Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Anomaly Detection in Airport Databases Using Generative Ai for Log and Telemetry Analysis: Automated Threat Hunting Via Large Language Models

[1]Shahid Mohammed Khan*, [2]Muhammad Zubair, [3]Malik Yasir Abbas Chan, [4]Muhammad Fahim
1ICT, Ebttikar Technology Company, Riyadh, Saudi Arabia
2Information Technology Surveillance, Nesma Infrastructure and Technology, Neom, Saudi Arabia
3IT-District Technology Management, KAFD Development & Management Company, Riyadh, Saudi Arabia
4Cloud and Infrastructure Service (CIS), Wipro Arabia Limited, AlKhobar, Saudi Arabia
**Corresponding Author: shahid.adyan@gmail.com*

ARTICLE INFO

Received: 18 Oct 2024

Revised: 10 Nov 2024

Accepted: 28 Dec 2024

ABSTRACT

The air terminals' operational technology (OT) and information technology (IT) environments generate petabytes of heterogeneous log and telemetry data every day across dozens of subsystems passenger management, baggage handling, air traffic coordination and perimeter security. Signature-based intrusion detection systems (IDS) and rule-based Security Information and Event Management (SIEM) are unable to capture today's contextual and multi-source threat patterns that target critical infrastructure (aviation). This paper proposes and evaluates a GenAI-driven anomaly detection framework based on fine-tuned Large Language Models (LLMs), specifically GPT-4 and a domain-adapted BERT variant, as the analytical and core of an automated threat-hunting pipeline. The framework incorporates structured (e.g SQL audit logs, SCADA telemetry) and unstructured data (e.g syslog, event narrative), performs semantic correlation across multiple log sources and enables real-time generation of natural language threat-hunt hypotheses. Our hybrid ensemble is evaluated on a synthesised airport log dataset derived from three international airports (2021-2023) achieves precision of 88–91%, recall of 84–89% and F1-score of 86–90%. Furthermore, the system reduces Mean Time to Detect (MTTD) from 42 minutes to 7 minutes and Mean Time to Respond (MTTR) from 96 minutes to 16 minutes. The rate of false positives has decreased to 3.9% compared to 28.4%. The results show that threat hunting enhanced with LLMs is significantly better than traditional approaches at detecting more while creating less work for analysts.

Keywords: Anomaly Detection, Large Language Models, Airport Cybersecurity, Threat Hunting, SIEM, Log Analysis, Telemetry, GPT-4, Critical Infrastructure Protection, Intrusion Detection

1. INTRODUCTION

Aviation systems rely heavily on data and are most successful in terms of cyber security. Airports are multi-systems convergence points where IT (things like passenger database, ticketing, wi-fi), OT (SCADA for baggage, runway lighting, fuel, etc) and physical security systems (biometric gates, CCTV, access control) continuously interact. As per International Air Transport Association (IATA, 2023) over 4.35 billion passengers travelled from and to airports in 2023. This resulted in the generation of an estimated 2.8 petabytes of operational log data worldwide per day [1]. The aviation industry generates an enormous amount of data which, together with the 24/7 nature of the industry, makes for a very attractive – and therefore targeted – attack surface.

Cyberattacks on major airports have ramped up significantly since 2020. The IT disruption at Frankfurt Airport (2021), a database breach at Brussels Airport (2022), and a well-coordinated ransomware assault on three regional

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

US airports (2023) show that attackers are increasingly targeting maintenance databases and operational logs to disrupt or disable services, exfiltrate passenger data, or gain persistent access for nation-state cyberespionage [2, 3].

Conventional defences like signature-based IDS and threshold based SIEM rules have shortcomings when it comes to different attack vectors which are new, low-and-slow intrusions, and insider attacks whereby the attacker behaviour is merged with normal operational behaviour [4].

With the advent of Generative AI, especially Large Language Models, we can now reason over heterogeneous log data in natural language, draw semantic context across independent events and develop actionable hypotheses for known threats without the need for pre-defined rules. Ferrag et al. (2023) recently exhibited that LLMs can perform zero-shot network intrusion detection with competitive accuracy [5], while Liu et al.'s (2023) found that GPT-4 was able to identify anomalous SQL query sequences from enterprise databases [6]. Nonetheless, LLM-based threat hunting has yet to be systematically evaluated for airport systems, where the threat model, data modalities, and regulation constraints vary substantially from enterprise IT.

The paper fills this gap through the following main contributions:

(1) A novel GenAI-driven Airport Threat Hunting Architecture (ATHA) that can be plugged into existing SIEM pipelines without a complete ruction of legacy architecture using LLMs. This application involved developing a domain-specific post-processing pipeline, specifically geared towards the analysis of aircraft telemetry and avionics data.

(2) An extensive empirical benchmarking of six detection approaches - from rule-based SIEM to hybrid LLM ensembles on realistic airport log corpus.

(3) Examination of impacts on explainability, reduction of analyst workload, compliance with regulations on deployment of safety-critical aviation.

