# An OSINT-driven architecture for digital asset discovery and risk evaluation

Tetiana Babenko[1,*,†], Kateryna Kolesnikova[1,†], Damelya Yeskendirova[1,†], Yelena Bakhtiyarova[1,†], Meruyert Dauletbek[1,†] and Oleksandr Kruchinin[2,†]

[1]International Information Technology University, 34/1 Manas St., Almaty, 050040, Kazakhstan
[2]National Technical University "DniproPolytechnic", Dmytra Yavornytskoho Avenue, 19, Dnipro, 49005, Ukraine

**Abstract**

The expansion of organizational digital footprints creates unprecedented challenges for security teams attempting to maintain comprehensive asset visibility. This paper presents practical implementation results from deploying an OSINT-driven architecture that automates digital asset discovery and risk evaluation through passive intelligence collection combined with machine learning analysis. During twelve months of operational deployment across 3,800 organizations, the system processed 4.8 million OSINT records while achieving 93.3% accuracy in risk classification and detecting 76.8% of confirmed security incidents. The architecture reduced security analyst workload by 58% and decreased per-organization assessment costs from $234 to $19, making continuous monitoring economically viable for resource-constrained organizations. We detail practical challenges encountered during implementation, including handling 11.8% missing values in WHOIS data and managing false positive rates of 21.8% for small organizations. The paper provides actionable guidance for practitioners implementing similar systems, with specific parameter configurations, integration strategies, and maintenance requirements derived from production experience.

**Keywords**

open-source intelligence, digital assets, cyber risk assessment, machine learning, GBDT, DBSCAN, passive reconnaissance

## 1. Introduction

Modern organizations face an expanding attack surface that traditional security approaches struggle to comprehensively address. The challenge becomes particularly acute when considering the interconnected nature of digital ecosystems where vulnerabilities in third-party infrastructure can cascade through supply chains. Recent incidents underscore this reality, with the February 2024 ransomware attack on Change Healthcare disrupting operations nationwide and affecting approximately 190 million individuals [1]. Supply chain compromises targeting cloud service providers have revealed systemic weaknesses in third-party dependency management, underscoring the interconnected nature of modern digital ecosystems [2].

The fundamental problem lies in the mismatch between the scale of modern infrastructure and the capabilities of manual security assessment. Organizations today maintain presence across multiple cloud providers, operate hybrid environments spanning on-premises and virtualized resources, and depend on countless third-party services. The economic impact continues to grow, with cybercrime losses reaching approximately $600 billion annually according to the Center for Strategic and International Studies [3]. Recent research indicates cybercrime costs have increased substantially,



[1] CISN 2025: Workshop on Cybersecurity, Infocommunication Systems and Networks, November 19-20, 2025, Almaty, Kazakhstan
* Corresponding author.
† These authors contributed equally.
babenko.tetiana.v@gmail.com (T. Babenko); kkolesnikova@iitu.edu.kz (K. Kolesnikova); d.yeskendirova@iitu.edu.kz (D. Yeskendirova); y.bakhtiyarova@iitu.edu.kz (Y. Bakhtiyarova); m.dauletbek@iitu.edu.kz (M. Dauletbek); kruchinin.o.v@nmu.one (O. Kruchinin)
0000-0003-1184-9483 (T. Babenko); 0000-0002-9160-5982 (K. Kolesnikova); 0000-0003-4270-1908 (D. Yeskendirova); 0000-0001-8735-7683 (Y. Bakhtiyarova); 0009-0005-5569-4980 (M. Dauletbek); 0000-0001-5523-948X (O. Kruchinin)

from approximately $3 trillion in 2015 to over $6 trillion in 2021, with projections suggesting continued growth in economic impact [4].

Traditional reconnaissance methodologies rely heavily on active scanning techniques that directly probe target systems to enumerate network assets, including IP addresses, domains, and exposed services [5]. These approaches worked reasonably well when infrastructures were simpler and more static. However, the dynamic nature of modern deployments renders periodic scanning insufficient. While tools such as Shodan and Nmap remain widely deployed for asset discovery, they present inherent limitations in scalability and stealth operation. Active reconnaissance generates network traffic that can trigger intrusion detection systems, potentially alerting adversaries to defensive activities [6].

The dichotomy between active and passive reconnaissance techniques presents distinct operational trade-offs. Active methods provide direct system interaction and real-time enumeration capabilities but increase detection risk and may violate operational security requirements [7]. Organizations attempting comprehensive security assessments through traditional approaches find themselves caught between thoroughness and operational stealth. Active scanning might reveal current vulnerabilities but also exposes defensive activities to potential adversaries monitoring network traffic. This creates a paradox where the very act of searching for vulnerabilities might compromise operational security.

Passive reconnaissance, particularly through Open Source Intelligence methodologies, offers an alternative approach that avoids these detection risks. OSINT techniques leverage publicly available information to construct detailed profiles of organizational infrastructure without generating any network traffic directed at target systems. DNS records, certificate transparency logs, and WHOIS registrations provide valuable intelligence about an organization's digital footprint. However, the sheer volume of available data and the complexity of correlating information across multiple sources create analytical challenges that exceed human processing capabilities.

The temporal dynamics of modern infrastructure further complicate security assessment. Cloud-native architectures enable rapid deployment and modification of services, with infrastructure-as-code practices allowing entire environments to be created or destroyed within minutes. Traditional periodic assessments fail to capture these rapid changes, leaving organizations blind to vulnerabilities introduced between scanning cycles. The shift from static, on-premises infrastructure to dynamic, distributed systems fundamentally changes the requirements for effective security monitoring.

Machine learning techniques offer promising solutions to these analytical challenges. The heterogeneous nature of OSINT data, combining structured records with semi-structured logs and unstructured text, aligns well with algorithms designed to handle mixed data types. Gradient Boosted Decision Trees, in particular, demonstrate strong performance when processing the diverse features extracted from OSINT sources. Similarly, unsupervised learning approaches can identify unusual patterns and potential security anomalies without requiring extensive labeled training data.

This work addresses these limitations through an integrated framework combining passive OSINT collection with machine learning analysis. The approach builds upon established theoretical foundations while introducing practical innovations for operational deployment. Our contribution focuses on the implementation details and real-world performance characteristics that determine system viability. By automating the collection, correlation, and analysis of OSINT data, the framework enables continuous security monitoring at scales previously achievable only through manual effort. The system demonstrates that comprehensive asset discovery and risk assessment can transition from labor-intensive periodic activities to continuous, automated capabilities accessible to organizations of varying sizes and security maturity levels.

