Scientific Research Publishing

Journal of Computer and Communications, 2026, 14(6), 27-37
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Artificial Intelligence as a Security Mediation Layer in Enterprise Environments

Ustinovich Viktor Mikhailovich

LLC "R-Alpha Lab", Moscow, Russia
Email: science.field.work1@gmail.com

How to cite this paper: Ustinovich, V.M. (2026) Artificial Intelligence as a Security Mediation Layer in Enterprise Environments. *Journal of Computer and Communications*, **14**, 27-37.
https://doi.org/10.4236/jcc.2026.146003

Received: May 6, 2026
Accepted: June 15, 2026
Published: June 18, 2026

Abstract

Traditional perimeter-based security models have proven structurally inadequate against the adaptive, multi-vector attack strategies that characterize contemporary enterprise threats. This article examines the theoretical and architectural foundations of artificial intelligence as a security mediation layer in large organizational environments, focusing on behavioral analytics, anomaly detection, automated policy enforcement, and zero-trust integration. Drawing on peer-reviewed scholarship and practitioner-research contributions, the study traces how recent architectural work engages problems that academic research has identified but rarely resolved at the deployment level, among them temporal threat modeling, distributed policy consistency, and the governance of autonomous decision systems. The analysis reveals persistent tensions between detection sensitivity and operational manageability, between model sophistication and institutional accountability, and between autonomous response and the interpretive role of human analysts. Resolving these tensions requires coherent architectural and governance frameworks sustained over the operational lifecycle of deployed systems.

Keywords

Artificial Intelligence, Enterprise Cybersecurity, Behavioral Analytics, Anomaly Detection, Zero-Trust Architecture, Machine Learning, Security Operations Center, Autonomous Threat Response, Distributed Systems Security

1. Introduction

Enterprise cybersecurity has undergone a fundamental reconceptualization over the past decade. The organizations that digital infrastructure underpins, including financial institutions, healthcare systems, industrial operators, and government

agencies, face adversaries whose capabilities and operational patience have consistently outpaced the defensive paradigms deployed against them. The inadequacy of perimeter defense, signature matching, and rule-based alerting is structural: the threat landscape these methods were designed to address has changed in ways that undermine the assumptions on which they rest [1] [2].

The term security mediation layer as used in this article denotes a software architectural tier that sits between raw telemetry streams and enforcement infrastructure, performing three functions that no single existing tool class combines: continuous behavioral risk scoring, dynamic policy generation, and coordinated enforcement orchestration across heterogeneous network segments. It is distinct from intrusion detection systems, which produce alerts but do not enforce policy; from SOAR platforms, which automate predefined response playbooks but do not generate policies adaptively from behavioral signals; and from zero-trust policy engines, which evaluate access requests against stated rules but do not themselves produce those rules from learned behavioral baselines. The mediation layer integrates and orchestrates these functions within a unified architectural framework.

Modern enterprise environments operate across fragmented, heterogeneous infrastructure. Microservice architectures, containerized workloads, and multi-cloud deployments have dissolved the network boundary that perimeter defense presupposes, multiplying inter-service communication channels and creating lateral movement surfaces that conventional controls cannot adequately monitor [3]. The average dwell time of advanced persistent threats in enterprise networks, measured from initial compromise to detection, has been documented at over two hundred days in threat intelligence analyses, a figure that points to structural limitations in alert-driven, reactive detection workflows [4].

Behavioral analytics reframes the detection question: the primary signal is deviation from established normal operation, which enables identification of novel threat vectors without prior knowledge of their specific form. Realizing this principle demands more than algorithm selection. It requires architectural coherence between data ingestion and model inference, between risk scoring and policy enforcement, and between automated response and human oversight.

This article examines those requirements, situating recent practitioner-research contributions within the scholarly literature that motivates and contextualizes them. Three interrelated bodies of work by Dazhyma Oiun provide the principal applied reference points: an architectural framework for autonomous, AI-driven security layers in distributed systems grounded in cross-proof verification and zero-trust enforcement [3]; a monographic treatment of cybersecurity automation spanning manual SOC operations through autonomous threat intelligence [4]; and a systematic methodological analysis of modern approaches to AI-driven intrusion detection, addressing evaluation validity, deep learning architectures, federated training, and adversarial robustness [5].

