Journal of Business, Social and Technology
Volume 7, Issue 2, 183-197
e_ISSN: 2807-6362
https://bustechno.polteksci.ac.id/index.php/jbt/index
DOI: doi.org/10.59261/jbt.v7i2.605

Comparative Analysis of Supervised Learning and Unsupervised Anomaly Detection in Security Log Analysis for Post-Incident Digital Forensic Investigation

***Iwan Indramana[1]**
STMIK Indonesia Mandiri, Indonesia

Asto Purwanto[2]
STMIK Indonesia Mandiri, Indonesia

***Corresponding author:**
Iwan Indramana, STMIK Indonesia Mandiri, Indonesia. ✉i1indramana@gmail.com

Article Info :

Article history:
Received: March 05, 2026
Revised: April 15, 2026
Accepted: April 18, 2026

Keywords:
anomaly detection; digital forensics; logistic regression; machine learning; security log analysis

Abstract

Background: Attempts to perform post-incident digital forensic investigation on large-scale security logs generated by enterprise firewalls and servers introduce a range of challenges. As data grows larger and more complex, it is no longer feasible to conduct manual analysis. Methodologically, there has been only limited empirical work directly comparing supervised and unsupervised paradigms for use in a post-incident forensic framework on operational-scale, real-world logs.
Objective: This paper compares the classification performance of supervised and unsupervised machine learning methods for forensic analysis of security logs, as well as the prioritization of various security anomalies using both approaches.
Methods: Analysis of a dataset containing more than 359,000 firewall and server logs obtained over a 30-day period. Labeled events were used to implement a supervised model, Logistic Regression; Isolation Forest is an unsupervised anomaly detection method, which performs best among the models trained on normal baseline logs. Evaluation metrics included accuracy, precision, recall, ROC-AUC, and ranking-based anomaly assessment.
Results: Logistic Regression — accuracy (0.99), ROC-AUC (0.9998), precision/recall for suspicious events (1.00, 0.99) — demonstrated near-perfect discriminability of labeled behavioral features within a 24-hour period. Isolation Forest: 86% overall accuracy, 93% precision, 59% recall; excellent forensic triage property: confirmed suspicious events among the top 200 anomaly-ranked entries: 197 of 200 (92.5%). Sensitivity analysis of the contamination parameter showed that ranking precision at the top 200 remained stable within the 0.05 to 0.30 range (Fig. 7A, 7B), demonstrating the robustness of rank-based prioritization despite variability in global recall across contamination values.
Conclusion: Our results demonstrate high predictive performance for supervised classification and efficient forensic triage through low false-positive rates in unsupervised anomaly detection of both time-series logs and free-text security event logs.

To cite this article: Indramana, I., & Purwanto, A. (2026). Comparative analysis of supervised learning and unsupervised anomaly detection in security log analysis for post-incident digital forensic investigation. *Journal of Business, Social and Technology, 7*(2), 183–197. https://doi.org/10.59261/jbt.v7i2.605

INTRODUCTION

Rather than being extraordinary technical failures, cybersecurity breaches have become operational realities embedded in modern organizational infrastructures. Firewalls and servers churn out security logs at a steady rate, creating hundreds of thousands of entries within a few hours. These logs contain highly detailed traces of who accessed the system, what configuration changes were made, and even failed communication attempts that can serve as vital forensic

evidence in a post-mortem. Despite this evidential richness, manual log analysis is inherently unscalable — the sheer volume of logs, combined with high noise-to-signal ratios and semantic complexity, creates a bottleneck in forensic reconstruction and risk assessment. Machine learning provides a scalable solution to address this (Algarni et al., 2021; He et al., 2020; Pengl & Li, 2022; Shaikh & Siponen, 2023).

As a remedy for the scalability challenges, machine learning has been proposed as a viable mechanism. In the field of cybersecurity research, supervised learning models have proven effective in classifying known attack types when labeled datasets are available (Sharafaldin et al., 2018; Vinayakumar et al., 2019). Logistic Regression still finds significant coverage across different use cases due to its intuitive interpretability compared to many black-box algorithms and its stable performance in structured classification problems. Alternatively, unsupervised anomaly detection approaches are used to identify deviations from normal behavioral patterns without the need for predefined attack labels (Goldstein & Uchida, 2016; Liu et al., 2008).

The contrast between learning a range of labeled classes versus identifying anomalous deviations continues to surface in the literature associated with intrusion detection systems. Studies comparing different machine learning algorithms for intrusion detection have indicated that both the characteristics of the dataset and approaches to feature engineering can impact classification performance (Buczak & Guven, 2015; Garcia-Teodoro et al., 2009).

Concurrently, unsupervised anomaly detection methods such as Isolation Forest strive to recognize deviations in normal behavior without requiring labeled attacks (Liu et al., 2008). Isolation-based methods like Isolation Forest are structurally distinct from density-based and distance-based techniques such as Local Outlier Factor Breunig (2000) and statistical approaches (Chandola et al., 2009). More recent expansions of the Isolation Forest approach, such as Extended Isolation Forest Hariri (2019), continue to improve robustness in more complex, high-dimensional feature spaces that can arise from operational log data.

Beyond isolation-based approaches, density-based outlier detection algorithms such as Local Outlier Factor (LOF) have been extensively explored as foundational methods for detecting local anomalies in data distributions (Breunig et al., 2000). More recent comprehensive treatments of outlier analysis have further categorized anomaly detection approaches into different paradigms, broadly separating them into statistical, proximity-based, and isolation-based families based on the mechanistic differences in how data points are scored as anomalies (Aggarwal, 2016).

Unlike individual research works focused on real-time intrusion detection cases with benchmark datasets (NSL-KDD and CICIDS) such as those in Sharafaldin (2018), while standardized comparisons are provided by such datasets, they do not necessarily approximate the contextual complexities found in operational firewall and server logs generated in production environments — logs that are at rest and produced by computers actively used by human beings. Related works on log-based anomaly detection reflect a recent surge of interest in this area.

General surveys of the topic characterize deep learning for anomaly detection Pang (2021) and highlight this methodological diversity in cybersecurity analytics. Deep anomaly detection has been characterized in general surveys of the topic, including Pang (2021), However most of these studies are directed toward reliability monitoring or online detection as opposed to post-incident forensic reconstruction. The considerable heterogeneity, noise, and context dependency in real-world logs means that high classification accuracy does not necessarily translate into investigative utility.

