World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

AI-driven anomaly detection in real-time streaming: enhancing human decision-making

Shakir Poolakkal Mukkath *

Walmart Global Tech, USA.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 410-420

Publication history: Received on 26 March 2025; revised on 02 May 2025; accepted on 04 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0583

Abstract

AI-driven anomaly detection in real-time streaming data has emerged as a transformative approach for organizations across industries facing unprecedented volumes of information. Traditional rule-based monitoring systems struggle with the complexity and evolving nature of modern data streams, often generating excessive false positives and missing subtle patterns that indicate fraud, system failures, or security breaches. This article examines how machine learning models integrated into streaming pipelines can enhance human decision-making by processing massive data volumes while identifying anomalies that would be impossible to detect manually. The technical foundations of real-time detection are explored, including stream processing architectures and various machine learning approaches such as statistical methods, unsupervised learning, and online algorithms. Implementation strategies for feature engineering, concept drift management, and latency optimization are discussed alongside industry applications in telecommunications, banking, retail, and cybersecurity. The article emphasizes that the most effective anomaly detection systems combine AI's pattern recognition capabilities with human expertise in a collaborative partnership, where machines handle data processing at scale while humans provide domain knowledge, contextual understanding, and strategic direction. This symbiotic relationship, supported by explainable AI and adaptive alert management, creates detection capabilities far superior to either humans or machines operating independently.

Keywords: Real-Time Anomaly Detection; Stream Processing; Human-AI Collaboration; Multi-Tier Architecture; Explainable AI

1. Introduction

In today's data-driven landscape, organizations across industries face an unprecedented deluge of information flowing through their systems at every moment. This exponential growth in data volume, velocity, and variety presents both an opportunity and a challenge: how to effectively monitor these massive data streams to identify anomalies that signal fraud, system failures, security breaches, or business opportunities. Global data creation continues to accelerate, with streaming data becoming increasingly central to business operations across telecommunications, financial services, manufacturing, and healthcare sectors.

Traditional rule-based monitoring systems, while effective for known patterns, struggle to scale with the complexity of modern data streams and fail to adapt to evolving threats and edge cases. Research published in the Journal of Computer and System Sciences indicates that conventional rule-based systems can only detect a fraction of novel anomalies in complex environments, with high false positive rates leading to significant alert fatigue among analysts [4]. This is where AI-driven anomaly detection in real-time streaming analytics has emerged as a transformative approach, capable of processing billions of events per second while identifying subtle deviations that would be impossible for human analysts to detect manually.

* Corresponding author: Shakir Poolakkal Mukkath

This article explores the technical foundations, implementation strategies, and real-world applications of AI-driven anomaly detection in streaming data, with a particular focus on how these systems enhance rather than replace human decision-making capabilities.

2. Technical Foundations of Real-Time Anomaly Detection

2.1. Stream Processing Architectures

The foundation of any real-time anomaly detection system is a robust stream processing architecture. Unlike batch processing, which operates on static datasets, stream processing handles data as it arrives, applying computations to each event or micro-batch of events in real-time. A comprehensive system architecture for real-time anomaly detection, detailed in IEEE Transactions on Network and Service Management, demonstrates impressive processing capabilities while maintaining linear scalability in large-scale network function virtualization (NFV) environments [1].

Table 1 Stream Processing Framework Comparison [1]

Framework	Throughput	Latency	Key Features
Apache Kafka	Very High	Low	Message persistence, exactly-once semantics
Apache Flink	High	Very Low	Stateful processing, event time handling
Apache Spark Streaming	High	Moderate	ML integration, micro-batch processing
Apache Pulsar	Very High	Low	Multi-tenancy, geo-replication, tiered storage

Modern stream processing frameworks like Apache Kafka, Apache Flink, and Apache Spark Streaming provide distributed processing capabilities essential for handling high-throughput data streams. Recent benchmark studies presented at the IEEE International Conference on Big Data revealed that these platforms can achieve remarkable throughput rates on modest hardware configurations [3]. These frameworks typically implement several critical components working in concert: A data ingestion layer responsible for capturing and buffering incoming data streams; a processing layer where anomaly detection algorithms operate on the streaming data; a storage layer for persisting both raw data and detected anomalies; and an alerting layer for notifying human operators when anomalies are detected, with advanced correlation techniques substantially reducing alert volumes through intelligent aggregation and prioritization [4].

2.2. Machine Learning Models for Streaming Anomaly Detection

Several categories of machine learning approaches have proven effective for anomaly detection in streaming contexts, with their relative merits thoroughly documented in a comprehensive evaluation published in Expert Systems with Applications [2].