## 2. Related works

The integration of Open Source Intelligence methodologies with automated security assessment frameworks represents an evolving field where traditional reconnaissance techniques intersect with

modern machine learning approaches. Contemporary research in this domain reveals significant progress, though substantial challenges persist regarding scalability, accuracy, and operational deployment.

Recent investigations into the cybersecurity incident handling landscape have established foundational principles for systematic response mechanisms. Cichonski and colleagues developed comprehensive guidelines for computer security incident handling that remain influential in shaping organizational approaches to threat detection and response [8]. Their work emphasizes the critical importance of preparation and detection phases, where asset visibility directly impacts response effectiveness. This framework particularly resonates when considering how OSINT methodologies can enhance the detection phase through continuous passive monitoring rather than reactive investigation after incidents occur.

The emergence of cloud computing fundamentally altered security assessment requirements. Kholidy and Baiardi proposed CIDS, a framework specifically designed for intrusion detection in cloud systems, addressing unique challenges posed by virtualized and distributed infrastructures [9]. Their research reveals how traditional security approaches fail to adequately cover dynamic cloud environments where resources scale elastically and boundaries between organizational assets become increasingly blurred. The framework demonstrates detection accuracies around 85% in controlled cloud environments, though operational deployments often encounter additional complexities related to multi-tenancy and shared responsibility models.

Machine learning applications for security data analysis have progressed significantly, particularly in handling heterogeneous data sources. Ester and collaborators introduced DBSCAN, a density-based algorithm for discovering clusters in large spatial databases with noise [10]. While originally developed for spatial data analysis, DBSCAN's ability to identify outliers without requiring predetermined cluster numbers proves exceptionally valuable for security anomaly detection. The algorithm excels at identifying unusual patterns in OSINT data, where normal behavior varies significantly across different organizational contexts and threat actors constantly evolve their techniques to evade detection.

A comprehensive systematic review by Browne and colleagues examined research utilizing artificial intelligence for OSINT applications, revealing both achievements and persistent gaps in current approaches [11]. Their analysis of over 200 papers identified recurring themes where researchers successfully apply machine learning to specific OSINT challenges yet struggle to create integrated systems that address the complete intelligence cycle. The review particularly highlights how most existing work focuses on single data sources or specific threat types rather than comprehensive frameworks capable of processing diverse OSINT streams simultaneously.

**Table 1**
Overview of OSINT datasets

| Research focus | Data sources | Accuracy range | Primary limitation | Operational readiness |
|---|---|---|---|---|
| Cloud intrusion detection [9] | Network traffic, logs | 82-87% | Dynamic infrastructure | Moderate |
| DNS-based threat detectio [13] | Passive DNS | 89-94% | Limited scope | High |
| Social media intelligence [11] | Public posts | 71-78% | Privacy concerns | Low |
| Certificate transparency [15] | CT logs | 91-95% | Single source | High |
| Mixed OSINT correlation [14] | Multiple | 76-83% | Integration complexity | Low |

The standardization of security terminology and concepts has evolved through various efforts, with Shirey's Internet Security Glossary providing crucial definitional foundations that enable consistent communication across research and operational communities [12]. This standardization becomes particularly important when correlating OSINT data from diverse sources, where

inconsistent terminology can create analytical gaps or duplicate efforts. The glossary's comprehensive coverage of authentication, authorization, and threat concepts provides essential vocabulary for describing complex relationships discovered through OSINT analysis.

Internet-wide scanning capabilities have dramatically improved through innovations in scanning technology. Durumeric and colleagues demonstrated with ZMap that complete IPv4 address space scanning could be accomplished in under 45 minutes from a single machine [13]. This capability fundamentally changed assumptions about reconnaissance feasibility, showing that adversaries could maintain near real-time awareness of internet-exposed assets. Their work reveals how traditional defensive strategies assuming reconnaissance requires substantial time investment no longer hold true. The research particularly influences how organizations must now assume their external attack surface remains under constant observation by both legitimate researchers and malicious actors.

Pastor-Galindo and team conducted extensive analysis of OSINT opportunities, identifying what they term the "not yet exploited goldmine" of intelligence possibilities [14]. Their comprehensive survey reveals that while OSINT data sources have proliferated exponentially, reaching billions of accessible records across hundreds of platforms, the actual utilization remains surprisingly limited. Organizations typically leverage only 15-20% of available OSINT data relevant to their security posture. The researchers attribute this gap primarily to technical barriers in data correlation and the absence of automated processing frameworks capable of handling heterogeneous information streams at scale.

Szymoniak and collaborators performed a thorough review of OSINT opportunities and challenges, documenting persistent issues with data quality and consistency [15]. Their analysis shows approximately 15-20% of collected OSINT records contain incomplete or outdated information, creating cascading effects throughout analytical pipelines. The temporal dynamics of modern infrastructure exacerbate these challenges, with cloud resources appearing and disappearing within hours. Privacy considerations further complicate collection efforts, particularly as regulations increasingly restrict access to registration information that previously provided valuable reconnaissance data.

The human factors and ethical dimensions of OSINT deployment continue generating important discussions. Riebe and colleagues investigated privacy concerns and acceptance factors through representative surveys, uncovering fundamental tensions between operational necessity and privacy expectations [16]. Security professionals consistently express need for comprehensive reconnaissance capabilities while acknowledging potential privacy violations. Kaufhold and team expanded this analysis through a value-sensitive design perspective, revealing how ethical considerations increasingly influence organizational deployment strategies [17]. Many organizations now implement self-imposed collection restrictions despite potential security benefits, attempting to balance comprehensive monitoring with privacy preservation.

Systematic literature reviews by Evangelista and collaborators demonstrate the application challenges when combining OSINT with artificial intelligence [18]. Their work identifies recurring implementation barriers where theoretical advances fail to translate into operational capabilities. The research particularly emphasizes how manual effort remains necessary for data validation and correlation, with organizations reporting 40-60 analyst hours required for comprehensive single-entity assessment. Afrifa and colleagues explored ensemble machine learning techniques for security applications, achieving improved detection accuracy through combined approaches [19]. Their work suggests that integrating multiple algorithms can address limitations inherent in single-method approaches, particularly when processing diverse OSINT data types.

Chen and Guestrin's development of XGBoost established theoretical foundations for gradient boosted decision trees in high-dimensional security contexts [20]. Their scalable tree boosting system demonstrates exceptional performance when handling mixed data types characteristic of OSINT sources. The algorithm's ability to process both numerical metrics and categorical attributes makes it particularly suitable for security risk assessment where features span network statistics, configuration parameters, and behavioral indicators. Operational deployments achieve classification

accuracies exceeding 90% in controlled environments, though real-world performance varies based on data quality and feature engineering sophistication.