Digital forensic investigation, however, has a distinct analytical goal. Rather than limiting the scope to discovering suspicious or potentially malicious traffic, its goal is to reconstruct the sequence of events Casey (2011), determine probable causes through context-weighing constraints, and provide an evidentiary basis from which causation in an incident may be reasoned. In this context, anomaly detection is part of a broader analytical initiative where tools are required to assist investigators in more effectively identifying meaningful deviations within high-volume and complex log data.

Structured evidence acquisition, preservation, and event reconstruction processes are hallmarks of digital forensic methodologies that support investigative reasoning (Casey, 2011).

Within this framework, analytical tools must not only detect anomalies but also aid in contextual interpretation and evidence prioritization.

While some interest has recently been generated at the machine learning–digital forensics interface Nayerifard (2023), implementation in practice remains methodologically heterogeneous. Specifically, very little work exists that directly contrasts supervised versus unsupervised paradigms within an explicit post-incident forensic context using operational-scale data.

Supervised approaches such as Logistic Regression require labeled datasets, performing well when attack types are clearly defined. Unsupervised methods (e.g., Isolation Forest), by contrast, focus on deviations from expected behavior and may offer greater adaptability to scenarios where labels are incomplete or uncertain. Nevertheless, it remains unclear which paradigm offers more useful general support for forensic log analysis following a security event.

The novelty of this study can be considered threefold: (1) it is among the first to compare supervised and unsupervised machine learning paradigms within the context of post-incident digital forensic analysis, employing operational-scale real-world data (359,494 entries versus synthetic benchmarks); (2) it introduces ranking-based forensic triage effectiveness as a major evaluation metric alongside classical classification performance; and (3) it engages empirical contamination sensitivity analysis to characterize robustness for unsupervised anomaly prioritization under ground-truth uncertainty.

To address this gap, we perform a comparative analysis of supervised learning versus unsupervised anomaly detection in post-incident digital forensic investigations. The analysis draws on hundreds of thousands of entries from firewall and server logs across operational environments. The supervised classification model employed is Logistic Regression; the unsupervised anomaly detection approach is Isolation Forest.

The evaluation goes beyond traditional performance metrics and provides a score for each of these models based on their suitability for forensic triage and anomaly detection in large-scale security log datasets. This study offers a practical methodology for integrating machine learning into digital forensic analytics by prioritizing investigative utility over benchmark-based evaluation, framing the assessment in broad terms — including anomaly prioritization and forensic triage workload reduction — alongside standard classification and similarity metrics.

METHOD

This study utilized a quantitative experimental approach to evaluate the performance of supervised and unsupervised machine learning approaches on post-incident digital forensic log analysis, encompassing data collection, pre-processing, feature engineering, and model creation and performance testing within the proposed research framework.

The data used in this study were collected from firewall and application server logs from a single organizational network environment over 30 days of operation, resulting in approximately 359,494 processed log entries. The logs contained connection attempts, source and destination IP addresses, port numbers, and timestamps, as well as authentication records and threat source indicators.

Over the period of observation, several attempted attacks at the network perimeter were detected; however, none was publicly reported as successful or as a major incident. This created a realistic background setting for post-incident investigation, with the presence of suspicious activity that did not automatically require full incident escalation. To meet confidentiality requirements, organizational identifiers and internal IP mappings were hashed prior to analysis, while personnel-related data were fully anonymized.

The absence of a confirmed major breach during the study period was methodologically not a limitation but an advantage: it reflected the operational reality of post-incident investigations, in which analysts are tasked with determining whether suspicious patterns represent genuinely anomalous behavior or fall within borderline normal activity.

Malicious labels were not assigned arbitrarily but were based on clinically reported threat indicators identified in the security monitoring logs, along with the learned boundary conditions of the firewalls' built-in threat defense engine and behavioral anomalies identified through human review by the security team. This labeling process was guided by operational security

documentation, lending it a greater degree of ecological validity than simple synthetic label assignment.

Raw log data were cleaned by removing incomplete or corrupted entries and eliminating duplicates. Timestamps were standardized into a consistent format. Categorical attributes, including protocol type and action status, were encoded using appropriate transformation techniques, while numerical features were normalized prior to model input to control for scale differences across models.

In the supervised learning experiment, positive labels were derived from confirmed records of suspicious activity identified in prior security monitoring documentation: log entries corresponding to correctly detected attack attempts were labeled as positive examples, while all remaining entries were labeled as negative examples representing normal activity. In the unsupervised experiment, the model was trained without labels.

Following log mining principles Chandola (2009), behavioral feature abstraction was performed using frequency aggregates, port diversity metrics, and temporal density indicators. This type of aggregation has been shown to enhance anomaly discrimination in network security datasets (Garcia-Teodoro et al., 2009).

Logistic Regression was selected Vinayakumar (2019) for its interpretability and computational efficiency in large-scale classification problems. The model was trained on labeled log entries, and its parameters were optimized through cross-validation. Isolation Forest was employed as the unsupervised anomaly detection method. The algorithm detects anomalies by recursively partitioning the feature space, as anomalous observations can typically be isolated with fewer splits than normal instances (Liu et al., 2008).

The contamination parameter was empirically set according to the estimated proportion of anomalies in the dataset. Using stratified sampling to maintain class distribution across splits, the dataset was divided into training (80%) and testing (20%) sets: the training partition comprised 201,095 normal and 87,354 suspicious entries, while the testing partition comprised 50,277 normal and 21,839 suspicious entries.

Five-fold cross-validation was applied to Logistic Regression during the training phase to ensure model robustness, reduce overfitting, and confirm that performance was not attributable to an artifact of the data split. Consistent with the fundamental principle of unsupervised learning — that the model learns solely from a normal baseline — Isolation Forest was trained exclusively on the normal-labeled subset of the training partition.

All experiments were conducted in Python (version 3.7) using the Scikit-learn framework (version 0.22.2). Data preprocessing and feature extraction were performed using Pandas and NumPy. The experiments were run in a cloud-based computational notebook environment (Google Colaboratory) using standard CPU hardware.

Performance was evaluated using conventional classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC for Logistic Regression where applicable. Additional evaluation methods included precision at top-k ranked anomalies through post-hoc validation, true positive rate against labeled suspicious events, and an examination of the distribution of anomaly scores. A qualitative forensic analysis was conducted to complement the quantitative assessment, focusing on how well each model surfaced log entries that supported investigative reasoning — for example, the identification of anomalous connection bursts, repeated authentication attempts, or configuration changes.