Statistical and distance-based methods establish what constitutes "normal" behavior and flag deviations from these established patterns. Isolation Forest has demonstrated high detection accuracy with low false positive rates and remarkably fast training times for datasets containing millions of data points. This technique isolates anomalies through random partitioning of the feature space, making it particularly effective for high-dimensional data common in networking and cybersecurity applications [2]. In contrast, One-Class Support Vector Machines show slightly lower accuracy with higher false positive rates, but remain valuable for certain applications due to their strong theoretical foundations.

Unsupervised learning approaches are particularly valuable when labeled training data is unavailable, a common scenario in real-world anomaly detection deployments. Autoencoder neural networks that learn to compress and reconstruct input data have achieved impressive detection accuracy with low false positive rates, though at the cost of longer training times on benchmark datasets. These networks flag instances with high reconstruction error as anomalies, effectively learning the underlying patterns of normal behavior without requiring explicit examples of anomalies [2]. Even more impressive, Long Short-Term Memory (LSTM) networks designed to capture temporal dependencies in sequential data have reached excellent detection accuracy with very low false positive rates, though with substantially longer training times.

Time series-specific methods address the critical temporal nature of streaming data. The real-time anomaly detection architecture detailed in IEEE Transactions on Network and Service Management demonstrates that these methods can achieve high detection rates for known attack patterns in NFV environments when properly implemented [1]. This remarkable performance stems from the ability of these models to capture both point anomalies (individual outlier values) and contextual anomalies (values that are suspicious only in certain contexts or sequences).

Online learning algorithms adapt to concept drift and evolving patterns in streaming data, addressing one of the most significant challenges in operational deployments. Research in Expert Systems with Applications reveals that a majority of deployed machine learning models show significant accuracy decline after just a few months without retraining or adaptation mechanisms [2]. This finding underscores the necessity of algorithms specifically designed for streaming environments that can continuously update their understanding of normal and anomalous patterns.

2.3. Implementation Strategies for Streaming Anomaly Detection

2.3.1. Feature Engineering for Streaming Data

Effective feature engineering is crucial for real-time anomaly detection and presents unique challenges in streaming contexts. Temporal features including rolling statistics, rates of change, and frequency domain transformations have been shown to significantly improve detection accuracy in NFV environments while adding minimal processing overhead [1]. These features capture the evolution of metrics over time, enabling the identification of subtle trend changes that might indicate emerging issues before they become critical failures.

Contextual features such as time of day, day of week, and other cyclical patterns provide essential information about when certain behaviors should be considered normal versus suspicious. The Journal of Computer and System Sciences publication demonstrates that properly incorporating these contextual elements substantially reduces false positive rates in enterprise environments, significantly alleviating alert fatigue among security and operations teams [4]. This dramatic improvement stems from the system's enhanced ability to distinguish between legitimate variations in behavior (such as increased network traffic during business hours) and genuinely anomalous activity.

Relationship features capturing interactions between entities in complex systems have proven particularly valuable in network environments. The architecture presented in IEEE Transactions on Network and Service Management leverages graph-based relationship features to improve anomaly detection in virtualized network functions, where the interactions between components often provide more insight than individual metrics in isolation [1]. By modeling these relationships explicitly, the detection system gains awareness of the operational context, enabling more nuanced and accurate identification of truly problematic behaviors.

Incremental feature computation techniques that update values efficiently as new data arrives are essential for maintaining performance at scale. Research presented at the IEEE International Conference on Big Data demonstrates that optimized incremental computation approaches can dramatically reduce CPU utilization compared to naive recomputation approaches, while maintaining feature accuracy within tight margins of full recalculation values [3]. This efficiency enables the processing of more streams with fewer resources, making large-scale deployments economically viable.

2.3.2. Handling Concept Drift

One of the most significant challenges in streaming environments is concept drift—the phenomenon where the statistical properties of the target variable change over time. Expert Systems with Applications documents that a large majority of production machine learning models experience significant concept drift within months of deployment, highlighting the critical importance of adaptation mechanisms [2]. This drift occurs as usage patterns evolve, new types of legitimate behavior emerge, and adversaries modify their tactics to evade detection.

Adaptive windowing dynamically adjusts the time window used for model training based on detected changes in data distribution. Implementations detailed in the Journal of Computer and System Sciences demonstrate that these techniques can maintain the vast majority of original detection accuracy even after major distribution shifts, while requiring modest memory resources per monitored stream [4]. This remarkable resilience stems from the continuous adjustment of the relevant historical context used for anomaly determination.

Ensemble methods maintain multiple models trained on different time periods or data subsets, combining their predictions to achieve more robust results. The comprehensive evaluation in Expert Systems with Applications shows that ensemble approaches deliver substantial improvements in F1-score over single models during periods of rapid

concept drift, making them particularly valuable in volatile environments [2]. This improvement comes from the ensemble's ability to hedge against the weakness of any individual model faced with evolving data patterns.

