defining alert conditions. Grafana provides visualization dashboards that indicate anomalies. The ELK stack (Elasticsearch, Logstash, Kibana) provides capabilities to search and analyze anomalous patterns in log data. Evaluations examining monitoring architectures within microservice environments highlight substantial operational challenges in maintaining visibility across distributed streaming systems (Henning, S., & Hasselbring, W. 2021). Traditional monitoring approaches frequently struggle to address dynamic characteristics inherent within containerized deployments and extreme cardinality within modern observability data. Prometheus has emerged as a preferred solution for metric-based monitoring through its pull-based collection model and dimensional data representation, aligning effectively with dynamic infrastructure patterns. Visualization platforms, including Grafana, have evolved to support specialized anomaly visualization capabilities encompassing heatmaps, outlier highlighting, and dedicated anomaly panels. Integration between detection systems and monitoring platforms remains challenging, particularly regarding complex anomalies defying representation through traditional alerting rules, driving interest toward sophisticated notification systems providing contextual information alongside suggested remediation approaches.

REAL-WORLD USE CASES

Organizations spanning diverse sectors have implemented intelligent anomaly detection systems, yielding substantial results:

Within Financial Services, major payment processors analyze hundreds of billions of annual transactions, evaluating each against hundreds of potential fraud indicators within milliseconds. These systems have significantly reduced fraud losses while minimizing false declines affecting legitimate customers. Comprehensive examination of network anomaly detection techniques documents evolution from basic statistical approaches toward sophisticated hybrid models combining multiple detection methodologies (Ahmed, M. *et al.,* 2016). Contemporary financial fraud detection implementations typically layer complementary techniques, including statistical profiling, machine learning classification, and information theoretic approaches, maximizing detection accuracy while maintaining strict performance requirements. These hybrid architectures enable systems to identify both known fraud patterns alongside novel attack vectors previously unobserved. Financial service providers demonstrate particular innovation in developing specialized techniques addressing class imbalance and concept drift challenges, acknowledging that fraudulent transactions represent minuscule fractions among transaction volumes, while attack patterns continuously evolve to counter detection mechanisms.

Across Manufacturing Operations, predictive maintenance systems continuously monitor equipment sensor data, detecting subtle failure precursors. A prominent automotive manufacturer reduced unplanned downtime by twenty-seven percent through early identification of abnormal vibration patterns within assembly line robotics before catastrophic failures manifested. Recent healthcare anomaly detection research provides insights increasingly applied within manufacturing contexts, particularly regarding integration between domain knowledge and data-driven techniques (Samariya, D. *et al.,* 2023). These cross-domain applications leverage similar methodological foundations while adapting to industry-specific requirements. Advanced manufacturing implementations typically combine physics-based models encoding known failure modes alongside unsupervised learning approaches capable of identifying novel degradation patterns. Integration between these complementary approaches proves particularly effective for monitoring complex equipment with multiple potential failure modes and limited historical failure data. Organizations implementing these systems report substantial improvements regarding maintenance planning efficiency and spare parts inventory optimization beyond direct benefits through reduced unplanned downtime.

For Cybersecurity Applications, network defense systems analyze packet flows, authentication attempts, and access patterns, identifying potential intrusions. One telecommunications provider successfully detected and mitigated sophisticated zero-day attacks through the identification of anomalous lateral movement patterns within network infrastructure. Survey research examining network anomaly detection categorizes techniques

Publisher: SARC Publisher

into statistical-based, knowledge-based, and machine learning approaches, with modern implementations typically combining elements across categories (Ahmed, M. *et al.,* 2016). These systems face unique challenges, including adversarial detection contexts where attackers actively attempt to evade detection through mimicking normal behavior patterns. Most effective cybersecurity implementations employ ensemble approaches that analyze network traffic from multiple perspectives simultaneously, including volumetric analysis, protocol behavior examination, and communication pattern analysis between endpoints. This multi-faceted approach significantly increases the difficulty of evading detection compared to single-technique systems.

Within Healthcare Monitoring, intelligent systems process patient telemetry, detecting deteriorating conditions before reaching critical states. One hospital network implementing such technology reduced cardiac arrest incidents by twenty-three percent through the identification of subtle pattern changes within vital signs hours before traditional monitoring would trigger alarms. Systematic reviews examining healthcare anomaly detection document increasing application of advanced techniques across clinical domains, including physiological monitoring, medical imaging, and electronic health record analysis (Samariya, D. *et al.,* 2023). These implementations face unique challenges, including highly heterogeneous data sources, stringent privacy requirements, and critical consequences from false negative errors. Modern healthcare anomaly detection systems increasingly incorporate techniques handling missing data, integrating contextual information, and providing explainable outputs supporting clinical decision-making rather than merely generating alerts. Research demonstrates that successful implementations carefully balance automation alongside human expertise, using anomaly detection to augment rather than replace clinical judgment.