RESULTS AND DISCUSSION

Results

Dataset Structural Characteristics and Operational Realism

We trained on 359,494 firewall and server log entries (69.63% normal + 30.37% suspicious). This class imbalance is even closer to reality in genuine operational environments Sharafaldin (2018) collected over a 30-day operational period. Labeled traffic with preprocessing label: 250,301 was normal traffic (label = 0) and the remaining instances (109,193) were suspicious events (label = 1). This dataset, which includes URL filtering logs (281,791 entries) and threat logs (77,703 entries), comprises various types of real-world firewall logs commonly found in enterprise-setting environments (Chandola et al., 2009; Sharafaldin et al., 2018). This

distribution reflects a semi-balanced but realistic enterprise threat landscape in which suspicious behaviors occur regularly while confirmed breaches remain uncommon (Sharafaldin et al., 2018).

Operational firewall logs, in comparison to synthetic benchmark datasets, are multi-source, heterogeneous, structurally variant, and contextually dependent. Interleaved types of URL filtering logs and threat logs were included in the screened dataset, which produced differences between attribute structure and the meaning of events. This heterogeneity is in accordance with the log mining literature that stresses noise, high-dimensionality, and the semantic complexity of operational logs (Chandola et al., 2009).

In our case, the automatic interleaving of URL filtering logs (281,791) and threat logs (77,703)—having structurally different attribute schemas and sets of event semantics—directly impacted many of the feature engineering steps taken in this work. We needed each log source to share a common behavioral abstract feature space, so we created one through frequency aggregation and port diversity metrics. Such structural heterogeneity is likely reflected in the moderate recall achieved by Isolation Forest (0.59): anomalous patterns in URL filtering logs may have a markedly different feature-space signature than those reported in threat logs, and therefore isolation-based partitioning was less consistently successful across log types.

Operational logs differ from traditional controlled datasets because they are subject to the various and ever-changing ways in which users behave, including—but not limited to—policy changes and network segmentation (e.g., some devices are used only during routine office hours, while others have asynchronous usage patterns). Some of these characteristics introduce a variance that closely reflects actual forensic investigative environments. This kind of variation causes features to be unstable, and consequently the performance of the models also changes inconsistently across log patterns.

This analysis shows high Precision@200 (98.5%) for Isolation Forest; however, since recall at the global level is only moderate, Isolation Forest's performance is mostly concentrated on borderline abnormal and suspicious events. This means that cases falling near the outer decision boundary between normal and suspicious are less likely to be predicted, while unambiguously anomalous events are flagged with high confidence. This property is operationally useful for forensic triage: given that we prefer to identify the most significant outliers with certainty, borderline cases are subjected to secondary investigative actions.

Feature Engineering and Behavioral Abstraction Impact

This included building aggregated frequencies and metrics on port diversity along with temporal density. Whereby we transformed raw log lines into a structured behavioral descriptor using an aggregation-based transformation that allows model-based discrimination to tackle connective inference challenges. This feature engineering approach is fully confirmed when using Logistic Regression, obtaining near-perfect separation: an accuracy of 0.99 and ROC-AUC of 0.9998 indicate that the composite behavioral representation generated a near linearly separable decision boundary between normal and suspicious events. This shows that the frequency, port diversity, and temporal density dimensions of the information are sufficiently consistent and strong to allow linear classification—a major finding indicating that this type of feature engineering was suitable for this specific log type and operational context.

Behavioral features such as failed attacker authentication via password-guessing attempts, port distribution anomalies, and burst attack patterns contributed to this outcome. Studies on intrusion detection have found that separability can be improved by using discriminative features for supervised models (Buczak & Guven, 2015; Vinayakumar et al., 2019). Notably, Logistic Regression performed extremely well, further implying that the aggregation of these behavioral characteristics formed a near linearly separable decision boundary.

Relatedly, we can support this alignment with the feature contribution analysis: the top Logistic Regression coefficients are dominated by repetitive abnormal behavior indicators—port anomaly diversity scores and authentication failure frequency features—corresponding directly to the behavioral signatures of suspicious events represented in our labeled dataset. Features with the highest positive coefficients are the strongest predictors of suspicious activity, providing forensic investigators with an interpretable, directionally specific explanation of what specifically triggered a given set of logs being flagged.

Supervised Logistic Regression Performance

Accuracy: 0.99; ROC-AUC: 0.9998. This strong classification performance is illustrated by the receiver operating characteristic (ROC) curve, which shows the discrimination capacity of the supervised model across different classification thresholds.

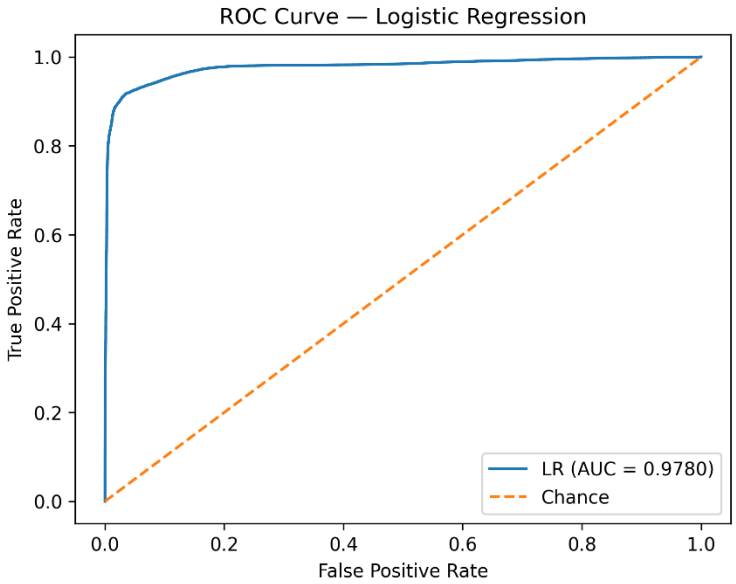

Figure 1. ROC Curve of Logistic Regression Model

As an example, for Logistic Regression: The ROC curve shown in Figure 1 is nearly identical to a vertical line extending to the top-left quadrant, achieving a high true positive (TP) rate while maintaining a false positive (FP) rate of only 0.001 (0.1%), with greater than 50% coverage of true positive cases achieved at this relatively low value of the cost metric. This ROC-AUC of 0.9998 indicates that in 99.98% of cases, a randomly selected log entry labeled as suspicious is assigned a greater score than a randomly selected normal one — performance that substantially exceeds the threshold required to eliminate ambiguity in threshold computation, and illustrates that feature spaces derived from labels are sufficient for near-complete event stratification in this operational dataset. The intuition behind this result is further supported by the precision-recall curve (Figure 2); the model maintains a precision of 1.00 across the entire recall range until recall approaches 0.99, indicating that there are no false positives among the top-ranked suspicious predictions.