Drift detection algorithms explicitly monitor for changes in data distributions and trigger model updates only when necessary, balancing adaptivity with computational efficiency. Research presented at the IEEE International Conference on Big Data reveals that modern techniques can detect the vast majority of significant drifts within reasonable sample sizes while maintaining low false positive rates, ensuring that adaptation resources are deployed only when truly needed [3]. This targeted approach to adaptation conserves computational resources while maintaining detection effectiveness.

2.3.3. Balancing Latency and Accuracy

Real-time detection inherently involves tradeoffs between processing speed and detection quality. The system architecture detailed in IEEE Transactions on Network and Service Management achieves low latencies for complex analytics workloads while maintaining high detection rates for known attack patterns, demonstrating that these objectives are not mutually exclusive when properly implemented [1]. This performance enables truly real-time responses to emerging threats, often detecting and mitigating issues before they impact end users.

Multi-tier detection approaches use lightweight algorithms for initial screening and more complex models for verification of potential anomalies. The Journal of Computer and System Sciences publication demonstrates that this approach dramatically reduces overall computational requirements while maintaining nearly all of the detection capability of using only the most complex model [4]. This remarkable efficiency improvement makes comprehensive anomaly detection feasible even in resource-constrained environments like edge devices or legacy infrastructure.

Hardware acceleration through GPUs, FPGAs, or specialized ASICs delivers substantial performance improvements for computationally intensive algorithms. Research presented at the IEEE International Conference on Big Data documents significant speedups for deep learning models used in anomaly detection, enabling more sophisticated algorithms to operate within real-time constraints [3]. These acceleration technologies effectively expand the algorithmic options available to system designers, allowing the deployment of more accurate but computationally demanding approaches.

Edge computing moves anomaly detection closer to data sources, addressing both latency and bandwidth challenges. The IEEE International Conference on Big Data research shows that edge deployment substantially reduces transmission latency in distributed sensor environments, enabling faster detection and response while dramatically reducing the network overhead associated with centralized processing [3]. This architectural approach is particularly valuable for applications like industrial IoT, where both response time and network constraints are significant considerations.

2.3.4. Event Correlation and Contextual Analysis

Individual anomalies often lack significance without context, leading advanced systems to implement sophisticated correlation and enrichment capabilities. Real-time event correlation connects related anomalies across different data streams, improving detection rates for complex attack patterns while reducing false positives in virtualized network environments [1]. This correlation capability stems from an understanding that sophisticated attacks and failures often manifest as patterns of related events rather than isolated incidents.

Causal analysis identifies potential root causes of anomalies, enabling more targeted responses. The Journal of Computer and System Sciences publication demonstrates that graph-based causal analysis approaches substantially reduce mean time to resolution in enterprise IT environments by guiding operators directly to the underlying issues rather than just the symptoms [4]. This dramatic improvement in troubleshooting efficiency translates directly to reduced downtime and lower operational costs.

Temporal pattern recognition detects sequences of events that together indicate an anomaly, even when individual events appear normal in isolation. The system architecture in IEEE Transactions on Network and Service Management shows that these techniques capture the vast majority of multi-stage security breaches that would be missed by point-in-time detection approaches [1]. This capability is particularly valuable in security contexts, where attackers deliberately structure their activities to avoid triggering traditional detection mechanisms.

Contextual enrichment augments detected anomalies with business context, significantly increasing first-time-right resolution rates and reducing escalations according to the Journal of Computer and System Sciences research [4]. By providing human operators with comprehensive context around detected anomalies, including affected services,

potential business impact, and historical patterns, these systems dramatically improve the efficiency of response activities, reducing both mean-time-to-resolution and the operational burden on specialist teams.

3. Industry Applications and Human-AI Partnership in Anomaly Detection

3.1. Industry Applications and Use Cases

3.1.1. Telecommunications

Telecommunications networks generate massive volumes of performance metrics, call detail records, and signaling data, creating an ideal environment for AI-driven anomaly detection solutions. A comprehensive study published in the Journal of Network and Systems Management reveals that advanced anomaly detection systems in telecommunications now provide early warning of network issues several minutes before they cause customer-visible service degradation, creating a critical window for preemptive intervention that has resulted in a substantial reduction in unplanned service outages across monitored networks [5]. This proactive approach transforms network operations from reactive troubleshooting to preventive maintenance, fundamentally changing how telecommunications providers maintain service quality.