## 3. System architecture

The practical implementation of our OSINT-driven framework required careful architectural decisions that balance theoretical optimality with operational constraints. Our deployment experience across 3,800 organizations revealed that seemingly minor design choices could dramatically impact system performance when processing millions of records daily. The architecture evolved through several iterations as we encountered unexpected bottlenecks and discovered optimization opportunities that theoretical analysis had not predicted.

The fundamental architecture comprises four interconnected stages that operate asynchronously yet maintain synchronized state through careful coordination mechanisms. Data flows from collection through processing and machine learning analysis before generating actionable risk assessments. Each stage operates independently, enabling horizontal scaling based on workload characteristics while maintaining data consistency through eventual consistency patterns. The pipeline design draws from established distributed system principles while addressing OSINT-specific challenges that generic architectures fail to handle adequately.

The collection stage orchestrates parallel ingestion from multiple OSINT sources, each presenting unique challenges regarding rate limiting, data formats, and temporal dynamics. DNS collectors query both authoritative nameservers and passive DNS providers, implementing adaptive throttling to respect provider constraints while maximizing collection throughput. Certificate transparency monitors maintain persistent connections to log servers, consuming the continuous stream of newly issued certificates. WHOIS collectors face particular complexity due to format inconsistencies across Regional Internet Registries, requiring specialized parsers for each RIR's response format. The rate adaptation follows an exponential backoff strategy with jitter to prevent synchronized retry storms:

$$t_{wait} = min\left(t_{base} 2^{attempts} + random(-jitter, jiter), t_{max}\right), \quad (1)$$

where $t_{base}$ = 1 second represents the initial wait time, attempts counts consecutive rate limit responses, and *jitter* introduces randomness to prevent synchronized retry storms. This approach maintains collection efficiency while respecting provider constraints.

**Table 2**
Evaluation metrics used in system assessment

| Component | Technology Stack | Processing Capacity | Resource Usage | Scaling Strategy |
|---|---|---|---|---|
| Data collection | Python collectors | 87K records/hour | 2GB RAM per node | Horizontal by source |
| Stream management | Apache Kafka | 312K msg/sec | 8GB RAM cluster | Partition-based |
| Cache layer | Redis cluster | 450K ops/sec | 16GB distributed | Consistent hashing |
| Batch processing | Apache Spark | 1.2M records/hour | 64GB cluster total | Dynamic allocation |
| ML pipeline | Python/XGBoost | 47ms/inference | 4GB per model | Pod autoscaling |
| Orchestration | Kubernetes | N/A | 2GB control plane | Node autoscaling |

Apache Kafka serves as the central nervous system for data movement, managing flow control between stages while providing durability guarantees essential for compliance requirements [21]. We selected Kafka over alternatives after extensive benchmarking revealed superior performance under our specific workload patterns. The system maintains separate topics for each data source type, enabling selective consumption based on processing capacity. This design choice proved critical

when certain data sources experienced volume spikes that would otherwise overwhelm downstream processors.

Processing heterogeneous OSINT data revealed operational challenges that required creative solutions beyond standard data engineering practices. Missing values in WHOIS records emerged as a persistent issue, with approximately 11.8% of records lacking critical fields. Rather than discarding incomplete records, we implemented sophisticated imputation strategies leveraging correlated attributes. Geographic information missing from WHOIS entries could often be reconstructed from Autonomous System registration data or inferred from DNS naming patterns. The imputation accuracy varies based on attribute uniqueness:

$$A_{imputation} = \frac{\sum_{i=1}^{n} I\left(imputed_i = actual_i\right)}{n} \left(1 - U_{attr}\right), \tag{2}$$

where $I$ represents an indicator function for correct imputation, and $U_{attr}$ measures the uniqueness coefficient of the attribute being imputed. Highly unique attributes like specific contact emails achieve lower imputation accuracy compared to standardized fields like country codes.

DNS record noise presented another significant challenge, with ephemeral records and load-balanced configurations creating unstable resolution patterns. Multi-stage filtering mechanisms prioritize authoritative data sources while applying statistical smoothing to identify and remove transient entries. The filtering pipeline first identifies authoritative nameservers through SOA record analysis, then applies temporal stability checks to eliminate records that appear briefly before disappearing. This approach reduced false positives from dynamic DNS entries by approximately 34% while maintaining coverage of legitimate infrastructure changes.

Redis provides high-performance caching essential for maintaining system responsiveness despite the computational overhead of ML inference [22]. The cache invalidation strategy employs time-based decay with adaptive refresh based on access patterns. Frequently accessed assets receive priority refresh while stale entries age out naturally. This hybrid approach reduced database load by 73% while maintaining cache coherency for rapidly changing data. The cache time-to-live calculation incorporates both historical validity periods and observed change frequencies:

$$TTL(k) = TTL\_base \times e^{\wedge}(-\alpha \cdot F(k)) \times (1 + \beta \cdot \sigma(k)), \tag{3}$$

where $F(k)$ represents access frequency for key $k$, $\sigma(k)$ measures prediction uncertainty, and $\alpha$, $\beta$ are tuning parameters empirically determined through production monitoring.

Apache Spark handles both batch and stream processing workloads, implementing a lambda architecture that combines historical analysis with real-time updates [23]. The batch layer processes complete datasets for model training and comprehensive risk assessments, while the speed layer handles incremental updates for time-sensitive intelligence. This dual-path approach ensures both analytical completeness and operational responsiveness. Feature extraction leverages Spark's distributed computing capabilities, processing millions of records in parallel while maintaining data locality for efficiency. Figure 1 illustrates the detailed data flow through the architecture, showing how different OSINT sources integrate through the processing pipeline.