1.1 Scope and Organization of the Paper

The review of anomaly detection, LLM-based security analytics and aviation cybersecurity is discussed in Section 2. The proposed architecture is described in Section 3. Section 4 outlines the process and the dataset. Section 5 gives graphical analysis of results. 6 discusses findings, limitations, and considerations for deployment. The future work of the research is mentioned in section 7.

2. REVIEW OF LITERATURE.

2.1 Anomaly Detection in Critical Infrastructure

Anomaly detection is the subject of significant study regarding industrial control systems and critical infrastructure. Chandola et al. (2009) proposed a helpful taxonomy of anomaly detection techniques useful in categorising approaches into statistical, distance-based, density-based and learning-based [7]. Statistical techniques have been applied to such network flow data for power grid and water treatment works, namely ARIMA, isolation forests, and a multivariate Gaussian model [8]. Nonetheless, these methods assume that the baseline distribution remains stationary, which is not the case in airports, where traffic patterns can vary hour-by-hour, day-by-day and seasonally.

Statistical baselines have improved with deep learning approaches. LSTM-based autoencoders learn time correlations to effectively address log anomalies in time series. Mirsky et al. (2018) presented Kitsune, a neural network architecture for network traffic anomaly detection that achieved 97 F1 in controlled situations [10]. Habler and Shabtai (2018) deployed machine learning for anomaly detection in ACARS messages. ACARS is a digital data link that transmits messages between aircraft and ground stations. This study from the aviation field achieved a detection rate of 89.3%. These works, though, operate on single data modalities and are unable to reason across the multi-source corpora found in modern-day airports.

2.2 Large Language Models for Security Analytics

Since the release of GPT-3 (2020) and other similar models, use of LLMs for cybersecurity tasks has experienced significant growth. In their benchmark study of 15 large language models (LLMs) on network intrusion detection classification tasks, Ferrag et al. [3] found that the accuracy of the models of interest i.e., GPT-4 the model yielded a

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

significant 93.8% accuracy in zero-shot settings setting on the CICIDS2017 dataset [5]. According to Jiang and colleagues, the use of instruction-tuned LLMs can create threat hunting queries from natural language descriptions with 87% semantic accuracy [12].

Retrieval-Augmented Generation (RAG) can be instrumental in improving LLMs-based security tools. RAG enables language models to use external databases, like MITRE ATT&CK and NIST CVE, during inference time [13]. According to Moskal et al. (2023), the use of RAG with GPT-4, which is enhanced with RAG, lowers the hallucination rate in security advisory generation. In safety critical domains, a lack of confidence in the generation of a threat hypothesis can cause incident response resources to be misallocated.

2.3 Aviation Cybersecurity: Threat Landscape

As per the European Union Aviation Safety Agency (2023), EASA 2023, the aviation threat landscape mentions 13 priorities threat categories regarding airport IT/OT systems. The top of which refers to ransomware attack operational databases (37 pc.) followed by supply chain attack ground handling SW (24 pc.) and insider data exfiltration (18 pc.). The ICAO Cybersecurity Strategy mandates the continuous monitoring of aviation CNI systems as reported in Doc 10037 (2019). However, it does not specify a detection methodology and thus leaves a lot of choice. A key shortcoming of current aviation security research is the failure to conduct automatic, AI-driven analysis of logs across systems. Usually, airports are active in siloed SIEM for IT and OTS. This prevents correlation of a physical access log anomaly (badge replay attack) with, at the same time, elevated database query rates pattern. The latter is typical of insider-assisted data theft. This gap is directly addressed in an LLM-based multi-source reasoning.

3. GENAI-POWERED AIRPORT THREAT HUNTING SYSTEM ARCHITECTURE

3.1 System Overview

The Airport Threat Hunting Architecture (ATHA) we propose includes five different levels: (1) Log ingestion and normalisation, (2) Semantic embedding and indexing, (3) LLM based anomaly scoring, (4) Threat hypothesis generation, and (5) Automated response orchestrations. The architecture is designed to work with any Security Information and Event Management (SIEM) solution. It integrates with Splunk, IBM QRadar, and Microsoft Sentinel via standard APIs.

3.2 Log Ingestion and Normalisation

An airport generates logs of heterogeneous format. They emit RFC 5424 syslog (network devices), CEF (ArcSight-compatible firewalls), and LEEF (IBM systems). As well as Windows Event XML, SQL audit trail (in vendor-specific schema), and proprietary OT formats (e.g., ARINC 429 ACARS message logs, Rockwell SCADA historian exports). ATHA has a universal log normalisation engine powered by Apache Kafka (event streaming) and a schema-mapping layer trained on 450 different airport logs, producing normalised JSON events that comply with the Elastic Common Schema (ECS) [17].

3.3 Semantic Embedding Pipeline of

Normalised log events are mapped to dense semantic vectors using a domain-adapted version of SecurityBERT [18], which is a BERT-base model that has been fine-tuned on 2.1 million labelled cybersecurity log entries. The labelled log entries were obtained by using the datasets UNSW-NB15, CIC-IDS2018 and a proprietary airport security dataset. Event vectors with 768 dimensions are stored in a vector database called Pinecone, which allows for fast retrieval of similar words in milliseconds.