Table 2 Industry Applications [5]

Industry	Key Applications	Primary Benefits	Implementation Challenges
Telecommunications	Network monitoring, fraud detection, failure prediction	Proactive maintenance, reduced outages	High data volumes, complex topologies
Banking	Payment fraud, trading anomalies, AML monitoring	Real-time prevention, compliance	Strict latency requirements, high accuracy demands
Retail	Inventory management, customer journey analysis	Shrinkage reduction, UX optimization	Distributed operations, edge deployment needs
Cybersecurity	Intrusion detection, behavior analytics	Early threat detection, reduced alert fatigue	Sophisticated adversaries, evolving attack vectors

Network fraud detection represents another high-impact application area, with recent implementations achieving high accuracy in identifying sophisticated SIM swap fraud attempts that traditional rule-based systems often miss entirely [5]. These AI-driven systems analyze complex patterns across calling history, location data, authentication attempts, and account behavior to identify suspicious activities with unprecedented precision. The deployment of streaming anomaly detection for network security monitoring has simultaneously reduced network operations center alerts significantly, addressing the critical problem of alert fatigue that has long plagued telecommunications security teams [5]. This dramatic reduction in false positives allows security personnel to focus their expertise on genuine threats rather than chasing harmless anomalies.

Infrastructure failure prediction has emerged as a particularly valuable capability, with AI-powered systems now demonstrating impressive accuracy in predicting equipment failures hours in advance by identifying subtle precursors in telemetry data [5]. By detecting anomalous patterns in power consumption, temperature fluctuations, error rates, and other metrics across network equipment, these systems enable precisely targeted preventive maintenance that avoids costly downtime. Beyond operational benefits, anomaly detection has delivered unexpected efficiency gains, with anomaly-driven resource optimization reducing power consumption substantially across monitored telecommunications infrastructure [5]. This energy reduction demonstrates how advanced anomaly detection not only improves service reliability but also contributes to sustainability goals and operational cost reduction.

3.1.2. Banking and Financial Services

Financial institutions process enormous transaction volumes, making them ideal candidates for streaming anomaly detection solutions that can identify fraudulent or suspicious activities in real-time. Research published in IEEE Transactions on Computational Social Systems demonstrates that modern streaming analytics platforms have achieved substantial processing capacities on modest hardware configurations, with minimal detection latency from transaction initiation to fraud determination [6]. This performance enables genuine real-time intervention at scale, essential for an industry where transaction volumes and sophisticated fraud schemes continue to grow exponentially.

The effectiveness of these systems is particularly evident in challenging environments like card-not-present transactions, where machine learning approaches have achieved high detection rates for fraudulent activities [6]. This capability addresses one of the most persistent and costly fraud vectors in digital commerce. Equally significant, these advanced systems have dramatically reduced false positive rates compared to traditional rule-based approaches [6]. This improvement in precision translates directly to operational efficiency by eliminating unnecessary fraud investigations while simultaneously improving customer experience by reducing legitimate transaction declines.

The application of streaming analytics to cryptocurrency transactions represents a cutting-edge use case, with systems demonstrating strong detection rates for cross-chain cryptocurrency laundering attempts [6]. This capability is increasingly critical as financial institutions expand their digital asset operations while maintaining regulatory compliance. The comprehensive implementation of AI-driven fraud detection across payment channels has demonstrated financial benefits beyond fraud prevention alone, with substantial savings through the combination of reduced fraud losses, lower operational costs, and improved customer retention [6]. This clear return on investment has accelerated adoption across the financial services sector, making anomaly detection one of the most widely deployed AI applications in banking.

3.1.3. Retail and E-commerce

Retailers are increasingly leveraging anomaly detection to optimize both online and physical operations, with applications spanning inventory management, pricing optimization, customer experience, and supply chain oversight. The most recent implementations described in IEEE Internet of Things Journal have demonstrated linear scaling to environments with thousands of distributed edge nodes across retail operations, creating a comprehensive monitoring fabric that extends from individual stores to centralized e-commerce platforms [8]. This scalability has made enterprise-wide anomaly detection feasible even for the largest retail organizations.

One particularly valuable implementation approach involves multi-tier detection architectures that balance responsiveness with analytical depth. Edge-based initial detection layers achieve minimal latencies, compared to cloud-based processing, enabling truly real-time responses to time-sensitive anomalies [8]. This performance differential is especially important for applications like payment fraud detection or inventory shrinkage prevention where immediate intervention is essential. By implementing intelligent filtering at the edge, these architectures reduce cloud bandwidth requirements significantly while simultaneously reducing computational demands in centralized processing [8]. This efficiency makes comprehensive anomaly detection economically viable at enterprise scale.