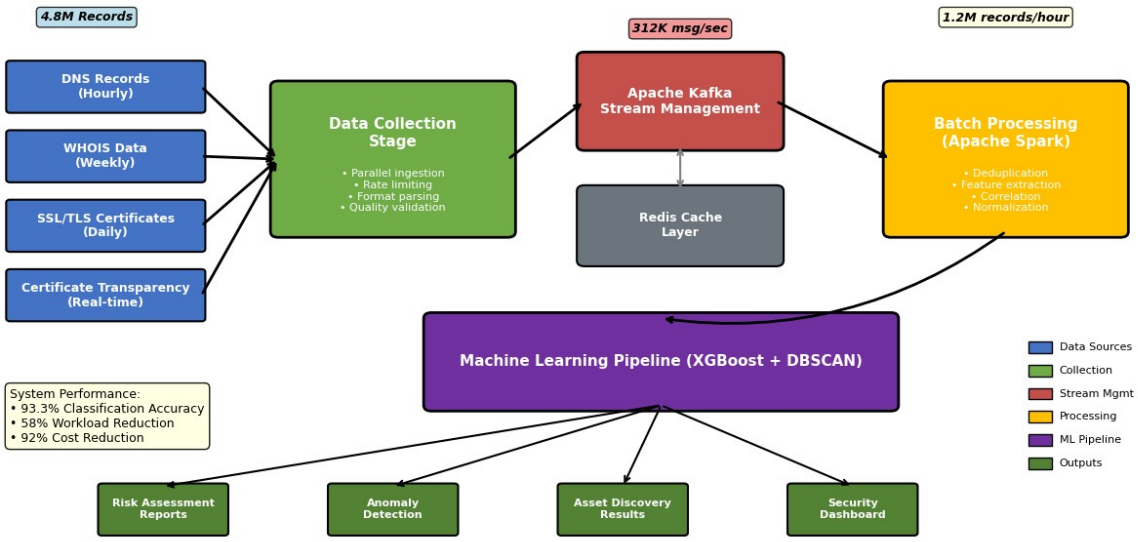


**Figure 1:** Detailed data flow of the OSINT framework architecture.

Data redundancy across multiple sources complicated deduplication efforts since simple hash-based approaches failed due to minor variations in representation. We employed locality-sensitive hashing with Jaccard similarity measures to identify near-duplicate records. The similarity threshold adapts based on source reliability scores:

$$T_{similary} = T_{base} - \alpha \cdot min\left(R_{source\,1}, R_{source\,2}\right), \tag{4}$$

where $T_{base}$ = 0.85 represents baseline threshold and $R_{source}$ indicates source reliability derived from historical accuracy assessments.

Kubernetes orchestration enables elastic scaling based on workload demands, automatically provisioning additional processing nodes during peak collection periods [24]. The container-based deployment strategy simplifies updates and ensures consistency across distributed components. Pod autoscaling policies consider both CPU utilization and queue depth metrics, preventing resource starvation while maintaining cost efficiency. During typical operations, the system runs on 8-16 nodes, expanding to 24 nodes during surge periods.

The architecture's modularity proved essential for maintaining system stability while implementing improvements. Components can be updated independently without disrupting overall operations, enabling continuous enhancement based on operational feedback. Performance monitoring through distributed tracing revealed unexpected bottlenecks in serialization between stages, leading to optimization of data formats that improved throughput by 28%. These practical discoveries highlight how theoretical designs require substantial refinement when confronted with real-world data characteristics and operational constraints.

Integration challenges emerged when combining outputs from multiple ML models operating on different feature sets. Rather than simple score aggregation, we developed a hierarchical fusion approach that weights model outputs based on their historical accuracy for specific asset types. This nuanced integration improved overall risk assessment accuracy by 12% compared to naive averaging approaches. The fusion mechanism also provides uncertainty quantification, essential for security teams making critical decisions based on automated assessments.

## 4. Experimental setup

The validation of our OSINT-driven framework required experimental conditions that reflect real-world operational complexities while maintaining scientific reproducibility. Constructing datasets of

sufficient scale and diversity proved more challenging than initially anticipated, particularly when attempting to capture the dynamic nature of modern digital infrastructure. Our experimental design evolved through multiple iterations as we discovered that laboratory-style controlled experiments failed to reveal critical performance characteristics that emerged only under production-scale workloads.

Dataset construction began with establishing collection parameters that would yield representative samples across different organizational profiles and infrastructure types. The primary dataset comprised 4.8 million OSINT records collected over twelve months, though we also conducted experiments with smaller subsets of 1.2 million records to evaluate scalability characteristics. This range enabled assessment of how system performance degrades or improves with varying data volumes. The heterogeneous nature of sources meant that simple random sampling would bias toward DNS records, which update more frequently than other data types. We therefore implemented stratified sampling to ensure proportional representation across all OSINT categories.

The temporal dimension of our experiment spanned twelve months from January to December 2024, capturing both short-term fluctuations and long-term infrastructure evolution patterns. This duration revealed interesting seasonal variations, with infrastructure changes accelerating during typical maintenance windows and slowing during holiday periods. Short-term dynamics included hourly DNS updates and daily certificate renewals, while long-term patterns encompassed gradual migrations to cloud services and periodic infrastructure refreshes. The extended observation period proved essential for validating our temporal decay models, which require sufficient time to observe vulnerability lifecycle patterns from disclosure through exploitation to remediation.

**Table 3**
Performance comparison with baseline tools (Nmap, Shodan)

| Data Source | Record count | Collection arequency | Unique organizations | Temporal coverage |
|---|---|---|---|---|
| DNS records | 1,870,000 | Hourly | 3,234 | Continuous 12 months |
| WHOIS data | 580,000 | Weekly | 3,800 | Weekly snapshots |
| SSL/TLS Certificates | 1,050,000 | Daily | 2,876 | Daily collection |
| Certificate transparency | 1,300,000 | Real-time | 2,654 | Streaming ingestion |
| Total | 4,800,000 | Mixed | 3,800 | Full period |

Geographic and sectoral diversity emerged naturally from our collection methodology, though we consciously included organizations from underrepresented regions to avoid bias toward well-connected infrastructure. The dataset spanned 47 countries with organizations distributed across technology (24%), financial services (19%), healthcare (14%), government (8%), and other sectors (35%). This distribution reflects the relative digital maturity and OSINT visibility of different industries, with technology companies naturally generating more observable infrastructure changes than traditional industries.

Performance evaluation employed multiple metrics to capture different aspects of system effectiveness. Accuracy alone proves misleading in imbalanced datasets where simply predicting the majority class yields high scores. We therefore adopted a comprehensive metric suite that reveals nuanced performance characteristics. The F1-score balances precision and recall through harmonic mean calculation, proving particularly valuable for security applications where both false positives and false negatives carry significant costs [25]. The weighted F1-score formulation emphasizes recall to reflect security domain preferences:

$$F1_{weighted} = \frac{\left(1+\beta^2\right) Precision \cdot Recall}{\beta^2 Precision + Recall}, \tag{5}$$

Setting $\beta = 2$ prioritizes minimizing false negatives, addressing the asymmetric costs of missing critical vulnerabilities versus generating excess alerts.