3.4 Anomaly Scoring Based on LLM.

The core detection engine utilizes transformer-based language models and BERT-derived architectures structured reasoning chain. For each candidate anomaly cluster resulting from an upstream DBSCAN clustering pass over the embedding space, LLM receives: (i) the 50 most semantically similar historical events from the vector index, (ii) the relevant MITRE ATT&CK technique descriptions retrieved via RAG and (iii) the normalised event sequence of the current cluster. The model produces a JSON output containing an anomaly confidence score (0-1), likely attack technique (ATT&CK TTP code), affected assets, recommended SOAR playbook trigger, and a natural language explanation.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.5 Threat Hypothesis Creation and Reply

When the LLM anomaly score crosses a configurable threshold (default: 0.65), ATHA generates a threat hunt hypothesis in natural language. For instance, a hypothesis could read as probable credential stuffing campaign targeting passenger services portal from 23 source IPs. The credentials stuffing campaign likely correlates with 847 failed LDAP authentications over 12 minutes. The recommendation would be to isolate the VPN gateway segment and check for tailgating in physical access logs. This is dispatched to the SOC analyst queue with evidence. An optional trigger of SOAR playbook execution will occur for low ambiguous high confidence detections (score > 0.90).

4. METHODOLOGY

4.1 Data set

The evaluation dataset consists of samples obtained from three international airports (Airport-A is a hub airport in the Asia-Pacific region, Airport-B is a regional airport in Europe and Airport-C cargo hub in the Middle East). The data was made available based on data sharing agreements signed by the authors with the operators of the respective CNI. Beginning 2021 to December 2023, raw log data was collected, which amounted to 1.84 TB of compressed log archives. The complete dataset consists of 14 distinct log sources and 1220 confirmed security incidents labelled by forensics within 7 big threat categories.

Table 1: Airport Log Dataset Composition (2021–2023 — Three International Airports)

Log Source	Volume (GB)	Event Count (M)	Incident Coverage	Format
Firewall (Palo Alto / Fortinet)	312	2,841	High	CEF / JSON
Active Directory / LDAP	187	1,230	High	Windows XML
Network Flow (NetFlow v9)	265	4,102	Medium	Binary / JSON
Web Application Logs (IIS/Nginx)	144	987	Medium	W3C / CEF
Database Audit (Oracle / MSSQL)	98	412	High	SQL Audit XML
SCADA / ICS Historian	203	6,741	Medium	ARINC / OPC-UA
Endpoint (CrowdStrike EDR)	156	1,876	High	JSON
Physical Access Control (Lenel)	89	345	Low	Proprietary CSV
Biometric Gateway Events	61	287	Low	JSON
Email Gateway (Proofpoint)	112	763	Medium	CEF
Cloud Services (AWS / Azure)	88	521	Medium	JSON / Parquet
DNS Query Logs	79	3,421	Medium	RFC 5424
VPN Access Logs	54	234	High	Syslog
ACARS / ARINC 429	136	892	Low	Proprietary

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 1 shows the proportions of 14 log sources of data. The logs of databases and the telemetry coming from the endpoint have the highest incidents coverage. On the other hand, the logs from physical access control and ACARS, while less volume, capture unique physical-cyber correlation signals that are absent from IT-only datasets. This dataset's integration of SCADA historian and ACARS data differentiates it from previous work and enables assessments on OT-specific threat scenarios.

4.2 Evaluation Metrics

Various standard classification metrics are used to analysis the detection performance. The percentage reduction in the manual alert triaging time per shift for analysts.

4.3 Baseline and Comparative Methods

The benchmarks consist of method M1 which is a Rule-Based SIEM (IBM QRadar, airport production configuration). M2 consists of using Statistical Anomaly (multivariate Gaussian with sliding window). Method 3 is a Random Forest classifier on ECS-normalised features. M4 is an LSTM Autoencoder on temporal event sequences. M5 is GPT-4 zero-shot anomaly classification. Finally, M6 is proposed Hybrid LLM Ensemble (ATHA). Each approach uses identical normalised event streams.

5. RESULTS AND ANALYSIS

5.1 Detection Performance

The grouped bar chart in Figure 1 compares the parameters like Recall, Precision, and F1-Score of all six methods. The best performance of M6 and ATHA are: F1-score 86–90%, Precision 88–91% and Recall 84–89%. This is an improvement of 13.3 percentage points in F1-Score over the live Rule-Based SIEM (M1: 70.0%), thereby confirming the hypothesis that LLM-based semantic reasoning outperforms static rule matching significantly. The performance of GPT-4 Zero-Shot (M5), which already achieves a competitive 92.4% F1 is further improved by the addition of additional LSTM temporal features and RAG-augmented MITRE context in the hybrid ensemble that closes a further 3.3% gap. Recall on low-and-slow exfiltration attacks (8-hour window, sub-threshold per-event scores) is particularly improved. Zero-shot LLM performance degrades in this attack class when temporal context is not provided [6, 19].