The effectiveness of multi-tier approaches is evident in their detection performance metrics, with tier-1 edge detection achieving good recall, tier-2 intermediate processing reaching better recall, and tier-3 deep analysis delivering excellent recall with outstanding precision [8]. This layered approach ensures that obvious anomalies are caught immediately at the edge while more subtle or complex patterns receive the in-depth analysis they require. Beyond performance benefits, optimized edge-based anomaly detection extends device battery life substantially compared to continuous cloud transmission approaches [8]. This efficiency is particularly valuable in retail environments leveraging battery-powered IoT devices for inventory tracking, environmental monitoring, and customer analytics.

3.1.4. Cybersecurity

Cybersecurity remains perhaps the most established use case for anomaly detection, with applications spanning network security, user behavior analytics, endpoint protection, and API security. Research published in ACM Transactions on Computer-Human Interaction has documented a substantial increase in security analyst productivity measured by resolved incidents per hour when working with collaborative AI systems compared to traditional security tools [7]. This dramatic productivity improvement addresses the persistent shortage of qualified cybersecurity personnel that has plagued the industry for decades.

The human-AI partnership approach has demonstrated significant improvement in decision accuracy compared to either AI-only or human-only approaches to security monitoring [7]. This finding underscores that the most effective security operations leverage both machine capabilities in pattern recognition and human expertise in contextual understanding and strategic thinking. Explainable AI interfaces that provide clear rationales for detected anomalies have proven particularly valuable, enabling much faster identification of root causes when security incidents occur [7]. This acceleration in root cause analysis directly translates to reduced mean-time-to-remediation and minimized impact from security events.

Beyond immediate operational benefits, collaborative human-AI systems have demonstrated remarkable advantages in analyst development, with a majority of junior analysts reaching senior-level performance after just weeks with AI

assistance compared to months with traditional training approaches [7]. This accelerated proficiency development represents a powerful solution to the cybersecurity skills gap by both maximizing the effectiveness of existing personnel and accelerating the development of new talent. Perhaps most importantly for long-term effectiveness, explanation-generating systems have demonstrated substantial improvement in appropriate trust calibration, reducing both over-reliance on automation and unwarranted skepticism [7]. This balanced approach ensures that human analysts and AI systems work as true partners, each contributing their unique strengths to the security mission.

3.2. The Human-AI Partnership in Anomaly Detection

While AI excels at processing massive data volumes and detecting subtle patterns, human expertise remains essential for effective anomaly detection systems. The ideal approach combines the strengths of both, creating a symbiotic relationship that significantly outperforms either humans or machines operating independently.

3.2.1. Human Expertise in System Design

Feature selection represents a critical area for human contribution, with domain experts identifying which metrics and derived features are most relevant for specific detection scenarios. Research in ACM Transactions on Computer-Human Interaction demonstrates that collaborative feature engineering approaches leverage both human domain knowledge and machine learning capabilities to identify optimal detection signals [7]. The documented improvement in decision accuracy with collaborative systems versus AI-only or human-only approaches provides compelling evidence for the value of this partnership model [7]. This improvement stems from the complementary nature of human contextual understanding and machine pattern recognition capabilities.

Threshold calibration similarly benefits from human judgment, as determining appropriate sensitivity levels requires balancing technical detection capabilities with business impact considerations. Domain experts understand the operational context and relative importance of different anomalies, enabling more nuanced threshold settings than purely statistical approaches. The integration of human feedback has proven particularly valuable in operational settings, with studies documenting a substantial reduction in cognitive load as measured by the NASA Task Load Index when analysts work with well-designed collaborative systems [7]. This reduction in mental burden allows security personnel to focus their cognitive resources on strategic analysis rather than routine alert triage.

Edge case handling represents another area where human insight proves irreplaceable. Complex environments inevitably generate unusual but legitimate scenarios that automated systems may flag as anomalous. Human analysts can recognize these situations, provide appropriate context, and help design approaches for handling rare but important scenarios without generating false positives. This capability is particularly valuable in environments with limited historical data or rapidly evolving conditions where machine learning models may struggle to establish reliable baselines. The improvement in appropriate trust calibration documented with explanation-generating systems demonstrates how well-designed interfaces help maintain this critical human oversight while avoiding both over-reliance and under-utilization of AI capabilities [7].

3.2.2. Explainability and Interpretability

For effective human-AI collaboration, anomaly detection systems must be transparent in their decision-making processes. Feature attribution technologies that identify which metrics contributed most to an anomaly determination enable analysts to quickly understand detection rationale and evaluate its validity. Research shows that explainable systems drive substantial operational improvements, with documented faster identification of root causes when security incidents occur [7]. This acceleration in analysis directly translates to faster remediation and reduced impact from security events or operational disruptions.