CASE STUDY

Intelligent Anomaly Detection in High-Speed Railway Networks

This case study presents empirical validation through a real-time anomaly detection system implemented for monitoring high-speed railway networks in collaboration with a major European transportation authority. The system was deployed across a network of sensors covering mechanical components, track infrastructure, and signaling systems. Over the six-month evaluation period, the system processed data streams from more than 800 distributed sensors, representing approximately 12,000 time-series metrics updated continuously.

The detection architecture implemented a multi-stage pipeline approach that addressed several key challenges in real-time anomaly detection identified in a comprehensive survey of explainable anomaly detection techniques (Li, Z. *et al.,* 2023). This research emphasizes that most anomaly detection algorithms, particularly those based on deep learning, lack interpretable results, which significantly limits practical applicability in critical infrastructure environments where operators must make rapid decisions based on alerts. Furthermore, the survey identifies that real-world deployment scenarios introduce unique challenges including temporal context preservation, adaptive threshold adjustment, and multi-dimensional correlation analysis that laboratory environments often fail to capture.

The implementation directly addressed these challenges through a hybrid approach combining statistical pre-filtering with ensemble machine learning models. The first processing stage employed statistical pattern recognition to identify potential anomalies, followed by a secondary analysis using both isolation forest and temporal convolutional network models. This architecture achieved substantial performance improvements over the previously deployed rule-based system. Specifically, time-to-detection measurements showed that potential equipment failures were identified significantly earlier, providing maintenance teams with extended response windows. The evaluation methodology employed time-series cross-validation using data partitioned into weekly segments throughout the deployment period, with statistical significance established through appropriate hypothesis testing.

The system infrastructure was built on open-source stream processing frameworks with customized extensions for state management and temporal correlation detection. Performance monitoring documented that the system consistently maintained throughput requirements while operating within acceptable latency parameters. A particularly noteworthy aspect of the implementation was the integration of explainability components, which transformed the system from a black-box detector into an operational decision support tool that enhanced rather than replaced human expertise.

Publisher: SARC Publisher

Research in anomaly detection approaches for real-world environments (Ahmed, M. *et al.,* 2016) indicates systems must contend with class imbalance, concept drift, and context-dependent normal patterns. This survey examines various techniques for anomaly detection across domains, noting that while financial applications historically received substantial attention, the underlying methodological approaches transfer effectively to other critical infrastructure scenarios including transportation networks. The framework categorizes detection techniques across statistical, knowledge-based, and machine learning domains, observing that most production deployments require elements from each category to achieve optimal performance.

The economic assessment of the railway anomaly detection system documented significant operational improvements, with maintenance scheduling efficiency substantially increased due to early warning capabilities. The return on investment calculations incorporated both direct cost reductions from prevented failures and indirect benefits from improved service reliability.

FUTURE DIRECTIONS

The intelligent anomaly detection field continues advancing along several promising trajectories:
Explainable AI (XAI) addresses inherent opacity within complex models, providing human-comprehensible explanations for anomaly determinations. This transparency becomes particularly crucial within regulated industries where decisions require justification to auditors and customers. Comprehensive surveys examining explainable anomaly detection approaches categorize techniques across multiple dimensions, including explanation scope (local versus global), explanation timing (ante-hoc versus post-hoc), and underlying detection methods supported (Li, Z. *et al.,* 2023). These analyses identify several key challenges specific to explaining anomalies, including the inherent rarity of anomalous instances, diverse manifestations anomalies exhibit, and domain-specific characteristics defining actionable explanations. Current research directions focus on developing explanation methods specifically designed for unsupervised anomaly detection scenarios, where labeled examples complicate traditional explainability approaches. Practical benefits extend beyond regulatory compliance toward improved operator trust, accelerated incident resolution, and more effective model refinement through expert feedback.

Federated Learning enables anomaly detection models to learn from distributed data sources without centralizing sensitive information. This approach preserves privacy while allowing models to benefit from broader patterns across organizational boundaries. Recent research explores specific challenges applying federated learning paradigms toward anomaly detection tasks, where extreme class imbalance and diverse anomaly manifestations create unique technical obstacles (Li, Z. *et al.,* 2023). These approaches show particular promise in scenarios where anomalies share common characteristics across organizational boundaries while maintaining domain-specific variations. Through enabling collaborative model training without data sharing, these techniques address both privacy concerns alongside fundamental data scarcity challenges limiting many detection systems. Emerging architectures incorporate mechanisms addressing data heterogeneity across participating organizations while maintaining privacy guarantees, making federated learning attractive within sensitive application domains.

Edge AI moves anomaly detection directly toward data sources, eliminating network latency while reducing bandwidth requirements. As edge devices become increasingly powerful, sophisticated models operate locally, enabling ultra-fast responses to detected anomalies. Research examining explainable anomaly detection highlights particular importance regarding edge deployments, where human operators may have limited access to supporting information when investigating alerts (Li, Z. *et al.,* 2023). These implementations typically employ specialized model compression techniques, deploying sophisticated detection capabilities within constrained computational environments characterizing edge devices. Resulting systems achieve detection latencies measured in milliseconds rather than seconds or minutes required by cloud-based approaches. Beyond performance advantages, edge deployment offers enhanced reliability through continued operation during network disruptions alongside reduced cloud computing costs through substantial decreases in data transmission volumes. As edge hardware capabilities continue advancing, boundaries between cloud processing requirements versus edge analysis capabilities continuously shift.