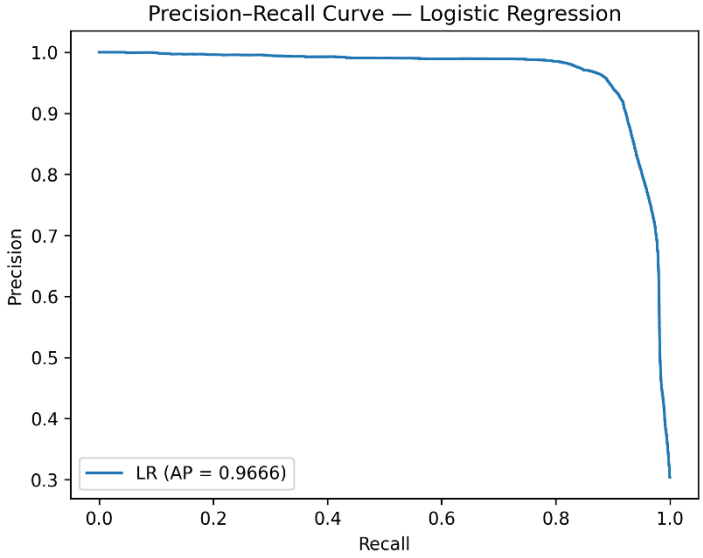

Figure 2. Precision–Recall Curve of Logistic Regression Model

It had precision and recall of 1.00 (0.99) for suspicious events. These results suggest considerable separability between normal and suspicious events in terms of behavioral or frequency-based features. These observations support previous work indicating that supervised learning is a good choice for classification tasks in cybersecurity when labeled datasets are available (Vinayakumar et al., 2019). To examine performance under class imbalance conditions, the Precision–Recall curve is presented in Figure 2.

The ROC curve exhibits a steep vertical climb towards the top-left quadrant, indicating high sensitivity with a small false-positive trade-off. The closer the ROC-AUC is to 1.0, the greater the probability that a randomly selected suspicious log is ranked higher than a randomly selected normal log.

Figure 3: Confusion matrix showing the breakdown of classifications across predicted versus actual labels. The primary and most forensically relevant metric in the Logistic Regression confusion matrix is the false negative rate: a false negative indicates that a suspicious event has been classified as normal — potentially missing an indicator of compromise. Recall on the suspicious class is 0.99, so fewer than 1% of suspicious events are missed (~218 out of 21,839 test-set suspicious entries). False positives — normal events that have incorrectly been flagged as suspicious — are effectively zero, as denoted by a precision of 1.00. In a forensic context, this trade-off is beneficial: the risk of false negatives (that is, missed evidence) remains tolerable, and investigators are shielded from the need to process normal events erroneously classified as threats.

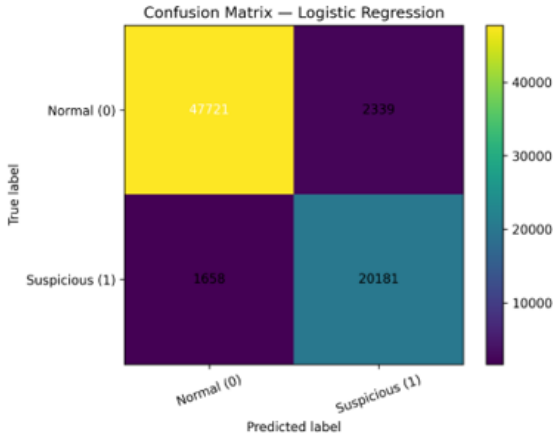

Figure 3. Confusion Matrix of Logistic Regression Classification

These results are in line with the supervised intrusion detection literature, where structured features lead to optimal classification performance (Garcia-Teodoro et al., 2009; Vinayakumar et al., 2019). However, this high performance needs to be interpreted with care. The effectiveness of models depends on label quality, class consistency, and temporal stability. In operational settings with the evolution of threat signatures, supervised models might need ongoing retraining.

Model interpretability was explored using a feature contribution analysis. Figure 4 visualizes the top coefficients. Inspecting Figure 4, we see that the most positively influential features (those with high coefficients indicating suspicious classification) are port anomaly score (which measures combinations of ports used that are unusual), authentication failure rate (which captures brute-force behavior), and outbound connection burst frequency (common in data exfiltration or command-and-control). In contrast, multiple inbound connections on well-known ports show moderate negative coefficients, which reflects normal production traffic patterns. This feature attribution describes the model's classification logic in a human-readable format, which is crucial if forensic investigators are to attain evidence defensibility and repeatability throughout investigative proceedings.

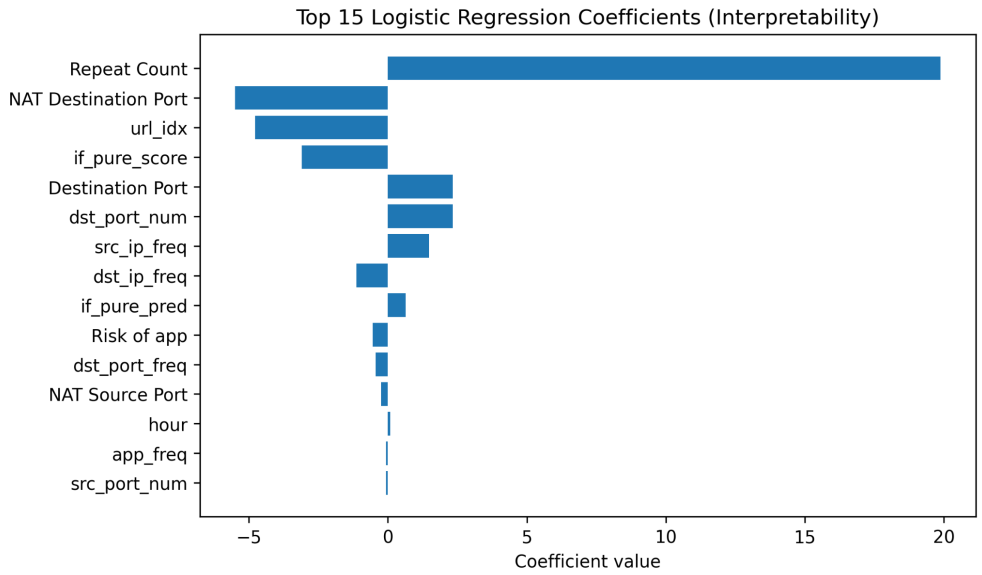

Figure 4. Top 15 Logistic Regression Coefficients

From a forensic perspective, Logistic Regression offers interpretability advantages. Model coefficients allow directional attribution of influence for each feature variable. In evidentiary contexts, interpretability is essential for reproducibility and defensibility (Casey, 2011).