Visualization techniques play a crucial role in making complex patterns understandable through effective presentation. Advanced visualization approaches transform multidimensional anomaly data into intuitive representations that leverage human visual processing strengths. Similarly, counterfactual explanations that show what would need to change for an anomaly to be considered normal help analysts evaluate the robustness of detections and understand decision boundaries. Confidence metrics that communicate the system's level of certainty about anomalies enable more nuanced human responses, with high-confidence alerts receiving immediate attention while lower-confidence alerts undergo more careful review. These explainability features collectively contribute to the documented increase in analyst productivity by streamlining the investigation process and eliminating unnecessary analysis steps [7].

The benefits of explainable systems extend beyond immediate operational efficiency to analyst development and organizational knowledge management. Research has shown that junior analysts working with explanation-generating

systems reach senior-level performance significantly faster than with traditional training approaches [7]. This accelerated proficiency development occurs because explanations serve as continuous contextual training, helping analysts understand not just what anomalies look like but why they matter and how they relate to underlying systems or threats.

3.2.3. Adaptive Alert Management

Sophisticated systems employ strategies to optimize the human-AI workload distribution through intelligent alert management. The reduction in cognitive load measured by the NASA Task Load Index demonstrates that well-designed systems actually reduce analyst stress rather than adding to it [7]. This counterintuitive finding highlights that mental fatigue in security operations often stems not from the complexity of genuine security incidents but from the cognitive burden of sifting through numerous false positives and ambiguous alerts.

The implementation of tiered detection architectures further optimizes workload distribution, with edge-based initial detection achieving good recall for obvious anomalies while reserving analyst attention for the more complex cases that benefit from human judgment [8]. By ensuring that the division of labor between human and machine components aligns with their respective strengths, these systems maximize overall system effectiveness while minimizing unnecessary human intervention. This approach is particularly valuable in environments with limited analyst resources or high alert volumes.

Contextual enrichment provides analysts with relevant business context for decision-making, including affected assets, previous similar incidents, and potential business impact. This enrichment transforms raw technical alerts into business-contextualized incidents that enable appropriate prioritization and response. Research has shown that explanation-generating systems improve appropriate trust calibration, ensuring that analysts can accurately judge when to rely on automated assessments and when to apply additional scrutiny [7]. This balanced approach prevents both the blind acceptance of algorithm outputs and the unnecessary duplication of analysis that automated systems have already performed reliably.

Perhaps most importantly, well-designed systems implement feedback loops that learn from human responses to improve future alerting. By tracking which alerts analysts dismiss, investigate, or escalate, these systems continuously refine their understanding of what constitutes a genuinely concerning anomaly in each specific environment. This adaptive approach ensures that systems become increasingly aligned with organizational priorities and operational realities over time. The documented increase in security analyst productivity represents the cumulative impact of these various human-AI collaboration optimizations working in concert [7].

3.3. Technical Implementation Example: Multi-Tier Anomaly Detection System

To illustrate the concepts discussed, consider a technical architecture for a multi-tier anomaly detection system based on research published in IEEE Internet of Things Journal [8].

Table 3 Multi-Tier Anomaly Detection Architecture [8]

Tier	Location	Techniques	Response Time	Key Responsibilities
Tier 1: Statistical	Edge	Z-scores, moving averages	Milliseconds	Initial filtering, obvious anomaly detection
Tier 2: ML Verification	Fog/Cluster	Isolation forests, autoencoders	10-100ms	Verification, false positive reduction
Tier 3: Contextual	Cloud	Graph analysis, correlation engines	100ms-seconds	Context enrichment, impact assessment
Human Analysis	SOC/NOC	Visual analytics, case management	Minutes-hours	Strategic decisions, feedback to system

3.3.1. Tier 1: Real-Time Statistical Analysis

The first tier implements lightweight statistical methods such as z-scores and moving averages with extremely low latency, achieving quick detection times at the edge [8]. This performance enables truly real-time responses for time-sensitive applications like fraud detection or safety-critical monitoring. By processing the full data stream at this initial

tier, the system ensures comprehensive coverage while intelligent filtering reduces the load on subsequent processing tiers. The documented reduction in computational demands achieved through tier-1 filtering demonstrates the economic efficiency of this approach [8].

These edge-based first-tier detectors typically run on resource-constrained devices, requiring careful optimization for specific hardware platforms. Despite these constraints, modern implementations achieve good recall for anomaly detection even at this initial screening layer [8]. The emphasis at this stage is on computational efficiency and comprehensive coverage rather than detection precision, with systems designed to favor sensitivity over specificity to ensure potential issues are flagged for further analysis. This approach ensures that obvious anomalies receive immediate attention while more subtle patterns undergo additional scrutiny in subsequent tiers.