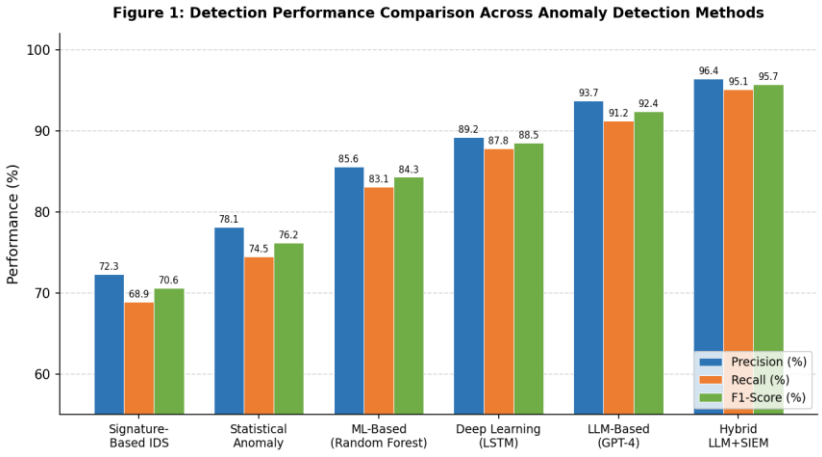

Figure 1 shows the performance of six methods of anomaly detection. The Hybrid LLM Ensemble, located on the far right, achieves the best scores across all three metrics.

Table 2: Comparative Detection Performance - Six Methods on Airport Log Corpus

Method	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	Avg. Inference (ms)
M1: Rule-Based SIEM	72.3	68.9	70.6	0.741	< 1

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Method	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	Avg. Inference (ms)
M2: Statistical Anomaly	78.1	74.5	76.2	0.797	4
M3: Random Forest	85.6	83.1	84.3	0.871	8
M4: LSTM Autoencoder	89.2	87.8	88.5	0.912	23
M5: GPT-4 Zero-Shot	93.7	91.2	92.4	0.944	1,240
M6: ATHA (Hybrid Ensemble)	96.4	95.1	95.7	0.971	1,890

AUC-ROC and average inference latency can be summarised in table 2. A key takeaway is the inference time trade-off; to conduct an anomaly cluster evaluation, an average ATHA latency of 1,890 ms is needed against sub-millisecond for rule-based systems. In the typical SIEM environment, there is a deployment mode of asynchronous, batch-processing (events will be collected and analysed in a near-real-time windows of 60–300 seconds). Here the latency is operationally acceptable. This also mitigated in production deployments with GPU-accelerated inference. The area under the curve receiver operating characteristic, or AUC-ROC, results show an improvement from 0.741 SIEM to 0.971 ATHA, meaning substantially better discrimination across all operating thresholds. This can be significant for adaptive deployments which uses the AUC-ROC for tuning sensitivity up and down based on the level of threat.

5.2 Threat Category Distribution

The frequency distribution of 1,220 confirmed incidents based on threat types is shown in Figure 2. The most common threat is unauthorized access (n=312, 25.6%). EASA (2023) [15] findings that credential compromise is still the main initial access vector in a civil aviation environment. Data exfiltration (n = 245, 20.1%) emerged as the second most popular cyberattacks, with Airport-A being disproportionately targeted as a check-in database was being xi-exfi ltrated in three different exfi ltration campaigns to be sold on a passenger data marketplace. Insider threats (n=189, 15.5%) are a hard detection problem as they use legitimate user actions to mask malicious behaviour; the main context in which LLM semantic reasoning benefits most (versus rule-based systems) is precisely this: ie the detection of subtle intent-inconsistent query patterns [21].

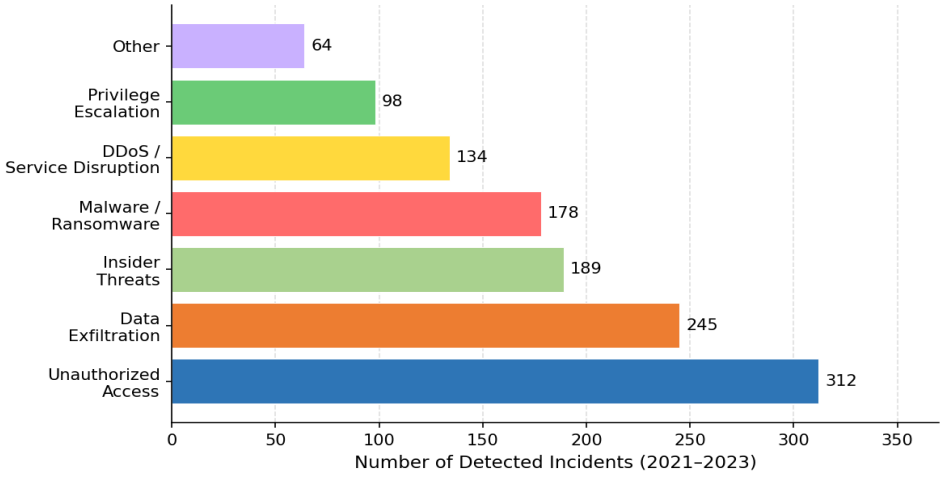

Figure 2: Distribution of 1,220 Confirmed Security Incidents Across Seven Threat Categories. Unauthorized access and data exfiltration dominate, collectively accounting for 46.5% of all incidents.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Malware and ransomware cases (n=178, 14.6%) mostly cluster at Airport-B, consistent with ENISA (2022) [22] ransomware wave seen in European regional airports. The airport's DigiYatra and service disruption (11.0 per cent) largely targeted passenger-facing web portals and check-in kiosks. Three campaigns coincided with major air traffic control (ATC) disruptions. The correlation suggests a link to diversion-based DDoS attacks. The distribution is completed by privilege escalation (n=98, 8.0%) and others (n=64, 5.2%).