2. Literature Review

Buczak and Guven, surveying machine learning methods across the intrusion detection literature, documented a consistent pattern: supervised models achieve strong results on training datasets, then fail to generalize when the distribution shifts, which it does continuously in any live enterprise environment [1]. The response that emerged from this recognition was a turn toward methods whose primary signal is deviation from modeled normality, enabling detection of unknown threats without prior enumeration of their form. This shift from classification to anomaly detection reflects a substantive reorientation of what the detection problem is taken to be [5].

The benchmark validity problem warrants separate attention. Datasets such as KDD Cup 99, which remained in widespread use long after their limitations were documented, contain substantial redundancy, severe class imbalance, and leakage between training and test partitions, properties that inflate reported accuracy figures in ways that do not survive deployment [5]. More recent datasets attempt greater ecological validity but introduce their own modeling challenges: elevated feature dimensionality, device heterogeneity, and attack taxonomies that only partially capture the structure of real-world campaigns.

The theoretical foundation of behavioral analytics rests on a well-established result in anomaly detection theory: entities exhibit statistical regularities in their behavior, and deviations from those regularities carry information. The taxonomy offered by Chandola, Banerjee, and Kumar, distinguishing point anomalies, contextual anomalies, and collective anomalies, remains the most useful conceptual framework for understanding what kind of threat each detection approach is equipped to address [6]. Collective anomalies are anomalous by virtue of their sequential structure, not by the character of any individual component, and they represent the threat class most directly relevant to sophisticated enterprise intrusions: credential abuse, lateral movement, data staging, and exfiltration are each individually unremarkable events that become significant only when read as a sequence.

The detection of collective anomalies demands temporal modeling. Recurrent architectures and, more recently, transformer-based models have demonstrated the capacity to encode long-range dependencies within event sequences, capturing contextual relationships that static feature vectors cannot represent [5]. Graph neural network approaches extend this to relational structure: representing the enterprise network as a dynamic graph, where hosts, users, and services are nodes and authenticated flows are edges, enables the identification of lateral movement through anomalous traversal paths, that is, communication between entities that historically had no direct interaction [3]. Temporal and graph-based methods address different structural dimensions of the detection problem and are most effective when deployed in combination [7].

Self-supervised learning has emerged as a response to the practical impossibility of maintaining comprehensive labeled attack data in live environments. By training on pretext tasks derived from the unlabeled telemetry stream itself, these

methods develop representations sensitive to normal behavioral structure, enabling drift tolerance and reducing the annotation burden that supervised approaches require. Systems that adapt continuously to incoming data are also susceptible to corruption of learned representations by adversarially engineered streams, a concern addressed through robustness mechanisms including update gating and uncertainty-controlled adaptation policies [5].

The zero-trust paradigm, formalized by NIST in Special Publication 800-207, addresses a complementary dimension of the enterprise security problem [8]. Where behavioral analytics characterizes what is happening across the network, zero-trust governs whether any entity, regardless of network location, should be permitted to act as it is attempting. The operating principle of never trust, always verify eliminates the architectural assumption that internal entities are benign, which is precisely the assumption that lateral movement exploits. Continuous verification of all entities, in real time, across a dynamic enterprise environment, requires risk assessment capabilities that static policy systems are structurally incapable of providing.

The integration of machine learning risk scoring with zero-trust policy enforcement is a functional requirement of the architecture: continuous verification demands dynamic, behavior-sensitive risk signals that access control lists operating on static attributes cannot provide [9]. Dynamic, model-generated risk scores fed into policy enforcement engines operationalize the zero-trust principle at enterprise scale. The architectural question then becomes how to design a system in which model outputs translate reliably and accountably into enforcement decisions, a question that deployment-oriented research must answer in concrete terms [3].

SOC analysts face conditions, including persistent alert fatigue, skill shortages, and the cognitive demands of real-time threat assessment, that systematically degrade performance under the volume of alerts that AI detection generates [10]. Automation concentrates human attention on decisions that genuinely require it, such as contextual interpretation, escalation judgments, and decisions with significant operational consequences, while handling routine alert triage at machine speed [4].

The question of where to draw the boundary between automated and human decision-making has both technical and institutional dimensions. Technically, the reliability required of an automated decision depends on the consequences of a wrong decision, a threshold calibration problem that varies by deployment context. Institutionally, automated security decisions must be accountable and auditable, particularly under regulatory frameworks that impose obligations of transparency on organizational decision-making. These considerations motivate the policy-as-code paradigm with version control, staged deployment, and automated rollback, an architecture that preserves institutional accountability by making automated decisions transparent, reversible, and subject to human review at each escalation point [3].