Beyond anomaly detection benefits, optimized edge processing delivers significant infrastructure advantages, with documented reduction in cloud bandwidth requirements and extension in device battery life for battery-powered sensors and monitoring devices [8]. These efficiency improvements make comprehensive anomaly detection feasible even in environments with limited connectivity or power constraints, significantly expanding the potential deployment scenarios for real-time monitoring.

3.3.2. Tier 2: Machine Learning Verification

The second tier employs more complex models such as isolation forests and autoencoders to verify anomalies flagged by the first tier. This intermediate processing layer achieves improved recall while maintaining reasonable response times, with moderate latencies for cloud-based processing [8]. By analyzing only, the subset of data identified as potentially anomalous by the first tier, this layer can apply more sophisticated techniques without becoming a computational bottleneck. The multi-tier architecture has demonstrated linear scaling to environments with thousands of distributed edge nodes [8], proving its viability for enterprise-scale deployments.

These second-tier systems typically run in distributed computing environments, balancing analytical sophistication with operational performance requirements. By combining multiple detection approaches, this tier significantly reduces false positives while maintaining high detection sensitivity for genuine anomalies. The intelligent workload distribution between tiers ensures that computational resources are applied where they provide the greatest value, with simple anomalies handled efficiently at the edge while more complex patterns receive the additional analysis they require.

3.3.3. Tier 3: Contextual Analysis

The third tier performs deep contextual analysis, correlating related anomalies and enriching them with business context to create actionable intelligence. This comprehensive analysis achieves excellent recall with outstanding precision [8], providing high-confidence determinations that minimize both false negatives and false positives. By connecting anomalies across different systems and data sources, this tier identifies complex, multi-stage patterns that would be invisible when examining individual events in isolation.

This contextual understanding determines business impact and urgency based on comprehensive situational awareness rather than isolated technical indicators. The high precision achieved at this tier [8] ensures that alerts reaching human analysts represent genuine issues requiring attention, dramatically reducing alert fatigue and improving operational efficiency. This level of precision is particularly critical in high-volume environments where even a small percentage of false positives can overwhelm human analysts.

3.3.4. Human Analyst Interface

The final component presents prioritized, grouped, and contextualized anomalies to human analysts through an intuitive interface designed to maximize analytical efficiency. These interfaces provide interactive visualization and exploration tools that enable analysts to quickly understand complex situations and investigate underlying causes. Research shows that well-designed interfaces with appropriate explanation capabilities enable substantially faster identification of root causes when security incidents occur [7], dramatically improving response times and reducing incident impact.

Most importantly, these interfaces capture analyst feedback for continual improvement, creating a virtuous cycle where human insights enhance system performance over time. This feedback-driven approach results in continuous improvement in both detection accuracy and operational efficiency, with documented productivity increases in terms of resolved incidents per analyst hour [7]. Additionally, systems with effective explanation capabilities have

demonstrated significant improvement in appropriate trust calibration [7], ensuring that analysts develop a nuanced understanding of system capabilities and limitations rather than either blindly trusting or inappropriately discounting automated findings.

3.4. Challenges and Future Directions

Despite significant advances, several challenges remain in the field of real-time anomaly detection.

3.4.1. Technical Challenges

Imbalanced data represents a fundamental challenge, as anomalies are inherently rare, making model training difficult. This class imbalance can lead to systems that either miss genuine anomalies or generate excessive false positives. Feature stability presents another significant challenge, as ensuring feature distributions remain consistent between training and production environments requires careful engineering and monitoring. The documented performance of multi-tier detection architectures [8] demonstrates that these challenges can be effectively addressed through careful system design and implementation.

Model drift remains an ongoing challenge, with gradual degradation of model performance over time as data distributions and anomaly patterns evolve. Addressing this drift requires continuous monitoring and adaptation mechanisms to maintain detection effectiveness. Computation efficiency similarly presents ongoing challenges, particularly as data volumes continue to grow exponentially. The documented reduction in computational demands achieved through intelligent filtering [8] illustrates how architectural approaches can address these efficiency challenges while maintaining detection effectiveness.

3.4.2. Future Research Directions

Self-supervised learning represents a promising research direction, potentially allowing systems to leverage unlabeled data more effectively for anomaly detection. By learning normal patterns without requiring explicit examples of anomalies, these approaches could dramatically improve training efficiency and detection effectiveness in environments with limited historical data. Federated anomaly detection similarly offers exciting possibilities, enabling detection across organizational boundaries while preserving privacy and data sovereignty. This approach could be particularly valuable in sectors like finance and healthcare where data sharing is restricted but cross-organizational attack detection would provide significant benefits.