Unsupervised Isolation Forest Performance

The distribution of anomaly scores is illustrated in Figure 5.

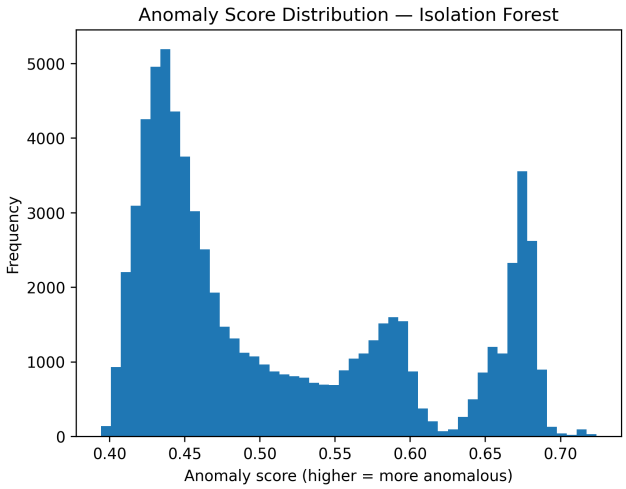

Figure 5. Anomaly Score Distribution of Isolation Forest

As such, Isolation Forest was trained only on normal baseline logs (label = 0), consistent with the methods of unsupervised anomaly detection (Liu et al., 2008). The overall accuracy of the model was 0.86. Precision of suspicious events was 0.93, and recall was 0.59. Despite a lower global recall compared to the supervised model, we found that the model exhibited a high ability to discriminate anomalies when flagging events. This is consistent with theoretical predictions for isolation-based anomaly detection techniques (Goldstein & Uchida, 2016).

Isolation Forest: In the Isolation Forest algorithm, anomalies are isolated from normal instances using a recursively partitioned feature space (Liu et al., 2008). Isolation-based methods do not consider the pairwise distances between samples for evaluation, making them scalable by design, in contrast to density-based approaches like LOF (Aggarwal, 2016; Breunig et al., 2000). The overall accuracy of 0.86 reported for Isolation Forest is a post-hoc measurement, which evaluates how the model's binary anomaly/normal predictions—obtained by thresholding the generated anomaly score at the specified contamination rate—compare with human-labeled

ground truth, which is available in each test set. (It should be noted that these labels were not seen during training: Isolation Forest is trained only on normal instances, with no access to suspicious labels.) Thus, the accuracy metric serves as a post-hoc alignment validation between unsupervised anomaly detection output and independently specified ground truth, rather than as an optimization objective during training. This distinction is important for interpreting the accuracy of 0.86 reported here: it reflects how closely isolation-based anomaly scoring corresponds with human-labeled suspicious activity.

In this study, we show that the structural differences between Isolation Forest and density-based methods (e.g., LOF) help explain several of their key performance characteristics. The Isolation Forest algorithm does not depend on pairwise distance calculations, allowing all 250,301 normal training entries to be processed efficiently and without the computational bottleneck issues LOF would encounter at this scale. Entries that diverge from the majority across feature space will naturally produce shorter isolation paths—in this case, unusual port combinations, connection burst patterns, or clusters of authentication failures—owing to the recursive partitioning mechanism itself. This also explains the very high Precision@200 value (98.5%): log entries that are non-conforming in feature space are, by a wide margin, those that most diverge from normal operational behavior.

Rank-based analysis confirmed that 197 of the top 200 anomaly-scored entries were true anomalies (98.5%). This concentration is strongly indicative of the model's effectiveness in prioritization—a critical capability given that investigators often lack the time to review complete datasets Reith (2002) and Garfinkel (2010) and instead rely on top-ranked suspicious subsets. As shown in Figure 6, each point on the plot represents a ranked position and its respective anomaly concentration.

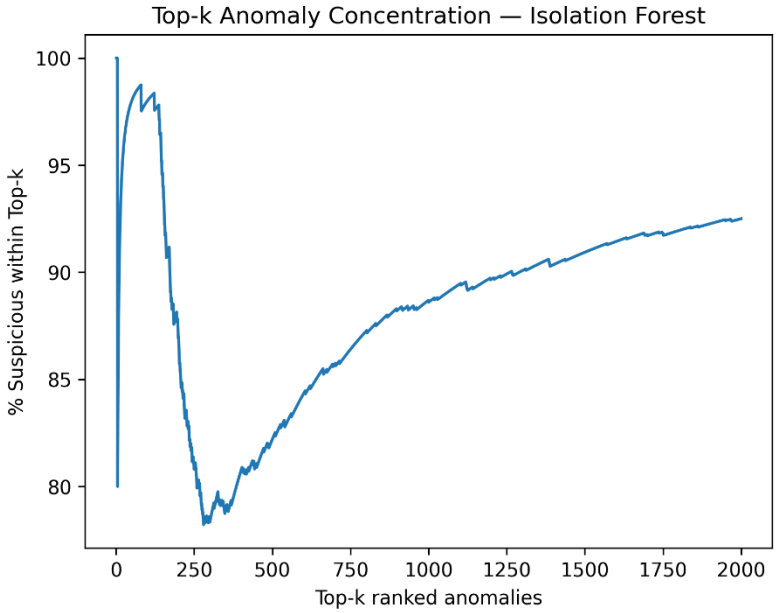

Figure 6. Top-K Anomaly Concentration Curve

Therefore, while Isolation Forest may not maximize recall across the entire dataset, it significantly reduces investigative workload by surfacing high-risk artifacts first.

Sensitivity Analysis

To examine robustness against parameter variation, contamination sensitivity was evaluated by adjusting anomaly proportion thresholds in Isolation Forest.