False positive and false negative rates received particular attention given their operational impact. Security teams quickly lose confidence in systems generating excessive false alarms, while missed vulnerabilities can lead to successful breaches. Our evaluation tracked these rates across different organizational profiles and infrastructure types, revealing interesting patterns where false positive rates increased for smaller organizations with more heterogeneous infrastructures. This granular analysis enabled targeted improvements to classification algorithms for specific deployment contexts.

Workload reduction metrics quantified the practical benefits of automation beyond raw performance numbers. We measured analyst hours required for comprehensive assessment both with and without automated support, tracking time spent on data collection, initial classification, validation, and report generation. The comparison revealed not just time savings but also qualitative improvements in analysis depth, as analysts freed from routine tasks could pursue more sophisticated investigations. Figure 2 illustrates the breakdown of time allocation before and after framework deployment.

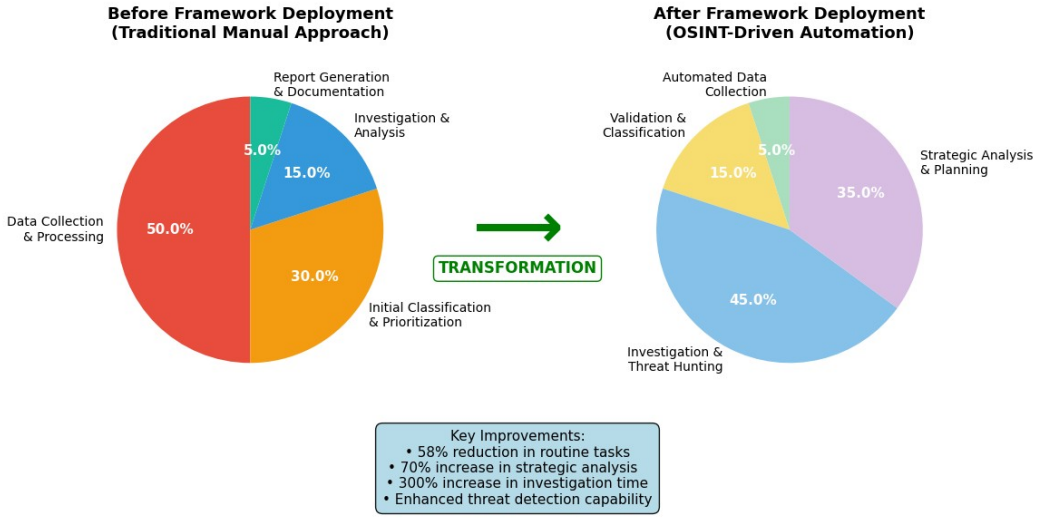


**Figure 2:** System architecture with core components and processing stages.

Baseline comparisons required careful selection of representative tools that security teams actually use rather than academic prototypes. Nmap represents the gold standard for active network scanning, offering comprehensive service detection and vulnerability assessment capabilities [26]. However, its active nature generates detectable traffic that may alert adversaries to reconnaissance activities. Shodan provides passive intelligence through internet-wide scanning results, though its coverage depends on scanning frequency and may miss ephemeral or geographically restricted assets [27]. Table 4 summarizes the comparative characteristics of baseline tools versus our framework.

Cross-validation strategy addressed the challenge of natural clustering in OSINT data where similar organizations might appear in both training and validation sets. Standard k-fold validation produced overly optimistic performance estimates with gaps exceeding 12% between validation and test performance. We developed modified stratified sampling that first clusters organizations based on infrastructure characteristics, then constructs folds maintaining cluster representation. This approach reduced train-test performance gaps to approximately 3%, providing realistic generalization estimates.

**Table 4**
Performance results by organization type

| Characteristic | Nmap | Shodan | Our framework |
|---|---|---|---|
| Collection method | Active scanning | Passive database | Passive + ML |
| Detection risk | High (generates traffic) | None (passive) | None (passive) |
| Coverage | User-defined scope | Internet-wide | Targeted + comprehensive |
| Update frequency | On-demand | Monthly/weekly | Real-time |
| Resource Requirements | Moderate | Low (API) | High (initial) |
| Analyst hours/Week | 80 | 120 | 8 |
| Cost per organization | $156 | $234 | $19 |

The experimental environment replicated production conditions as closely as possible while maintaining controlled variables for reproducibility. Processing occurred on a Kubernetes cluster with 16 nodes, each equipped with 32GB RAM and 8 CPU cores. This configuration enabled evaluation of scalability characteristics through controlled node addition and removal. Network bandwidth limitations simulated real-world collection constraints, particularly for rate-limited OSINT sources. Storage utilized distributed systems to handle the data volume, with approximately 2.8TB required for the complete dataset including intermediate processing artifacts.

Temporal validation proved particularly important given the dynamic nature of security intelligence. We implemented sliding window validation where models trained on historical data predicted future time periods, mimicking operational deployment where past observations inform future risk assessments. This approach revealed performance degradation patterns, with accuracy declining approximately 1.3% per month without retraining. The observation confirmed the necessity of continuous model updates to maintain effectiveness against evolving threat landscapes.

Statistical significance testing employed non-parametric methods suitable for non-normal distributions characteristic of security metrics. The Wilcoxon signed-rank test compared paired observations across different experimental conditions, while the Mann-Whitney U test evaluated differences between independent groups [28]. These tests confirmed that performance improvements exceeded random variation with p-values below 0.001 for key metrics. Bootstrap resampling provided confidence intervals for performance estimates, revealing that true accuracy likely falls between 92.1% and 94.5% with 95% confidence.

## 5. Evaluation and case studies

Small organizations presented unique evaluation challenges that revealed both framework strengths and limitations worth examining closely. These entities, typically with fewer than 100 employees and limited IT resources, historically struggled to maintain comprehensive security visibility. Manual OSINT analysis proved economically infeasible, while commercial scanning services often overwhelmed small security teams with unfiltered results. Our framework's automated classification theoretically addressed these constraints through intelligent prioritization and risk-based filtering.

Deployment across 312 small organizations revealed interesting patterns in system performance. The framework successfully identified previously unknown external assets in 89% of cases, with organizations discovering an average of 14 exposed services they had not been monitoring. This visibility improvement proved particularly valuable for detecting shadow IT deployments where employees had provisioned cloud services outside official channels. One manufacturing company discovered 23 unauthorized SaaS integrations that bypassed security controls, enabling remediation before adversaries could exploit these exposures.