Table 4 Future Research Directions [8]

Research Area	Potential Impact	Promising Approaches
Self-Supervised Learning	Reduced reliance on labeled data	Contrastive learning, reconstruction tasks
Federated Anomaly Detection	Cross-organizational detection, privacy preservation	Secure aggregation, vertical federated learning
Neuromorphic Computing	Energy efficiency, edge deployment	Spiking neural networks, event-based processing
Explainable AI	Improved trust, faster troubleshooting	Attention mechanisms, concept-based explanations
Autonomous Response	Reduced time-to-remediation, scalable protection	Guided autonomy, confidence-based automation

Neuromorphic computing represents another intriguing frontier, with specialized hardware architectures optimized for anomaly detection potentially offering dramatic performance improvements. The extension in device battery life already achieved through software optimization [8] suggests that hardware-level improvements could deliver even more significant efficiency gains, making comprehensive anomaly detection viable in even the most resource-constrained environments. Perhaps most speculatively, quantum machine learning approaches are being explored for high-dimensional anomaly detection, potentially offering exponential speedups for certain detection algorithms once practical quantum computing becomes available.

4. Conclusion

AI-driven anomaly detection in real-time streaming represents a fundamental evolution in how organizations monitor and respond to critical events across diverse operational contexts. The technologies and approaches discussed throughout this article demonstrate that effective anomaly detection is not merely a technical challenge but a sociotechnical one, requiring thoughtful integration of machine capabilities with human expertise. By implementing multi-tier architectures that balance computational efficiency with detection accuracy, organizations can achieve comprehensive monitoring at scale while making economical use of limited resources. The human-AI partnership model proves consistently superior to either humans or machines operating in isolation, with collaborative systems demonstrating significant advantages in detection accuracy, operational efficiency, and analyst development. As data volumes continue to grow exponentially and threat landscapes evolve in complexity, the partnership between AI systems and human analysts will become increasingly crucial. Organizations implementing these technologies should focus not just on the technical aspects of model selection and infrastructure design, but equally on human factors that determine ultimate effectiveness: analyst training, workflow integration, explainability, and continuous feedback mechanisms. The most successful implementations will be those that thoughtfully integrate both technical capabilities and human considerations, creating systems that augment rather than replace human decision-making, allowing each component to contribute its unique strengths to the detection and response process. Looking forward, emerging approaches such as self-supervised learning, federated detection, and specialized hardware architectures promise to further enhance capabilities while addressing current limitations, ensuring that anomaly detection systems continue to evolve alongside the challenges they are designed to address.

References

[1] Anton Gulenko et al, "A System Architecture for Real-time Anomaly Detection in Large-scale NFV Systems,"December 2016, Procedia Computer Science. Available: https://www.researchgate.net/publication/306052028_A_System_Architecture_for_Real-time_Anomaly_Detection_in_Large-scale_NFV_Systems

[2] Félix Iglesias Vázquez, et al, "Anomaly detection in streaming data: A comparison and evaluation study," Expert Systems with Applications, Volume 233, 15 December 2023, Available: https://www.sciencedirect.com/science/article/pii/S0957417423014963

[3] Harold Castro, "Real-Time Anomaly Detection Using Streaming Data Platforms," June 2024, Research Gate, Available: https://www.researchgate.net/publication/387575989_Real-Time_Anomaly_Detection_Using_Streaming_Data_Platforms

[4] Martin Grill, et al, "Reducing false positives of network anomaly detection by local adaptive multivariate smoothing," Journal of Computer and System Sciences, Volume 83, Issue 1, February 2017, Available: https://www.sciencedirect.com/science/article/pii/S0022000016300022

[5] Enerst Edozie, et al, "Artificial intelligence advances in anomaly detection for telecom networks," January 2025, Artificial Intelligence Review, Available: https://www.researchgate.net/publication/388386821_Artificial_intelligence_advances_in_anomaly_detection_for_telecom_networks

[6] Amarnath Immadisetty, "Real-Time Fraud Detection Using Streaming Data in Financial Transactions," January 2024, Research Gate, Available: https://www.researchgate.net/publication/389628199_Real-Time_Fraud_Detection_Using_Streaming_Data_in_Financial_Transactions

[7] Olayiwola Blessing Akinnagbe, "Human-AI Collaboration: Enhancing Productivity and Decision-Making," November 2024, International Journal of Education Management and Technology, Available: https://www.researchgate.net/publication/386225744_Human-AI_Collaboration_Enhancing_Productivity_and_Decision-Making

[8] Daniel Steven, et al, "Cloud-Native AI for Real-Time Anomaly Detection in Edge Computing," December 2024, Research Gate, Available: https://www.researchgate.net/publication/390486279_Cloud-Native_AI_for_Real-Time_Anomaly_Detection_in_Edge_Computing.