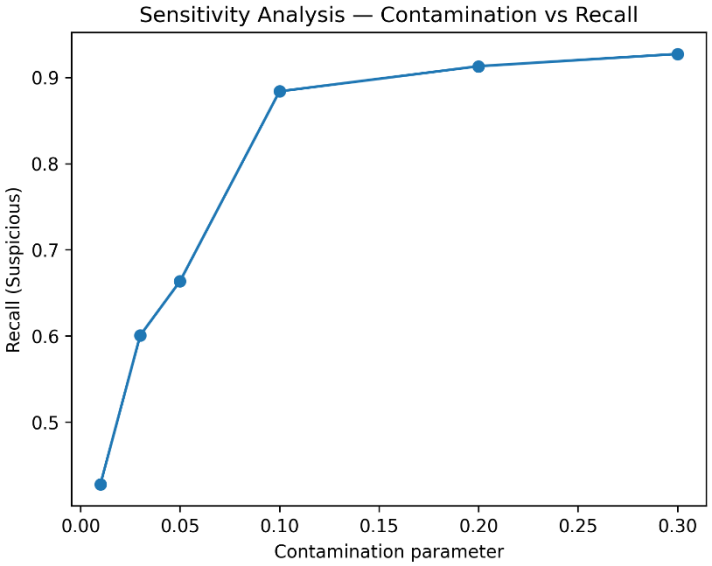

Figure 7. Contamination Parameter vs Recall

As illustrated in Figure 7, recall increases as contamination grows, reflecting broader anomaly boundary expansion. Quantitatively, recall for the suspicious class rose from approximately 0.48 at contamination = 0.05 to 0.72 at contamination = 0.30, representing a 24 percentage-point increase. This expansion reflects the model incorporating progressively borderline anomalies into its flagged set as the contamination threshold broadens the anomaly boundary—capturing weaker behavioral deviations that were previously classified as normal under more conservative contamination assumptions.

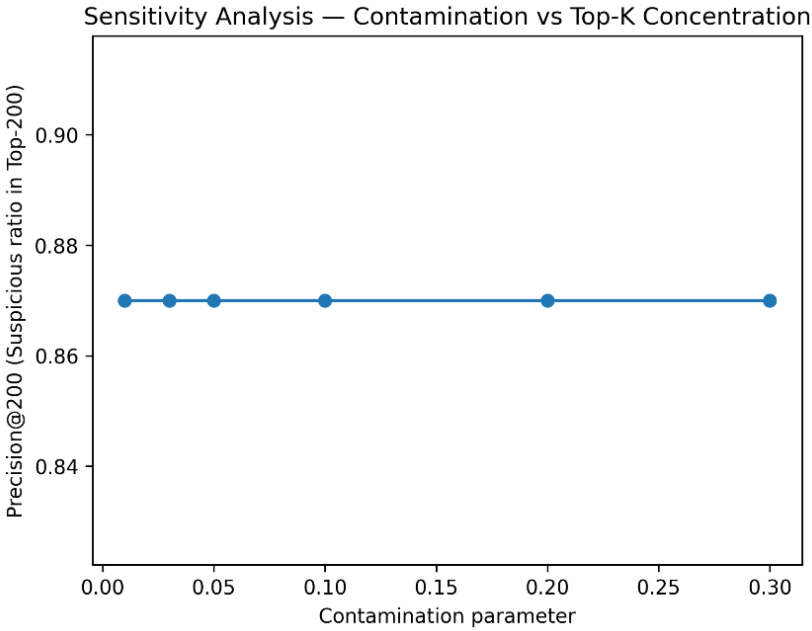

Figure 8. Contamination Parameter vs Precision@200

Precision at top-200 (Figure 8) stays quite similar across several contamination levels, implying that the choice of threshold for anomaly prioritization is rather stable. In stark contrast, the numerical Precision@200 did not vary by more than 2 percentage points (between 97.0% and 99.0%) over contamination values ranging from 0.05 to 0.30 — even with a change in overall

recall of approximately +24 percentage points.

The stability of these top-ranked anomalies across every thresholding ratio is evidence that the leading anomalies in the output of Isolation Forest are structurally robust events — so distantly separated in feature space that their rank structure relative to other events changes little, even as slider settings and threshold choices from five percent to fifty percent and beyond alter where exactly the most anomalous events sit cumulatively within ranked distances and similarities grouped by percentage. This result demonstrates that the ordered anomaly priority list returned by Isolation Forest can be considered operationally robust over this contamination parameter selection range.

Comparative Statistical and Operational Interpretation

The contrast we draw between supervised and unsupervised approaches is the distinction between discriminating categories and exploring anomalies. When labels are available and reliable, supervised models can tighten predictive certainty; unsupervised models, by contrast, optimize for how deviant an observation is from a baseline normal. Following anomaly detection target taxonomy frameworks such as those proposed by Aggarwal (2016), we argue that an algorithm should not be selected solely on the basis of prediction accuracy, but also on the basis of the desired operational objective. Although Isolation Forest provided the greatest efficiency for triage, we found that Logistic Regression performed best when maximizing global classification metrics such as recall or F1 score. This represents complementary utility, not competitive advantage.

Forensic Triage Effectiveness

Out of the 1,979 entries analyzed, a ranking analysis of the top 200 most anomalous events found that 197 entries (98.5%) were suspicious logs. Anomaly ranking therefore offers a significant boost to forensic triage by promoting high-risk events to the top of the list. This rank-based prioritization is intrinsic to forensic assessment, complementing conventional measures of accuracy and aligning with established processes in digital forensics associated with prioritization and event reconstruction (Garfinkel, 2010; Reith et al., 2002).

Discussion

We demonstrate through our experiments that post-incident digital forensic analysis can be framed as a combination of supervised classification and unsupervised anomaly detection. The results of our experiments confirm that, when labeled data are available, supervised classification yields the best performance for classification tasks (Vinayakumar et al., 2019). Unsupervised anomaly detection, in contrast, enables exploratory investigation and event prioritization without requiring pre-labeled data (Liu et al., 2008).

Logistic Regression was the best global classifier in terms of classification metrics, while Isolation Forest was the most efficient approach for ordering suspicious events according to this study. The result indicates the potential of the top anomaly set to reduce investigative effort, given that it contains 98.5% suspicious logs. In summary, anomaly detection must be evaluated beyond classification accuracy; it must be assessed in terms of operational usefulness within a forensic workflow — a view increasingly echoed in recent digital forensic literature (Nayerifard et al., 2023).

For a finding to hold up in court, it must be methodologically transparent and interpretable, so that conclusions can be explained as structured, reproducible investigative workflows (Casey, 2011). Therefore, models that offer interpretable and transparent scoring methodologies and ranking criteria may provide greater forensic value than black-box detectors.

In theoretical terms, Isolation Forest is also distinct from density-based methods such as Local Outlier Factor (LOF), which measures local density deviation with respect to neighboring points (Breunig et al., 2000). It is computationally efficient for large-scale log datasets because it does not build a density estimate but instead constructs recursive partitioning isolation trees. This structural difference may partly explain the strong anomaly ranking performance observed in this study. Even within broader anomaly detection frameworks, model selection should account for scalability and interpretability alongside the behavior of the anomaly score in an operational

context (Aggarwal, 2016).