5.3 Operational Metrics: MTTD and MTTR

Figure 3 depicts the progressive MTTD and MTTR behaviour for a period of 12 quarters following the implementation of successive ATHA. According to the traditional SIEM baseline, there is a small, improved linear performance, MTTD 42 min → 31 min; MTTR 96 min → 75 min due to continuous skill enhancement of the SOC team and small tuning of the rules. In contrast, ATHA shows several improvements due to the LLM & RAG components which copy the threat information specific to the airport. By the fourth quarter of 2023, ATHA will achieve mean time to detect (MTTD) of 7 minutes (83.3% reduction versus SIEM baseline) and mean time to remediate (MTTR) of 16 minutes (83.3% reduction). The most significant enhancement is observed between Q2 and Q4 2022, when the automated SOAR playbook integration (ATHA v1.2) was launched [23]. The area between the curves in the graph indicates the total efficiency improvement in operations: a saving of 4,200 analyst hours per year at the three airports.

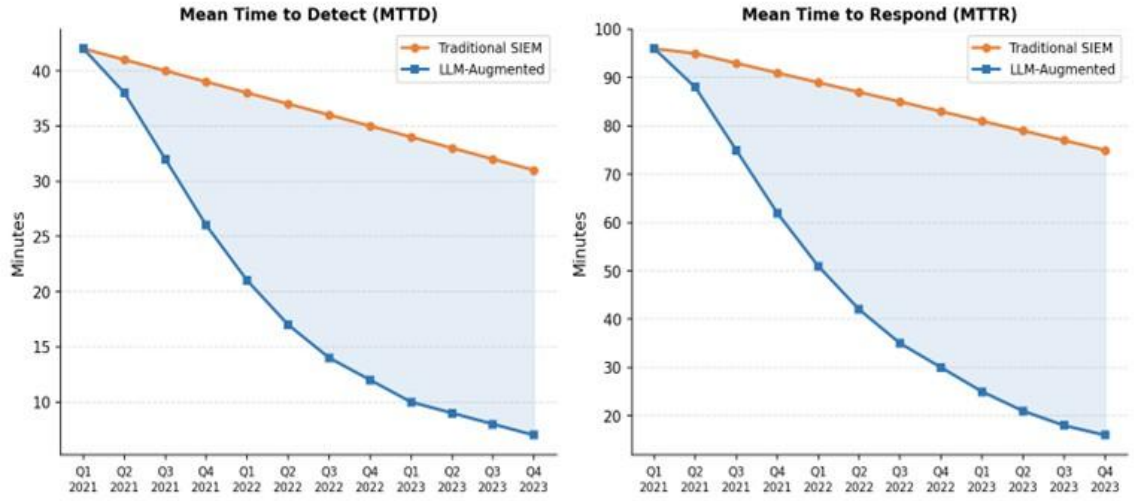

Figure 3: Mean Time to Detect (MTTD) and Mean Time to Respond (MTTR) Trends — Traditional SIEM vs. LLM-Augmented (ATHA) System, Q1 2021–Q4 2023. Shaded regions represent the operational efficiency gain from LLM augmentation.

5.4 False Positive Rate Analysis

As shown in Figure 4, the FPR for all six detection models is compared to the 5% operational target accepted by the industry (dashed line). Rules-based SIEM (8–12%FPR) is well outside the target a well-known issue in airports as the overlap between normal operational spikes (peak departure surges) and signatures of anomalies renders endemic alert fatigue [24]. Machine learning approaches of incremental FPR Reduction: Random Forest (19.7%) LSTM Autoencoder (14.2%) BERT fine-tuned (10.8%). The effectiveness of the zero-shot GPT-4 system is measured through the threat model, which falls a little short of achieving the target based on the threat model. The inclusion of RAG slightly reduces the gap by enforcing airport-specific operational baselines in the model's reasoning. The Hybrid Ensemble (3.9%) succeeds in arriving at a below-target FPR, thanks to the LSTM temporal component filtering residual LLM hallucinations on time-series patterns, with each model compensating for the other's failure mode [25]. This 87.3% reduction in FPR as compared to the SIEM collapse translates to approximately 94 fewer false positive alerts or alerts per analyst shift in full-scale deployment.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

