JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

AI-DRIVEN THREAT INTELLIGENCE FRAMEWORK FOR SECURING NEXT-GENERATION CLOUD NETWORKS

Muhammad Hanif Hussain Khan [1], Zerminey Saleem [2], Rab Nawaz Safi [3], Muhammad Uzair [4]

Affiliations

[1] Department of Computer Science, Brains Institute, Peshawar, Pakistan
Email: muhammad.hanif.hussain.khan@gmail.com

[2] Department of Computer Science, Bahria University, Karachi, Pakistan
Email: zermineysaleem@gmail.com

[3] Department of Telecommunication Engineering, University of Engineering and Technology, Mardan, Pakistan
Email: rabnawazsafi321@gmail.com

[4] Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar, Pakistan
Email: muhammaduzi761@gmail.com

Corresponding Author's Email

[1]
muhammad.hanif.hussain.khan@gmail.com

License:

ABSTRACT

Next-generation cloud networks require security approaches capable of analysing heterogeneous observations and supporting timely, context-aware threat assessment. This study presents and evaluates an AI-driven threat intelligence framework for securing cloud-native environments by integrating network, identity, application, and infrastructure telemetry with contextual threat information. The framework comprises six interconnected layers: telemetry collection, data processing and correlation, AI-assisted detection, contextual enrichment, threat prioritisation and explanation, and policy-controlled response recommendations. It combines supervised machine learning, behavioural anomaly detection, and rule-based analysis to identify suspicious activities and generate evidence-supported alerts, while retaining human approval for potentially disruptive response actions. The framework was evaluated in an isolated cloud-native testbed using 10,000 labelled observation windows, comprising 8,000 benign and 2,000 malicious observations. Four configurations were compared through detection-performance, component-ablation, workload-adaptation, robustness, and processing-load experiments. The full framework achieved an F1-score of 95%, compared with 80% for the rule-based baseline. At a matched recall of 90%, it reduced the false-positive rate to 1.50%, compared with 4.00% for the comparison configuration. Under changing workloads, the F1-score declined by 4 percentage points for the controlled-update configuration, compared with 11 percentage points for the fixed model. Robustness experiments further examined performance under incomplete and delayed telemetry, feature perturbation, and training-data contamination. Processing performance remained complete up to 2,000 incoming events per second, while throughput decreased to 3,800 events per second at an input rate of 4,000 events per second, with 5%event loss. Analyst-level improvements were not established because participant evaluation results were unavailable. The findings demonstrate promising detection, contextual integration, and workload-adaptation performance within the evaluated environment, while highlighting operational limitations and the need for further validation across broader cloud settings. framework with three components: a cost-effective institutional implementation model, a tiered teacher-training roadmap, and a privacy- and ethics-compliant governance protocol tailored to the Pakistani regulatory context. The expected contribution is twofold. First, the study provides a quantified, tool-specific evidence base on the differential impact of AI on teaching and learning outcomes in a developing-country setting. Second, it offers a practical, context-adapted framework that policymakers and university administrations can use to scale AI integration equitably and sustainably. This work narrows the gap between perception-based advocacy and evidence-driven AI adoption in resource-constrained higher education systems.

Keywords: Cloud Security; Artificial Intelligence; Threat Intelligence; Machine Learning; Anomaly Detection; Multi-Source Telemetry; Explainable AI; Cloud-Native Networks.

I. INTRODUCTION

A. Background

The development of cloud-native applications and distributed computing environments has created a need for security architectures that extend beyond conventional network boundaries. Applications and services

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

can operate across on-premises infrastructure, public clouds, and multiple cloud providers, requiring security policies that remain effective regardless of resource location. Chandramouli and Butcher (2023) explain that protecting these environments requires a shift from implicit trust based on network location towards authentication and authorisation grounded in user, application, and service identities. Their publication, *A Zero Trust Architecture Model for Access Control in Cloud-Native Applications in Multi-Location Environments* (NIST Special Publication 800-207A), provides an architectural foundation for addressing security requirements in hybrid and multi-cloud environments [1], [2].