3. Methodology

The article employs a thematic synthesis methodology [11]. The primary academic sources were identified through searches of IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar, supplemented by NIST technical publications. Search terms combined the following clusters: (*machine learning* OR *deep learning* OR *neural network*) AND (*intrusion detection* OR *anomaly detection* OR *threat detection*); (*zero trust* OR *zero-trust architecture*) AND (*enterprise security* OR *access control*); (*behavioral analytics* OR *behavioral baseline*) AND (*cybersecurity* OR *network security*); and (*security automation* OR *SOAR*) AND (*SOC* OR *security operations*). The search was limited to publications from 2009 onward; the lower bound was set by the publication of the Chandola, Banerjee, and Kumar anomaly detection taxonomy [6], which remains the theoretical reference point for much subsequent work. An initial pool of approximately 80 candidate sources was screened by title and abstract for relevance to at least one of the following: machine learning-based threat detection, behavioral analytics in enterprise environments, zero-trust architecture, AI-driven SOC automation, or automated policy enforcement in distributed systems. Sources reporting purely theoretical results without discussion of deployment constraints, and those whose experimental methodology relied exclusively on datasets documented as containing severe leakage or class imbalance without appropriate correction, were excluded. The final synthesis drew on 13 peer-reviewed and technical sources.

Three works by Dazhyma Oiun [3]-[5] are treated as primary applied references because they engage implementation-level problems, specifically benchmark validity in intrusion detection evaluation, SOC automation lifecycle design, and distributed policy enforcement architecture, that the academic literature identifies but rarely resolves in concrete architectural terms. Their inclusion supplements, rather than replaces, the peer-reviewed scholarship; claims derived exclusively from these sources are noted as such throughout the analysis.

Thematic synthesis organizes findings around conceptual structures and unresolved tensions in the field, forgoing a narrative inventory of individual studies. This approach is appropriate given the terminological heterogeneity and rapid empirical development that make simple narrative review inadequate. The principal limitation of the methodology is its reliance on experimental and benchmark-based evidence that may not transfer to adversarially contested production deployments, a constraint noted explicitly where it bears on the interpretation of specific findings.

4. Results

Across the studies reviewed, behavioral baseline modeling emerges as the most widely supported theoretical foundation for AI-driven enterprise security, though this characterization reflects a synthesis-based interpretation rather than settled consensus in the field. Supervised classification approaches retain advocates, par-

ticularly in contexts where labeled attack data is available and the threat distribution is relatively stable; signature-based methods remain operationally dominant in many deployed systems precisely because their failure modes are predictable. Behavioral approaches introduce their own documented problems: high false-positive rates on workloads with irregular but legitimate behavioral cycles, susceptibility to slow-drift poisoning by adversaries who gradually shift the learned baseline, and the operational cost of continuous model retraining. That said, whether the specific implementation involves statistical process control, unsupervised clustering, LSTM-based sequence modeling, or graph neural network approaches, the underlying logic of modeling normality and surfacing deviation from it has accumulated empirical support across independent research traditions in ways that classification-from-enumerated-attacks has not, and the convergence is substantive enough to function as a shared research platform [6] [7].

This convergence has been translated into architectural form in the design of behavioral threat signature engines that ingest continuous telemetry from heterogeneous sources, including authentication logs, API invocations, network flow data, and system-level process signals, normalize it across formats, and apply unsupervised clustering to generate threat signatures from statistically significant behavioral deviations [3]. Such engines generate signatures for novel, previously unseen attack patterns, a capacity that pre-enumerated attack taxonomies structurally cannot provide. The documented average signature generation latency of under thirty seconds from initial anomaly detection is operationally significant, since it determines whether the detection architecture is fast enough to matter in real incident timelines.

On benchmark performance, quantitative evidence is substantial but requires careful interpretation. Studies employing UNSW-NB15 and NSL-KDD datasets report detection accuracy in the range of 88 to 99 percent depending on attack category and architecture [12]. These figures depend critically on preprocessing decisions, class balance handling, temporal split construction, and the degree to which benchmark attack distributions match those encountered in deployment. A systematic analysis of these dependencies identifies contamination, class imbalance, and temporal leakage as the primary sources of inflated performance claims in the intrusion detection literature, a finding that establishes evaluation validity as a research concern of equal importance to algorithmic design [5].