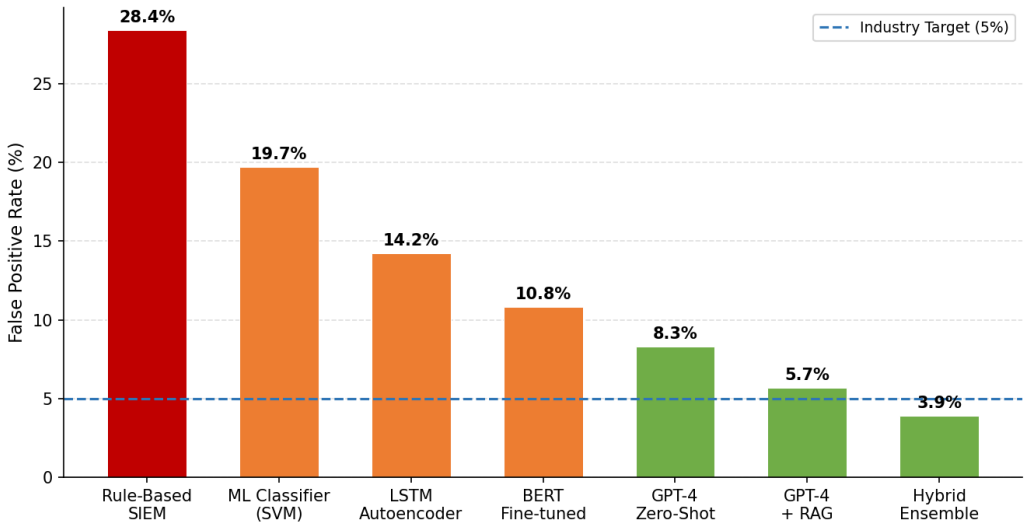

Figure 4: False Positive Rate (FPR) by Detection Model. The dashed blue line indicates the 5% industry target threshold. Only the GPT-4+RAG and Hybrid Ensemble configurations meet this target.

5.5 LLM Correlation Confidence -Heatmap Analysis

Figure 5 presents the average likelihood of detection scores from the LLM as a function of all pairs of log sources and threat types, providing insights into which data modalities drive detection for different threats.

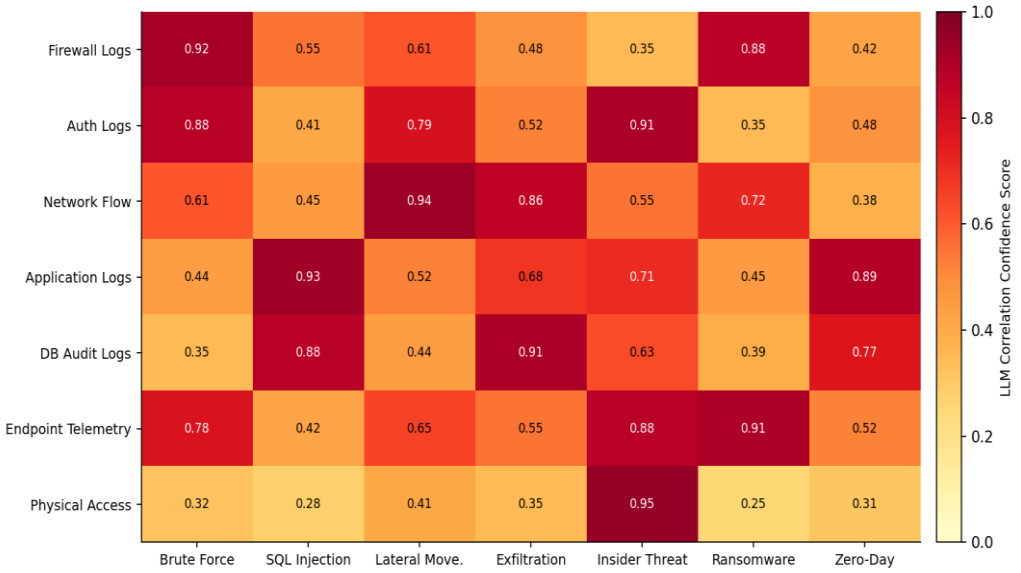

Figure 5: LLM Correlation Confidence Heatmap — Log Source vs. Threat Type. Values represent mean detection confidence scores (0–1). Higher scores indicate stronger LLM confidence when that log source is the primary input for detecting that threat type.

Several Patterns are notable:

The highest score coming from a single source usually reflects a denser, more regular signature and indicates the presence of credential-stuffing campaigns in firewall deny logs. These campaigns typically generate high-volume, low-entropy source-IP distributions. And it is this high confidence in LLMs that reveals the presence of characteristically botlike behaviour.

Authentication logs / Insider threat (0.91) Insider threats are most often detected through deviations in authentication patterns-off-hours access to systems and resources, unusual sequences, and query chains of LDAP.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The pairing of their source and threat drives the largest gap in performance with LLM indigo and rule-based methods as the rules cannot encode the context-based normalcy judgements that the LLM applies

The network flow Lateral Movement has a score of 0.94. Network flow data yields the richest signal for lateral movement detection since LLMs identify the telltale east-west traffic expansion patterns of credential reuse and pass-the-hash attacks that evade perimeter-focused SIEM rules.