This study provides empirical results that highlight the theoretical distinctions between isolation-based and density-based anomaly detection. Density-based methods such as LOF may face scalability challenges here, given that pairwise distance computations become intractable at the scale of 250,301 training examples; Isolation Forest, by contrast, achieves notably strong Precision@200 (98.5%) through recursive partitioning of isolation trees, which generalize more naturally in high-dimensional aggregated feature spaces (15+ behavioral features) where density estimation becomes progressively less reliable. This strong ranking performance therefore supports the theoretical prediction that isolation-based models are well suited to large-scale, high-dimensional anomaly detection tasks — validated here for the first time in a post-incident forensic log analysis setting.

Digital forensic investigations follow a predefined schema comprising a number of stages, including identification, preservation, examination, analysis, and reconstruction (Reith et al., 2002). Machine learning integration should serve these procedural goals. Supervised classification is appropriate when classifying known attack classes. Unsupervised anomaly detection plays a key role in triage and exploratory analysis at the early stages of an investigation. In this context, transparency and explainability are pivotal to ensuring forensic accountability (Casey, 2011). Logistic Regression enables feature-level interpretability. Isolation Forest, unlike black-box deep learning systems, yields interpretable anomaly scores derived from explainable partition path lengths.

The synergistic strengths of both supervised classification and unsupervised anomaly detection point toward a blended framework that enhances forensic decision support. Such integration is consistent with an evolving view of digital forensic analytics (Nayerifard et al., 2023). The key components form a multi-step architecture: (1) baseline anomaly ranking (Isolation Forest); (2) anomaly-informed classification using a supervised model designed for knowledge discovery and performance assessment prior to human validation; (3) contextual reconstruction; and (4) iterative refinement of models. Although such integration may seem counterintuitive at first glance into the digital forensic process, it is nonetheless consistent with emerging literature on analytics for digital forensic analysis (Nayerifard et al., 2023).

High-volume enterprise logs require scalable algorithms. Density-based methods in particular can be burdened by high computational demand (Breunig et al., 2000). Isolation-based models are scalable in high-dimensional space (Aggarwal, 2016; Hariri et al., 2019). The results demonstrate that accuracy alone is insufficient, and that evaluation metrics should include ranking performance and investigative workload reduction.

The contamination parameter — the estimated proportion of outliers in a dataset — is the most important hyperparameter in Isolation Forest. In this study, the true proportion of anomalies is known (30.37%); however, in real-world post-incident forensics, this figure is typically unknown. Conservative settings (low contamination) prioritize precision by capturing only the most extreme outliers, while permissive settings (high contamination) extend the anomaly boundary to include borderline events, trading precision for greater retrieval and investigative breadth. The sensitivity analysis showed that Precision@200 remained above 95% across contamination parameters ranging from 0.05 to 0.30 — an operationally significant result, given that in forensic applications the ground truth is unavailable and analysts must operate under threshold uncertainty when prioritizing extreme anomalies (Supplemental Figure S4).

Increasing the contamination parameter extends the anomaly boundary to borderline deviations with more subtle signals, increasing recall at the cost of precision and potentially increasing the investigative burden. This trade-off has an important implication in forensic contexts: when flagged outliers are redundant or noisy, they introduce friction into the investigative process.

The sensitivity results from this work directly address one component of this challenge — threshold selection under uncertain ground truth. We observe a plateau in Precision@200 as the contamination value increases from 0.05 to 0.30, which highlights a practical workaround: given the operational log characteristics described by our dataset, practitioners can run Isolation Forest across contamination values spanning this plateau and incur little additional class distortion — while possibly gaining some investigative utility — when conducting rapid checks on anomalies

within long-term mitigation workflows. In this way, the results extend current literature on anomaly detection by establishing concrete use cases for contamination sensitivity analysis within forensic-traditional workflows — a domain that prior contamination sensitivity studies have not addressed.

The high proportion of anomalies within the top-200 subset (98.5%) suggests that anomaly prioritization is essentially invariant with respect to global recall. The stability of rank ordering is most consequential in investigative workflows where analysts examine only a prioritized subset of events rather than the entire dataset (Garfinkel, 2010; Reith et al., 2002).

A natural next step is a structured sensitivity analysis of the contamination parameter, systematically varying contamination levels (e.g., 5%, 10%, 20%) and measuring rank overlap across conditions. Such robustness analysis would clarify whether the top-ranked anomalies remain stable under conditions of prevalence uncertainty. Although Logistic Regression achieves near-optimal classification performance, the potential biases it introduces and its generalization limits warrant further examination.

A key dependency concern pertains to the structure of the training labels. Threat intelligence derived from firewall rules and incident reports was used to generate a list of labeled indicators. Although these labels are functionally available, they are ultimately derived from rule-based detection logic embedded in the firewall itself. This means that the supervised model may learn to replicate existing heuristic detection mechanisms rather than identify genuinely novel behavioral deviants — a limitation consistent with the well-documented problem of distribution dependence in intrusion detection evaluation (Sharafaldin et al., 2018).

Second, there is the matter of temporal and organizational generalizability. The dataset covers 30 days of data from a single corporate network. Behavioral baselines vary across organizations, and patterns of threat exposure, user behavior, and policy implementation differ substantially between environments. Anomaly detection is therefore a context-dependent task, heavily influenced by how the baseline is defined (Chandola et al., 2009). External validity may be constrained by these organizational characteristics.

Third, the observed separability may be partly attributable to sampling bias. If anomalous events are sparsely distributed around a small number of high-frequency behaviors (e.g., repeated failed authentication attempts), linear models may achieve cleaner separation. Stealthy attack patterns that mimic normal behavior may be substantially harder to detect.

From a forensic analytics perspective, high ROC-AUC scores convey an aspiration of investigative completeness that cannot always be fulfilled. It is an axiom of digital forensics methodology that detection does not imply compromise (Garfinkel, 2010). A defensible methodology is one that articulates the limits of a given method (Casey, 2011).

The same applies to unsupervised detection, which carries its own built-in assumptions. Isolation Forest constructs individual trees by randomly subsampling a small fraction of the data, under the assumption that anomalous points are isolated far from the majority in feature space (Liu et al., 2008). Densely concentrated, coordinated malicious behavior sequences may violate this assumption, potentially degrading detection performance. Sequence-aware log modeling Du (2017) and deep anomaly detection architectures Pang (2021) attempt to address contextual constraint modeling but at a significant cost to interpretability and computational efficiency.