Detection performance is necessary but not sufficient for operational security. Between a model that produces a risk score and an enterprise that is better protected lies a substantial gap: telemetry pipelines, alert contextualization, policy generation and enforcement, and feedback mechanisms that sustain model performance over time. Bridging this gap requires frameworks that treat the full pipeline as a unified design object, with each component designed in relation to the others.

Performance figures are drawn from the validation reported by Dazhyma Oiun [3]; the study does not describe production deployment. A concrete instance of

such a framework is the AI-Driven Adaptive Security Layer (AASL), developed by Dazhyma Oiun [3]. As illustrated in **Figure 1**, the AASL comprises four tightly integrated components. The Behavioral Threat Signature Engine (BTSE) handles continuous telemetry analysis and signature generation. The Machine-Learning Anomaly Detection Engine (MADE) performs real-time entity risk scoring through unsupervised, semi-supervised, and graph-based models with interpretable feature-attribution outputs. The Auto-Policy Enforcement Orchestrator (APEO) manages automated policy generation, staging, and deployment, while the Zero-Trust Re-Routing Engine (ZTRR) provides immediate traffic containment and redirection to sandbox inspection environments upon detection of high-risk activity. Automated policy generation from high-risk alert to full enforcement deployment across distributed nodes averaged eighteen seconds in empirical validation, several orders of magnitude faster than manual procedures requiring hours to days. Staged deployment with monitoring-mode evaluation preceding full enforcement identified false positives before production impact in 94 percent of test scenarios, and automated rollback completed within approximately eight seconds [3].