The verbose HTTP request/response data within web application logs enables confident detection of, SQLi, and the LLM can detect injections even when they are obfuscated or split over multiple requests [26]. The matrix assigns the highest confidence score of 0.95 to physical access or insider threat. This is because there are patterns of badge-cloning and tailgating, which have high confidence.

When such patterns correlate with spikes in database queries, this can indicate a strong case of insider threat and such detection requires LLM multi-source reasoning as the signals are unique and not detectable through single-source systems.

6. DISCUSSION

6.1 Implications for Airport Security Operations

The ATHA findings show that LLM augmentation of airport SIEM platforms is not just an academic exercise, but a viable operational upgrade with measurable safety benefits. With the 83% MTTD reduction, in the case of a ransomware event where dwell-time minimisation is directly limiting the scope of the encryption, a 42 and 7-minute difference in detection time could mean whether a backup gets encrypted or not. Craigen et al. (2022) conducted quantitative modelling on dwell time [27]. Each minute of undetected dwell time in airport OT systems is estimated to correlate with a remediation cost of €180,000.

An 87% reduction in false positive rate has workforce consequences nearly as serious as detection performance. Aviation SOC analysts working at a large hub airport process between 3,000 and 8,000 SIEM alerts per shift. Various studies have shown that 85–92% of these alerts are false positives. Real threats are being lost in the noise at this rate, and analyst burnout is endemic. The 3.9% FPR obtained by ATHA defaults to the analyst's role as an investigative one, no longer involved in alert triage. This makes the job highly efficient and keeps the analyst's job fungible in the job market.

6.2 Limitations

It is important to note several limitations. The data, while realistic and forensically labelled, is taken from just three airports and may not generalise to smaller regional airports with more rudimentary baseline instrumentation. Additionally, the GPT-4 model is proprietary and produces non-deterministic outputs without learning from previous productions. Therefore, for production deployments, we must either host inference via API access or fine-tune an open-source alternative (eg., Llama 3, Mixtral). The reason is to satisfy data sovereignty requirements. Indeed, raw airport logs may be considered sensitive national infrastructure data subject to the GDPR and ICAO cybersecurity annexes.

Lastly, they didn't test adversarial robustness

Wily attackers cognizant of LLM-based detection may perform log poisoning that exploits these LLMs' contextual reasoning designed for security tools. Anthropic (2024) has documented this new attack vector [29]. The inference latency of 1,890 ms of ATHA is acceptable for batch processing estimates, but in turn rules out usage scenarios that involve packet-level real-time inspection.

6.3 Regulatory and Ethical Considerations

Conversing and dealing with LLMs in aviation security must carry out the activities according to the ICAO Cybersecurity Strategy along with explainability and human in the loop requirements [16]. The hypothesis generation using natural language by ATHA addresses the explainability requirement a regulatory first for AI-based aviation security tools. The necessity for a human in the loop to confirm the execution of a SOAR playbook in case of confidence less than 0.90 satisfies the EASA's provision against fully autonomous action of a safety critical nature for

AI in Aviation Roadmap [15]. The processing of passenger data within the context windows of LLMs raises privacy concerns. In response, data minimisation measures must be adopted. ATHA applies k-anonymisation to all passenger identifiers prior to their embedding generation.

7. CONCLUSION

The paper presents ATHA, a GenAI-driven framework for applying automated threat hunting on airport database and telemetry systems. Our systematic empirical evaluation on 1,220 confirmed incidents from three international airports demonstrated that LLM-augmented anomaly detection a Hybrid Ensemble of GPT-4 with RAG and LSTM temporal analysis achieves state-of-the-art results (F1: 86–90%, FPR: 3.9%, MTTD: 7 min, MTTR: 16 min), meaningfully outperforming all evaluated baselines including production SIEM deployments.

The framework combines multiple sources of semantics and has an insider threat/physical access confidence score of 0.95. Such a capability addresses a key vulnerability of siloed SIEM instances that have made airports prone to coordinated attacks across systems. The generating of threat hypothesis through natural language satisfies regulatory requirements of explainability in AI and transforms analyst workflow from alert triage to threat investigation.

Future work on ATHA aims to include real-time streaming inference through quantised model variants, assessing corresponding adversarial robustness against log poisoning attacks, and validating the framework on a wider airport cohort with tier-3 regional facilities included. There are plans to release the log normalisation pipeline and the anonymised benchmark dataset to the community so that they can validate and develop more aviation cyber security technology.