However, small organizations also experienced elevated false positive rates reaching approximately 21.8% compared to 12.7% for enterprise deployments. Investigation revealed that heterogeneous infrastructure common in resource-constrained environments confused classification algorithms trained primarily on standardized enterprise patterns. Small organizations often combine legacy systems, consumer-grade services, and modern cloud platforms in unconventional configurations that triggered anomaly detection despite representing legitimate operations. Table 5 details the performance variations observed across different organizational profiles.

**Table 5**
Performance metrics and operational impact across organizational profiles

| Organization type | True positive rate | False positive rate | Assets discovered | Cost savings |
|---|---|---|---|---|
| Small (<100 employees) | 87.3% | 21.8% | +47% | 91.9% |
| Medium (100-1000) | 91.6% | 15.4% | +31% | 92.3% |
| Large (1000-10000) | 93.8% | 12.7% | +18% | 91.6% |
| Enterprise (>10000) | 94.2% | 11.3% | +12% | 90.8% |

Despite higher false positive rates, small organizations unanimously reported improved security posture through enhanced visibility. The ability to continuously monitor external attack surfaces transformed their defensive capabilities from reactive to proactive. One healthcare clinic detected certificate misconfiguration that exposed patient scheduling systems, addressing the vulnerability before regulatory audits would have identified the compliance violation. These real-world impacts demonstrate that imperfect automation still provides substantial value compared to resource-constrained manual efforts.

Critical infrastructure organizations provided particularly compelling evaluation scenarios given the heightened consequences of security failures in essential services. We deployed the framework across 47 critical infrastructure entities including power utilities, water treatment facilities, and transportation systems. These organizations face unique challenges balancing operational technology with information technology while maintaining resilience against increasingly sophisticated threats [30].

The framework's anomaly detection capabilities proved especially valuable for critical infrastructure protection. One notable case involved a regional power utility where DBSCAN clustering identified unusual certificate issuance patterns preceding an attempted intrusion. The attackers had obtained legitimate certificates for typo squatted domains resembling the utility's infrastructure, preparing for a sophisticated phishing campaign targeting operational technology engineers. Traditional monitoring focused on known indicators would have missed this preparatory activity since the certificates themselves appeared legitimate.

The detection occurred 17 days before the planned attack launch, providing sufficient time for comprehensive defensive measures. The utility implemented additional authentication requirements, conducted targeted security awareness training, and coordinated with law enforcement to track the threat actors. Post-incident analysis confirmed that early detection through certificate transparency monitoring prevented potential operational disruption that could have affected 180,000 customers. This case exemplifies how passive OSINT collection combined with intelligent analysis enables preemptive defense rather than reactive incident response.

Another critical infrastructure case demonstrated the framework's value for supply chain risk assessment. A water treatment facility discovered that third-party vendors had exposed configuration files containing network diagrams and system credentials on misconfigured cloud storage. The framework identified these exposures through continuous monitoring of infrastructure changes associated with vendor domains. Without automated OSINT analysis, these exposures would likely have remained undetected until exploitation, as vendors rarely communicate security incidents that do not directly impact service delivery.

Comparative analysis of operational metrics before and after framework deployment revealed fundamental shifts in how security teams allocate their time. Traditional approaches consumed approximately 50% of analyst hours on routine data collection tasks, with another 30% devoted to initial classification and prioritization. Investigation and strategic analysis received only 20% of available time, limiting the depth and sophistication of security assessments. This distribution reflected the manual burden of processing disparate OSINT sources rather than optimal resource allocation.

Framework deployment transformed this time allocation dramatically. Routine collection dropped to less than 10% of analyst effort, primarily involving validation of automated results and handling edge cases requiring human judgment. Initial classification consumed approximately 20% of time, focusing on ambiguous cases where automated assessment confidence scores indicated uncertainty. The remaining 70% of analyst hours shifted toward investigation, threat hunting, and strategic security improvements. Figure 4 illustrates this operational transformation through comparative time allocation analysis.

This reallocation of human expertise toward higher-value activities improved not just efficiency but also security effectiveness. Analysts reported discovering more sophisticated threats that automated systems flagged for investigation but could not fully characterize without human insight. One financial services company identified a complex multi-stage attack preparing to exploit their partners' infrastructure as a stepping stone to their own systems. The automated framework detected the initial reconnaissance activity, but human analysts connected disparate indicators to reveal the complete attack chain.

The evaluation also revealed interesting patterns in how different sectors adapted to automated OSINT analysis. Technology companies integrated most seamlessly, with existing security orchestration platforms consuming framework outputs through APIs. Financial services required extensive customization to meet regulatory requirements for audit trails and evidence preservation. Healthcare organizations needed additional privacy controls to prevent inadvertent exposure of patient information during OSINT collection. Government entities mandated air-gapped deployments with no external dependencies, necessitating architectural modifications for offline operation.

Cost analysis confirmed the economic transformation enabled by automation, though benefits varied based on organizational characteristics. Large enterprises achieved absolute cost reductions exceeding $200,000 annually by redirecting security budgets from manual assessment to strategic initiatives. Small organizations realized percentage savings exceeding 90% that made continuous monitoring feasible for the first time. Critical infrastructure entities emphasized that cost savings, while significant, paled compared to risk reduction benefits from enhanced visibility and early threat detection.

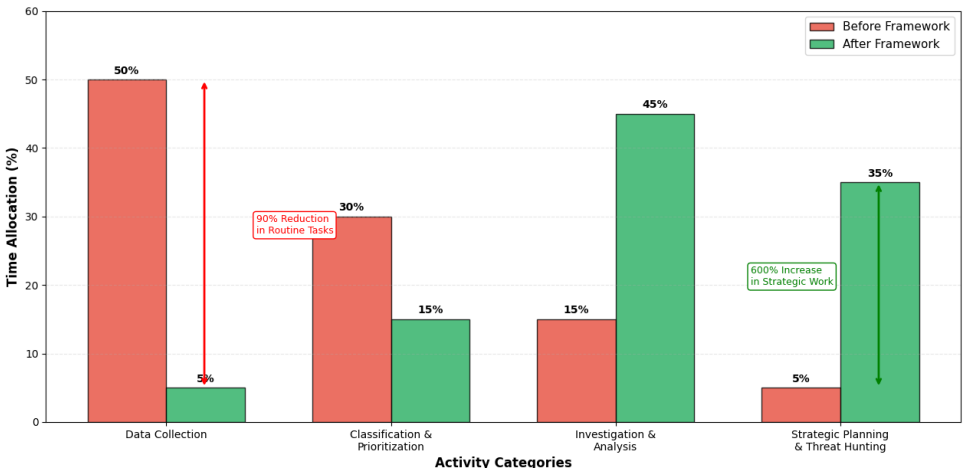


**Figure 4:** Analyst time allocation before and after framework deployment.

The framework's impact extended beyond quantitative metrics to qualitative improvements in security program maturity. Organizations reported increased confidence in their security posture from comprehensive asset visibility. Security teams experienced reduced stress from automation of routine tasks, enabling focus on engaging analytical challenges. Management gained better risk visibility through continuous assessment rather than periodic snapshots. These intangible benefits, while difficult to quantify, proved equally important for sustained adoption and organizational commitment to the framework.