Publisher: SARC Publisher

Generative AI Integration represents a frontier regarding automated response capabilities. Beyond merely detecting anomalies, these systems generate hypotheses explaining root causes, simulate potential remediation strategies, and potentially implement automated corrections addressing certain problem categories. Research examining explainable anomaly detection increasingly explores how generative models produce comprehensive explanations contextualizing detected anomalies within broader system behaviors (Li, Z. *et al.*, 2023). These capabilities extend anomaly detection systems beyond passive monitoring toward active response orchestration. Generative approaches show particular promise in addressing "unknown unknowns" challenges within anomaly detection—identifying novel anomaly types unanticipated during system design phases. Through modeling normal behavior patterns and then systematically exploring possible deviation spaces, these systems identify potential vulnerabilities before manifestation within production environments. As these technologies mature, implementation increasingly incorporates safety systems with appropriate human oversight to prevent unintended consequences of automated intervention.

CONCLUSION

Intelligent anomaly detection in real-time data streams has evolved from a specialized capability to an essential component of modern data architectures. By combining the velocity of stream processing with the sophistication of machine learning, organizations can transform their approach from reactive incident response to proactive risk management. As streaming data volumes continue to grow exponentially, the ability to instantly separate signal from noise will increasingly determine which organizations can effectively navigate an increasingly complex digital landscape. The most successful implementations will be those that balance technical sophistication with practical operational requirements, delivering insights that are not just accurate but actionable within the critical window when intervention remains possible. In this sense, intelligent anomaly detection represents not just a technological advancement but a fundamental shift in how organizations perceive and respond to emerging opportunities and threats.

REFERENCES

1. Samaila, Y. A., Sebastian, P., Shuaibu, A. N., Ali, S. S. A., Muhammad, S. A., Yahya, M. S., ... & Sani, I. M. "Real-Time Deep Anomaly Detection: An Overview of Benchmark Datasets and Performance Metrics." *Transportation Research Procedia* 84 (2025): 418-425.
2. Cherniack, M., Balakrishnan, H., Balazinska, M., Carney, D., Cetintemel, U., Xing, Y., & Zdonik, S. B. "Scalable Distributed Stream Processing." *CIDR*. Vol. 3. (2003).
3. Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. "Machine learning for predictive maintenance: A multiple classifier approach." *IEEE transactions on industrial informatics* 11.3 (2014): 812-820.
4. Stonebraker, M., Çetintemel, U., & Zdonik, S. "The 8 requirements of real-time stream processing." *ACM Sigmod Record* 34.4 (2005): 42-47.
5. Chandola, V., Banerjee, A., & Kumar, V. "Anomaly detection: A survey." *ACM computing surveys (CSUR)* 41.3 (2009): 1-58.
6. Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. "Outlier detection for temporal data: A survey." *IEEE Transactions on Knowledge and data Engineering* 26.9 (2013): 2250-2267.
7. Ahmad, S., Lavin, A., Purdy, S., & Agha, Z. "Unsupervised real-time anomaly detection for streaming data." *Neurocomputing* 262 (2017): 134-147.
8. Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. "Deep learning for anomaly detection: A review." *ACM computing surveys (CSUR)* 54.2 (2021): 1-38.
9. Flouris, I., Giatrakos, N., Deligiannakis, A., Garofalakis, M., Kamp, M., & Mock, M. "Issues in complex event processing: Status and prospects in the big data era." *Journal of Systems and Software* 127 (2017): 217-236.
10. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. "A survey on concept drift adaptation." *ACM computing surveys (CSUR)* 46.4 (2014): 1-37.
11. Henning, S., & Hasselbring, W. "Theodolite: Scalability benchmarking of distributed stream processing engines in microservice architectures." *Big Data Research* 25 (2021): 100209.
12. Zhao, Y., Nasrullah, Z., & Li, Z. "Pyod: A python toolbox for scalable outlier detection." *Journal of machine learning research* 20.96 (2019): 1-7.
13. Ahmed, M., Mahmood, A. N., & Hu, J. "A survey of network anomaly detection techniques." *Journal of Network and Computer Applications* 60 (2016): 19-31.

Publisher: SARC Publisher

14. Samariya, D., Ma, J., Aryal, S., & Zhao, X. "Detection and explanation of anomalies in healthcare data." *Health Information Science and Systems* 11.1 (2023): 20.
15. Li, Z., Zhu, Y., & Van Leeuwen, M. "A survey on explainable anomaly detection." *ACM Transactions on Knowledge Discovery from Data* 18.1 (2023): 1-54.
16. Ahmed, M., Mahmood, A. N., & Islam, M. R. "A survey of anomaly detection techniques in financial domain." *Future Generation Computer Systems* 55 (2016): 278-288.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Mayilsamy, M." Intelligent Anomaly Detection in Real-Time Big Data Engineering." *Sarcouncil Journal of Engineering and Computer Sciences* 4.9 (2025): pp 577-589.

Publisher: SARC Publisher