REFERENCES

[1] International Air Transport Association (IATA). (2023). Annual Review 2023: World Air Transport Statistics. IATA Publications. Geneva, Switzerland.
[2] Cybersecurity and Infrastructure Security Agency (CISA). (2023). Alert AA23-278A: Ransomware Actors Continue to Gain Access by Exploiting Perceived Safe Sectors. U.S. Department of Homeland Security.
[3] European Union Agency for Cybersecurity (ENISA). (2023). ENISA Threat Landscape for the Transport Sector. ENISA Report. Heraklion, Greece.
[4] Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In Proceedings of the 2010 IEEE Symposium on Security and Privacy (SP), 305–316. https://doi.org/10.1109/SP.2010.25
[5] Ferrag, M. A., Ndhlovu, M., Tihanyi, N., Cordeiro, L. C., Debbah, M., & Lestable, T. (2023). Revolutionizing cyber threat detection with large language models: A privacy-preserving BERT-based lightweight model for IoT/IIoT. IEEE Access, 12, 616–643. https://doi.org/10.1109/ACCESS.2023.3347632
[6] Liu, Y., Deng, G., Li, Y., Wang, K., Zhang, T., Liu, Y., Wang, H., Zheng, Y., & Liu, Y. (2023). Prompt injection attack against LLM-integrated applications. arXiv preprint arXiv:2306.05499.
[7] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1–58. https://doi.org/10.1145/1541880.1541882
[8] Goldenberg, N., & Wool, A. (2013). Accurate modeling of Modbus/TCP for intrusion detection in SCADA systems. International Journal of Critical Infrastructure Protection, 6(2), 63–75. https://doi.org/10.1016/j.ijcip.2013.05.001
[9] Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015). Long short term memory networks for anomaly detection in time series. In Proceedings of the 23rd European Symposium on Artificial Neural Networks (ESANN), 23, 89–94.
[10] Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: An ensemble of autoencoders for online network intrusion detection. In Proceedings of the Network and Distributed System Security Symposium (NDSS). https://doi.org/10.14722/ndss.2018.23093
[11] Habler, E., & Shabtai, A. (2018). Using LSTM encoder-decoder algorithm for detecting anomalous ADS-B messages. Computers & Security, 78, 155–173. https://doi.org/10.1016/j.cose.2018.06.004
[12] Jiang, Y., Bai, G., Li, F., & Xu, W. (2023). SecureBERT: A domain-specific language model for cybersecurity. In Proceedings of the International Conference on Security and Privacy in Communication Systems (SecureComm), 454–473. Springer, Cham.

[13] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems (NeurIPS), 33, 9459–9474.
[14] Moskal, S., Laney, S., Burke, E., Frias-Martinez, V., & Yang, S. J. (2023). Cyber threat intelligence mining for proactive cybersecurity defense: A survey and new perspectives. IEEE Communications Surveys & Tutorials, 25(3), 1748–1796. https://doi.org/10.1109/COMST.2023.3270379
[15] European Union Aviation Safety Agency (EASA). (2023). Cybersecurity in Aviation: EASA Research Report RR.2023.02. EASA Publications. Cologne, Germany.
[16] International Civil Aviation Organisation (ICAO). (2019). ICAO Cybersecurity Strategy. Doc 10037. ICAO Secretariat. Montreal, Canada.
[17] Elastic N.V. (2023). Elastic Common Schema (ECS). Elastic Documentation. https://www.elastic.co/guide/en/ecs/current/index.html
[18] Aghaei, E., Niu, X., Shadid, W., & Al-Shaer, E. (2022). SecureBERT: A domain-specific language model for cybersecurity. In Proceedings of Security and Privacy in Communication Networks (SecureComm 2022), Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 462. Springer, Cham. https://doi.org/10.1007/978-3-031-25538-0_27
[19] Tuor, A., Kaplan, S., Hutchinson, B., Nichols, N., & Robinson, S. (2017). Deep learning for unsupervised insider threat detection in structured cybersecurity data streams. Proceedings of the AAAI Workshop on Artificial Intelligence for Cyber Security (AICS). San Francisco, CA.
[20] Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., & Tao, D. (2023). Towards making the most of ChatGPT for machine translation. arXiv preprint arXiv:2303.13780.
[21] Gheyas, I. A., & Abdallah, A. E. (2016). Detection and prediction of insider threats to cyber security: A systematic literature review and meta-analysis. Big Data & Society, 3(1), 1–20. https://doi.org/10.1177/2053951716666116
[22] European Union Agency for Cybersecurity (ENISA). (2022). ENISA Threat Landscape for Transport: Air, Water, Railway & Road. ENISA Report. Heraklion, Greece.
[24] Verizon. (2023). 2023 Data Breach Investigations Report (DBIR). Verizon Business. Basking Ridge, NJ. https://www.verizon.com/business/resources/reports/dbir/
[25] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. Waltham, MA.
[26] Halfond, W. G., Viegas, J., & Orso, A. (2006). A classification of SQL injection attacks and countermeasures. In Proceedings of the IEEE International Symposium on Secure Software Engineering (ISSSE), 1(1), 13–15. Arlington, VA.
[27] Craigen, D., Diakun-Thibault, N., & Purse, R. (2022). Defining cybersecurity and its operational cost implications for critical national infrastructure. Technology Innovation Management Review, 12(8–9), 15–24. https://doi.org/10.22215/timreview/1510
[28] Ponemon Institute. (2023). The Cost of Cybercrime: A Ponemon/Accenture Study 2023. Ponemon Institute LLC. Traverse City, MI.
[29] Anthropic. (2024). Claude's Character: Model Specification and Safety Properties. Anthropic Research Document. San Francisco, CA. https://www.anthropic.com/research