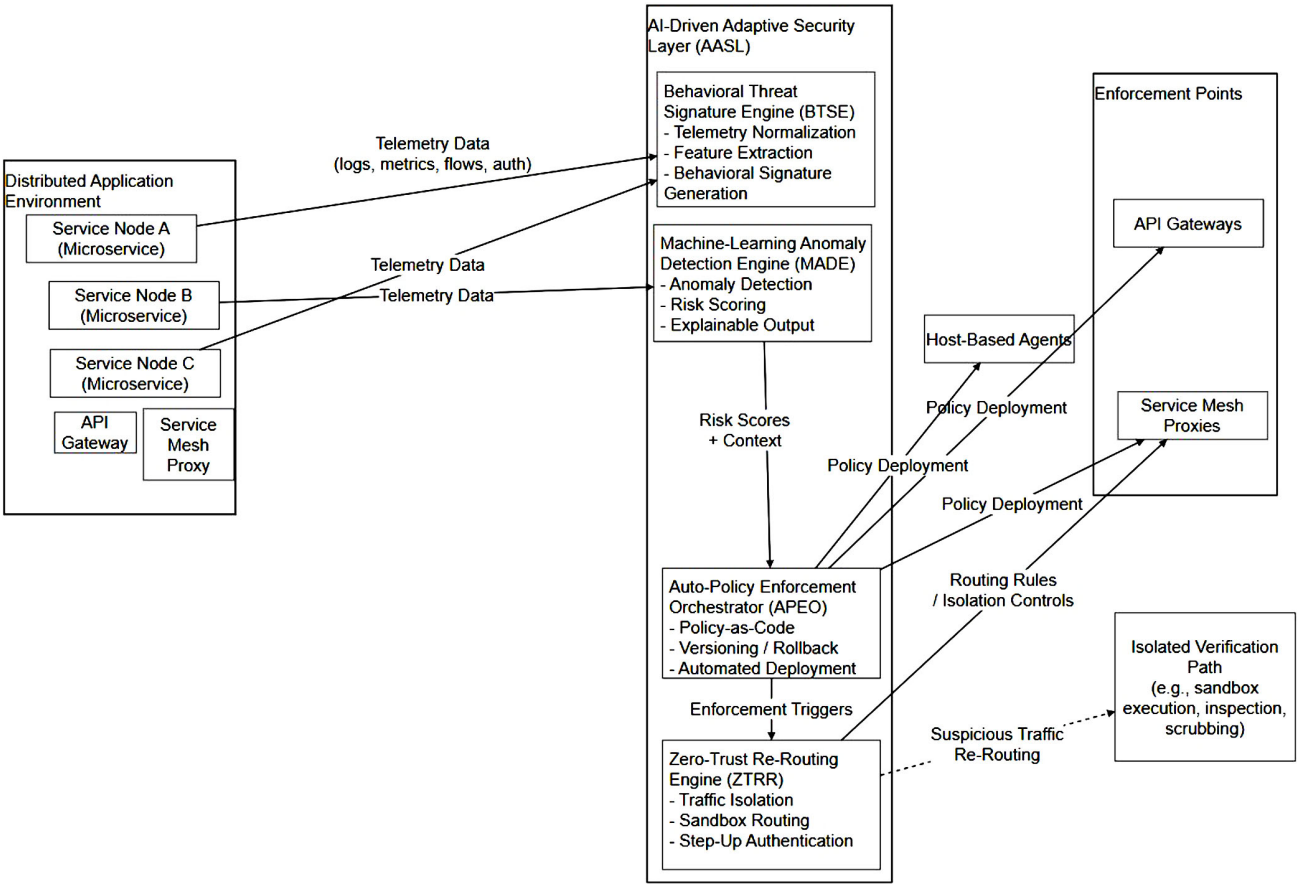

Figure 1. High-level architecture of the AI-Driven Adaptive Security Layer (AASL).

Policy consistency across distributed enforcement domains introduces a gov-

ernance challenge that centralized management cannot resolve. In multi-cloud and hybrid infrastructure deployments, security coverage gaps emerge at the boundaries between enforcement domains, a structural vulnerability arising from the absence of coordination mechanisms adequate to heterogeneous distributed environments. A cross-proof verification mechanism, combining cryptographic validation of policy artifacts with a distributed consensus protocol requiring quorum acknowledgment before deployment is considered complete, prevents both tampering during distribution and the partial deployment scenarios that create exploitable inconsistencies [3]. Scalability assessment documented sub-second response latencies across environments of up to five thousand nodes, with linear scaling in telemetry processing and logarithmic communication complexity in policy distribution.

Two directions that the literature has increasingly recognized as operationally necessary warrant documentation. Federated learning as a training paradigm for intrusion detection has gained practical relevance in enterprise contexts where regulatory constraints, administrative boundaries, or bandwidth costs preclude telemetry centralization. The challenges specific to security applications, including non-IID data distributions across federation participants, heterogeneous client hardware, variable participation, and the possibility of poisoned model updates, define the conditions of federated security deployments as a matter of course [5]. Cluster-based aggregation strategies and personalization layers offer partial mitigation, but the structural implication is that the training infrastructure must be considered part of the threat model.

Adversarial robustness constitutes the second direction. Unlike adversarial examples in computer vision, which are constrained by perceptual plausibility, adversarial perturbations of network traffic must simultaneously preserve protocol conformance and operational attack utility, a considerably tighter constraint that nevertheless leaves substantial room for evasion. Both evasion at inference time and poisoning at training time can materially degrade detection performance in ways that standard holdout evaluation will not detect [5].

5. Discussion

The results surveyed above reflect a field that has moved from methodological contestation toward architectural consolidation. Behavioral baseline modeling has accumulated sufficient empirical support across independent research traditions to function as a shared theoretical starting point, while the primary research frontier has shifted to the systems-level questions of how detection capability is sustained, governed, and integrated into operational environments.

Architectural design choices shape outcomes more decisively than the academic literature typically acknowledge, given its tendency to evaluate methods in isolation from the systems in which they are embedded. The quality of telemetry pipelines, the calibration of detection thresholds to organizational risk tolerance, the governance of model update cycles, and the design of human-machine interaction

patterns collectively determine whether a technically capable detection system produces real security improvements or merely generates alert volume. The empirical data from the AASL framework illustrates concretely what architectural integration means in operational terms: automated enforcement averages eighteen seconds from detection to deployment, while manual procedures require hours [3].

The governance dimension of AI security systems remains structurally underrepresented in academic literature relative to its operational importance. Zero-trust frameworks provide a policy logic within which AI-driven decisions can be embedded, but their implementation requires explicit structures specifying how model outputs become decisions, how decisions can be challenged or reversed, and how model performance is monitored over time [8]. In the architectural framework developed by Dazhyma Oiun, this requirement is addressed through a policy-as-code paradigm with staged deployment and automated rollback; the broader organizational conditions under which such automation delivers genuine security improvements are examined in the accompanying monographic treatment of SOC evolution [3].