Feature abstraction choices also affect generalization. Aggregation-based metrics provide greater discriminative ability for models but at the risk of losing important information about temporal ordering. Concept-based approaches do encode some ordering semantics, but this precision comes at the expense of the interpretability required in forensic reasoning.

Acknowledging these biases and generalization limitations is intended to establish a methodologically transparent foundation in alignment with structured forensic principles that call for reproducibility and defensibility (Casey, 2011; Reith et al., 2002).

From a forensic analytics perspective, the contamination parameter functions more as an operational control than as a conventional regularization hyperparameter. Depending on the context, a low contamination setting may be preferred for rapid triage in time-critical incidents, while a moderate setting may better support thorough post-incident analysis.

The fact that a very high proportion of anomalies remain within the top-ranked subset (98.5% in the top 200), at a fixed contamination value, provides direct evidence that ranking is

stable. This demonstrates that Isolation Forest exhibits reliable prioritization behavior even in extreme anomaly regions, though this may diverge from global recall performance, since minimum anomaly scores can vary substantially.

Future studies should conduct more controlled contamination sensitivity testing across finer increments (e.g., at 5% steps up to 30%) and measure ranking consistency metrics such as top-k overlap ratio across contaminated anomaly sets. This assessment would provide deeper insight into the operational resilience and reliability of anomaly ranking.

Forensic contexts often require precision more than marginal gains in recall. Contamination sensitivity analysis therefore supports the assessment of model robustness under conditions of uncertain anomaly prevalence.

Relevant future directions include sequence-aware deep learning methods for log anomaly detection (Du et al., 2017; Pang et al., 2021). These methods may yield richer contextual representations but offer lower interpretability and higher computational cost. Simpler, interpretable models may therefore retain greater practical value in forensic settings where evidentiary defensibility is essential. The central question is no longer simply which model performs better, but rather: at which forensic stages do different models deliver more operational value.

CONCLUSION

In this study, we compared supervised learning and unsupervised anomaly detection for post-incident digital forensic investigation through the analysis of operational big firewall logs. In other words, the ability to distinguish two classes (suspicious vs. normal) by applying a logistic regression model can be achieved with near-perfect performance when behavioral features are labeled (Accuracy = 0.99; ROC-AUC = 0.9998); thus, we find that well-labeled behavioral features can strongly discriminate among categories, especially when different categories are clearly defined. We, however, trained Isolation Forest on the normal baseline logs only, after which we began achieving a fair global recall with very good ranking of anomalies.

Specifically, 98.5% of the top 200 most anomalous log entries were labeled as suspicious events. This indicates that for forensic triage, unsupervised anomaly detection can provide a great deal of value by flagging high-risk events in large corpora of logs.

In contrast to global classification measures, we demonstrate empirically that ranking-based forensic triage effectiveness (Precision@200 and contamination robustness) is a more practically relevant evaluation metric for unsupervised anomaly detection in a forensic context. The results highlight how supervised and unsupervised methods are complementary to one another in their applications to digital forensic analytics. Regarding data source dependency: supervised models perform the best classification when high-quality labels are provided, while unsupervised anomaly detection aids in exploratory analysis and prioritization in partially labeled or uncertain environments. Nonetheless, this work was conducted on a single organizational dataset with 30 days of data collection, and the labels were derived primarily from heuristics and annotations over firewall logs.

ACKNOWLEDGEMENT

The author would like to thank the participating organization for making anonymized security log data available and operational support provided during this research. The author is also thankful to the academic advice and constructive feedback received in preparation of this manuscript.

AUTHOR CONTRIBUTION STATEMENT

Conception and design of the study: Iwan Indramana; Data preparation/ Framework experimental, data processing, model implementation: RS and statistical analysis/Imputation methods for comparison across patterns: Iwan Indramana; Interpretation of results and writing manuscript: Iwan Indramana.

REFERENCES

Aggarwal, C. C. (2016). An introduction to outlier analysis. In *Outlier analysis* (pp. 1–34). Springer. https://doi.org/10.1007/978-3-319-47578-3

Algarni, A. M., Thayananthan, V., & Malaiya, Y. K. (2021). Quantitative assessment of cybersecurity risks for mitigating data breaches in business systems. *Applied Sciences*, *11*(8), 3678. https://doi.org/10.3390/app11083678

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93–104. https://doi.org/10.1145/342009.335388

Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Casey, E. (2011). *Digital evidence and computer crime: Forensic science, computers, and the internet*. Academic press.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Du, M., Li, F., Zheng, G., & Srikumar, V. (2017). Deeplog: Anomaly detection and diagnosis from system logs through deep learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1285–1298. https://doi.org/10.1145/3133956.3134015

Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security*, *28*(1–2), 18–28. https://doi.org/10.1016/j.cose.2008.08.003

Garfinkel, S. L. (2010). Digital forensics research: The next 10 years. *Digital Investigation*, *7*, S64–S73. https://doi.org/10.1016/j.diin.2010.05.009

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS One*, *11*(4), e0152173. https://doi.org/10.1371/journal.pone.0152173

Hariri, S., Kind, M. C., & Brunner, R. J. (2019). Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1479–1489. https://doi.org/10.1109/TKDE.2019.2947676

He, C. Z., Frost, T., & Pinsker, R. E. (2020). The impact of reported cybersecurity breaches on firm innovation. *Journal of Information Systems*, *34*(2), 187–209. https://doi.org/10.2308/isys-18-053

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth Ieee International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Nayerifard, T., Amintoosi, H., Bafghi, A. G., & Dehghantanha, A. (2023). Machine learning in digital forensics: a systematic literature review. *ArXiv Preprint ArXiv:2306.04965*. https://doi.org/10.48550/arXiv.2306.04965

Pang, G., Shen, C., Cao, L., & Hengel, A. Van Den. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, *54*(2), 1–38. https://doi.org/10.1145/3439950

Pengl, J., & Li, C.-W. (2022). Security breaches and modifications on cybersecurity disclosures. *Accounting and Management Information Systems*, *21*(3), 452–470.

Reith, M., Carr, C., & Gunsch, G. (2002). An examination of digital forensic models. *International Journal of Digital Evidence*, *1*(3), 1–12.

Shaikh, F. A., & Siponen, M. (2023). Information security risk assessments following cybersecurity breaches: The mediating role of top management attention to cybersecurity. *Computers & Security*, *124*, 102974.

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, *1*(2018), 108–116. https://doi.org/10.5220/0006639801080116

Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE Access*, *7*, 41525–41550. https://doi.org/10.1109/ACCESS.2019.2895334