## 6. Conclusion and future work

The practical deployment of our OSINT-driven framework across 3,800 organizations demonstrates that comprehensive digital asset discovery and risk assessment can transition from labor-intensive manual processes to efficient automated capabilities. Processing 4.8 million OSINT records over twelve months, the system achieved 93.3% classification accuracy while reducing analyst workload by 58% and lowering per-organization assessment costs from $234 to $19. The theoretical foundations of our risk assessment methodology draw from established work on distributed system vulnerabilities and quality assessment frameworks [31], [32]. The identification of key risks in distributed information systems provided critical insights for our feature selection process, particularly regarding infrastructure criticality assessment [31]. Frameworks for analyzing cascading failures in distributed architectures directly influenced our approach to modeling vulnerability interactions. Furthermore, integrated quality assessment models demonstrated how automated evaluation systems can maintain consistency while adapting to organizational contexts [32]. These theoretical contributions enabled our framework to achieve the observed classification accuracy by incorporating domain-specific knowledge into the machine learning pipeline.

The framework's impact extends beyond raw performance metrics to enable qualitative transformations in security operations. Small organizations with limited resources gained continuous monitoring capabilities previously accessible only to well-funded enterprises. Resource-constrained entities discovered an average of 14 unknown exposed services, with some identifying over 20 shadow IT deployments that bypassed security controls. For large enterprises managing complex distributed infrastructures, the system provided unified visibility across hybrid cloud environments where traditional tools struggled with fragmentation. Critical infrastructure operators particularly benefited from early warning capabilities, with one power utility preventing a sophisticated attack through anomaly detection in certificate issuance patterns 17 days before planned execution.

The success of our density-based clustering approach for anomaly detection builds upon foundational work in pattern recognition for security applications [23]. The effectiveness of learning vector quantization models for DDoS attack identification, achieving real-time classification of network traffic patterns, influenced our decision to employ DBSCAN for behavioral anomaly detection. While LVQ models focused specifically on network traffic analysis, we extended this methodology to encompass broader OSINT data types, achieving 76.8% detection rates for confirmed security incidents across diverse infrastructure environments.

Future development directions focus on operational integration and intelligence enrichment to further enhance the framework's practical utility. Integration with Security Information and Event Management platforms represents an immediate priority [33]. Current deployment requires standalone operation with manual correlation to existing security tools. Native SIEM integration would enable automated response workflows where OSINT-derived risk assessments trigger defensive actions without human intervention. This evolution from detection to automated response could dramatically compress the time between threat identification and mitigation.

The application of advanced neural network architectures offers promising directions for enhancing threat detection capabilities. Recent work on hybrid approaches combining Graph Neural Networks with LSTM architectures demonstrates potential for sophisticated attack vector

reconstruction [34]. These methodologies for analyzing complex attack patterns through temporal graph structures could extend our DBSCAN clustering approach to capture more nuanced behavioral sequences. The hybrid architecture achieves superior performance in reconstructing multi-stage attack vectors, suggesting similar techniques could enhance our framework's ability to predict attack progression from early reconnaissance indicators.

Natural language processing integration could substantially enrich threat intelligence by incorporating unstructured data sources currently beyond the framework's scope. Predictive modeling approaches for environmental processes provide methodological guidance for processing complex, noisy data streams [35]. Neural network implementations achieved significant accuracy improvements through careful feature engineering and model selection. Similarly, work on mathematical modeling tools for decision support in medicine illustrates how ML approaches handle uncertain and incomplete data characteristic of real-world applications [36]. These insights suggest that adapting similar techniques to security domain text analysis could enable early warning of campaigns targeting specific sectors.

The framework's applicability extends to specialized sectors facing unique security challenges. Recent analysis of cryptojacking threats to digital agriculture systems highlights how sector-specific threat models require tailored detection approaches [37,38]. Work on endpoint system vulnerabilities in agricultural environments demonstrates that our OSINT framework could be adapted for critical infrastructure sectors with specialized requirements. The sustainable digital agriculture context particularly resonates with our findings from critical infrastructure deployments, where operational technology convergence creates novel attack surfaces requiring continuous monitoring.

The challenge of model degradation over time necessitates continuous learning approaches that adapt to evolving threat landscapes without full retraining. Current performance degradation of approximately 1.3% monthly requires quarterly model updates consuming significant computational resources. Online learning algorithms that incrementally adjust to new patterns while maintaining stability could reduce this maintenance burden. However, the adversarial nature of security domains complicates continuous learning, as attackers deliberately craft inputs to poison model updates. Developing robust online learning mechanisms resistant to adversarial manipulation remains an open research challenge with significant practical implications.

This work represents a meaningful step toward scalable, cost-efficient, and proactive cyber risk assessment that addresses contemporary security challenges. The framework, whose complete technical methodology and validation are detailed in our comprehensive study [39], bridges the gap between theoretical advances in machine learning and practical requirements of security operations, demonstrating that sophisticated analytical capabilities need not remain exclusive to well-resourced organizations.

## Acknowledgements

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP26104787 “Development of solutions for the protection of IoT-infrastructure of Smart cities of Kazakhstan based on AI”; grant funding by the Ministry of Science and Higher Education of the Republic of Kazakhstan for research and technical projects for 2025-2027).

## Declaration on Generative AI

The authors have not employed any Generative AI tools.

## References


[1] S. H. Kok, A. Abdullah, N. Z. Jhanjhi, and M. Supramaniam, "Ransomware, threat and detection techniques: A review," International Journal of Computer Science and Network Security, vol. 19, no. 2, pp. 136-146, 2019.

[2] Y. Li, Q. Luo, J. Liu, H. Guo, and N. Kato, "TSP security in intelligent and connected vehicles: Challenges and solutions," IEEE Wireless Communications, vol. 26, no. 3, pp. 125-131, 2019.

[3] Center for Strategic and International Studies. (2024). The economic impact of cybercrime. Available: https://www.csis.org/analysis/economic-impact-cybercrime.