Several deployment conditions bound the governance argument advanced here. First, the behavioral models discussed throughout assume telemetry of sufficient completeness, fidelity, and temporal regularity to produce stable baselines; in practice, log gaps from network appliances, inconsistent agent deployment, and clock skew across distributed sources degrade baseline quality in ways that inflate anomaly rates without producing actionable signals. Second, the performance figures reported for benchmark-evaluated systems and for the AASL lab validation may not transfer to adversarially contested production environments, where attack patterns are specifically engineered to evade detection and telemetry itself may be manipulated. Third, automated enforcement carries a category of risk absent from detection-only architectures: a false-positive enforcement decision can disrupt legitimate operations, and the 94 percent pre-production false-positive identification rate reported for the AASL still implies a residual 6 percent of false positives reaching full enforcement in test scenarios, a figure whose operational significance depends entirely on the volume of policy deployments in production. Fourth, adversarial drift represents a sustained attack surface: adversaries with knowledge of baseline update mechanisms can design campaigns that systematically erode detection sensitivity over time. These constraints do not invalidate the architectural approach, but they specify the operational conditions under which it delivers the security improvements its design intends.

One tension in the literature deserves continued attention: the relationship between autonomous response capability and the organizational conditions under which autonomous systems behave responsibly. Autonomous security architectures eliminate human latency from critical decision pathways, achieving faster response at the cost of reduced oversight. The appropriate balance depends on context, including threat severity, regulatory environment, and organizational

risk tolerance, and admits no universal calibration. What the architecture must provide is the capacity to set that calibration deliberately, making its consequences visible through audit trails, rollback mechanisms, and human review channels.

6. Conclusion

The transition from perimeter-centric to behavior-centric security involves a substantive reorientation in how the threat detection problem is defined and what counts as an adequate solution. The shift from signature-based detection to behavioral modeling, from static access controls to zero-trust continuous verification, from manual SOC operations to autonomous threat intelligence demands different tools, different architectural commitments, and different governance structures.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Buczak, A.L. and Guven, E. (2016) A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, **18**, 1153-1176. https://doi.org/10.1109/comst.2015.2494502

[2] Sommer, R. and Paxson, V. (2010) Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. *Proceedings of the IEEE Symposium on Security and Privacy*, Oakland, 16-19 May 2010, 305-316. https://doi.org/10.1109/sp.2010.25

[3] Oiun, D.A. (2026) Autonomous Security Layers for Global Distributed Systems: A Cross-Proof Architectural Framework. *Austrian Journal of Technical and Natural Sciences*, **3**-**4**, 106-111. https://doi.org/10.29013/AJT-26-3.4-106-111

[4] Oiun, D. (2026) Cybersecurity Automation: From Manual Defense to Autonomous Threat Intelligence. LAP Lambert Academic Publishing.

[5] Oiun, D. (2026) Modern Approaches to AI-Driven Intrusion Detection. In: *Proceedings of the* 78*th International Multidisciplinary Conference* "*Recent Scientific Investigation*", Primedia E-Launch, 113-124.

[6] Chandola, V., Banerjee, A. and Kumar, V. (2009) Anomaly Detection: A Survey. *ACM Computing Surveys*, **41**, 1-58. https://doi.org/10.1145/1541880.1541882

[7] Kwon, D., Kim, H., Kim, J., Suh, S.C., Kim, I. and Kim, K.J. (2019) A Survey of Deep Learning-Based Network Anomaly Detection. *Cluster Computing*, **22**, 949-961. https://doi.org/10.1007/s10586-017-1117-8

[8] Rose, S., Borchert, O., Mitchell, S. and Connelly, S. (2020) Zero Trust Architecture. NIST Special Publication 800-207. National Institute of Standards and Technology.

[9] Sarker, I.H., Furhad, M.H. and Nowrozy, R. (2021) AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. *SN Computer Science*, **2**, Article No. 173. https://doi.org/10.1007/s42979-021-00557-0

[10] Zimmerman, C. (2014) Ten Strategies of a World-Class Cybersecurity Operations Center. MITRE Corporation.

[11] Thomas, J. and Harden, A. (2008) Methods for the Thematic Synthesis of Qualitative

Research in Systematic Reviews. *BMC Medical Research Methodology*, **8**, Article No. 45. https://doi.org/10.1186/1471-2288-8-45

[12] Moustafa, N. and Slay, J. (2016) The Evaluation of Network Anomaly Detection Systems: Statistical Analysis of the UNSW-NB15 Data Set and the Comparison with the KDD99 Data Set. *Information Security Journal*: *A Global Perspective*, **25**, 18-31. https://doi.org/10.1080/19393555.2015.1125974