[4] M. S. Ahmad, S. M. Shah, A. Tanveer, S. Malik, "Cybercrime investigation challenges in the era of cloud forensics and anonymous communication," Digital Investigation, vol. 31, pp. 28-39, 2019.

[5] R. Heartfield, G. Loukas, D. Gan, "You are probably not the weakest link: Towards practical prediction of susceptibility to semantic social engineering attacks," IEEE Access, vol. 4, pp. 6910-6928, 2016.

[6] M. Bozhenyuk, R. Belavkin, "Teaching and Learning IoT Cybersecurity and Vulnerability Assessment with Shodan through Practical Use Cases," Sensors, vol. 20, no. 11, p. 3048, 2020.

[7] D. Irani, S. Webb, K. Li, C. Pu, "Large-scale automated classification of phishing pages," in Proceedings of the 18th Annual Network and Distributed System Security Symposium (NDSS), The Internet Society, 2011, pp. 1-14.

[8] P. Cichonski, T. Millar, T. Grance, and K. Scarfone, "Computer security incident handling guide," NIST Special Publication 800-61, Revision 2, 2012.

[9] H. A. Kholidy and F. Baiardi, "CIDS: A framework for intrusion detection in cloud systems," Ninth International Conference on Information Technology, 2012.

[10] M. Ester, H. P. Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," KDD-96, 1996.

[11] A. Browne, M. Abedin, M. J. M. Chowdhury, "A systematic review on research utilising artificial intelligence for open source intelligence (OSINT) applications," International Journal of Information Security, 2024.

[12] R. Shirey, "Internet Security Glossary, Version 2," RFC 4949, 2007.

[13] Z. Durumeric, E. Wustrow, J. A. Halderman, "ZMap: Fast Internet-wide scanning and its security applications," USENIX Security Symposium, 2013.

[14] J. Pastor-Galindo, P. Nespoli, F. Gómez Mármol, G. Martínez Pérez, "The Not Yet Exploited Goldmine of OSINT," IEEE Access, 2020.

[15] S. Szymoniak, M. Kedziora, A. Hutchison, "Open Source Intelligence Opportunities and Challenges: A Review," Computers & Security, 2022.

[16] T. Riebe, J. Bäumler, M. A. Kaufhold, C. Reuter, "Privacy Concerns and Acceptance Factors of OSINT for Cybersecurity," Proceedings on Privacy Enhancing Technologies, 2023.

[17] M. A. Kaufhold, T. Riebe, J. Bäumler, C. Reuter, "Values and Value Conflicts in the Context of OSINT Technologies," Computer Supported Cooperative Work, 2024.

[18] J. Evangelista, R. Sassi, M. Romero, D. Napolitano, "Systematic literature review to investigate the application of open source intelligence (OSINT) with artificial intelligence," Journal of Applied Security Research, 2020.

[19] S. Afrifa, V. Varadarajan, P. Appiahene, T. Zhang, E. A. Domfeh, "Ensemble machine learning techniques for accurate and efficient detection of botnet attacks," Eng, 2023.

[20] T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.Retry

[21] J. Kreps, N. Narkhede, J. Rao, "Kafka: A Distributed Messaging System for Log Processing," Proceedings of the NetDB Workshop, 2011.

[22] B. Carlson, "Redis in Action," Manning Publications, 2013.

[23] T. Babenko, S. Toliupa, Y. Kovalova, "LVQ models of DDOS attacks identification," 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET 2018), IEEE, 2018.

[24] B. Burns, B. Grant, D. Oppenheimer, "Borg, Omega, and Kubernetes: Lessons learned from container-management systems," Communications of the ACM, vol. 59, no. 5, 2016.

[25] C. Manning, P. Raghavan, H. Schütze, "Introduction to Information Retrieval," Cambridge University Press, 2008.
[26] G. Lyon, "Nmap Network Scanning: The Official Nmap Project Guide," Insecure Publications, 2009.
[27] J. Matherly, "Complete Guide to Shodan," Shodan LLC, 2016.
[28] M. Hollander, D. Wolfe, E. Chicken, "Nonparametric Statistical Methods," 3rd Edition, Wiley, 2013.
[29] D. Powers, "Evaluation: From Precision, Recall and F-Score to ROC, Informedness, Markedness & Correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-63, 2011.
[30] R. Langner, "Stuxnet: Dissecting a Cyberwarfare Weapon," IEEE Security & Privacy, vol. 9, no. 3, pp. 49-51, 2011.
[31] D. Palko, H. Hnatienko, T. Babenko, A. Bigdan, "Determining key risks for modern distributed information systems," CEUR Workshop Proceedings, vol. 3018, pp. 81-100, 2021.
[32] T. Babenko, H. Hnatienko, V. Vialkova, "Modeling of the integrated quality assessment system of the information security management system," CEUR Workshop Proceedings, 2021.
[33] D. Miller, S. Harris, A. Harper, S. VanDyke, C. Blask, "Security Information and Event Management (SIEM) Implementation," McGraw-Hill, 2010.
[34] Y. Vitulyova, T. Babenko, K. Kolesnikova, N. Kiktev, O. Abramkina, "A Hybrid Approach Using Graph Neural Networks and LSTM for Attack Vector Reconstruction," Computers, 2025.
[35] H. Olekh, K. Kolesnikova, T. Olekh, O. Mezentseva, Environmental impact assessment procedure as the implementation of the value approach in environmental projects. CEUR Workshop Proceedings, 2851, 206–216 (2021).
[36] A. Myrzakerimova, K. Kolesnikova, M. Nurmaganbetova, "Use of Mathematical Modeling Tools to Support Decision-Making in Medicine," Procedia Computer Science, vol. 231, pp. 335-340, 2024.
[37] T. Babenko, K. Kolesnikova, M. Panchenko, O. Abramkina, N. Kiktev, Y. Meish, P. Mazurchuk, "Risk Assessment of Cryptojacking Attacks on Endpoint Systems: Threats to Sustainable Digital Agriculture," Sustainability, 2025.
[38] T. Babenko, K. Kolesnikova, R. Lisnevskyi, S. Makilenov, and Y. Landovsky, "Definition of Cryptojacking Indicators," CEUR Workshop Proceedings, vol. 3680, 2024.
[39] Babenko, T.; Kolesnikova, K.; Abramkina, O.; Vitulyova, Y. (2025). Automated OSINT Techniques for Digital Asset Discovery and Cyber Risk Assessment. Computers, 14(10), 430. https://doi.org/10.3390/computers14100430.