For this study, next-generation cloud networks refer to distributed environments incorporating cloud-native applications, containerised workloads, programmable infrastructure, and hybrid or multi-cloud deployments. The research focuses on the security relationships among network activity, identity operations, application behaviour, and infrastructure changes. Rather than considering each security event independently, the study examines an intelligence-driven approach that correlates these observations to support threat detection, investigation, prioritisation, and response recommendations.

Artificial intelligence offers a potential means of analysing security data and identifying behavioural patterns associated with malicious activity. Maasaoui et al. (2024), in *Anomaly Based Intrusion Detection Using Large Language Models*, published in the proceedings of the ACS/IEEE 21st International Conference on Computer Systems and Applications (AICCSA 2024), investigated a transformer-based approach to network intrusion detection. Their experiments used the UNSW-NB15, TON-IoT, and Edge-IIoT datasets and reported improvements over the traditional methods evaluated in the study. These findings demonstrate the potential of AI-assisted detection under defined experimental conditions, although they do not establish equivalent performance across operational cloud environments [3], [4].

Building on this potential, the present study develops and evaluates an AI-driven threat intelligence framework that combines multi-source cloud telemetry, behavioural analysis, and contextual threat information. Its function is not merely to classify individual observations but also to relate suspicious activity to affected identities, applications, and resources. This distinction is important because an accurate prediction must also provide sufficient context to support an appropriate security decision. The framework therefore integrates detection, contextual enrichment, threat prioritisation, and evidence-supported response recommendations [5].However, incorporating AI introduces additional security concerns. Vassilev et al. (2025), in *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* (NIST AI 100-2 E2025), identify threats including data poisoning, evasion, and privacy breaches [4], [5]. Their taxonomy considers attacker objectives, capabilities, knowledge, and the stages at which attacks can affect machine-learning systems. Consequently, the framework also considers threats against its own data-processing and analytical components, in addition to threats against cloud resources.

Against this background, this study evaluates an integrated AI-driven threat intelligence framework through controlled experiments in an isolated cloud-native testbed. The empirical evaluation uses 10,000 labelled observation windows, comprising 8,000 benign and 2,000 malicious observations, and compares four configurations: a rule-based baseline, a network-only machine-learning model, a multi-source model without contextual enrichment, and the full framework [6], [7]. The evaluation examines detection performance, false-positive rates, detection and prioritisation delays, component contributions, performance under changing workloads, robustness under selected adverse conditions, and processing overhead. It also considers evidence-supported recommendations and human oversight while distinguishing the framework's analyst-support objectives from outcomes that have not been empirically established [7].

B. Problem Statement

The central problem addressed by this study is how to transform heterogeneous cloud security observations into timely, contextual, and dependable threat intelligence. Identity-aware security architectures provide a foundation for controlling access across distributed applications, but an additional challenge is determining how observations from these environments can be integrated to support threat assessment. Chandramouli and Butcher (2023) establish the importance of application and service identities in multi-location cloud security; this study extends that architectural concern into the design and empirical evaluation of an integrated threat-analysis process [1], [8].

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

Research on AI-assisted intrusion detection provides promising experimental evidence, but its relevance to an integrated cloud security framework requires careful interpretation. For example, the results reported by Maasaoui et al. (2024) concern particular datasets, model configurations, and classification tasks. Whether comparable benefits can be achieved when network, identity, application, and infrastructure observations are analysed together requires empirical assessment. In particular, the contribution of multi-source telemetry and contextual enrichment needs to be distinguished from the performance of individual detection components [2], [3].

A further challenge concerns the trustworthiness and operational reliability of the analytical system itself. As documented by Vassilev et al. (2025), machine-learning systems can be targeted through attacks that manipulate data or influence model behaviour. A cloud security framework that relies on AI must therefore be assessed not only in terms of detection performance but also through its behaviour under selected adverse conditions, including incomplete or delayed telemetry, feature perturbations, and training-data contamination [9]. Accordingly, this study addresses an integration and evaluation problem: whether a unified framework can combine multiple evidence sources and contextual information to improve threat detection, interpretation, and prioritisation without introducing excessive false alarms or operational costs. The empirical investigation compares the framework with selected baseline configurations and examines the contribution of individual components through ablation experiments. It also evaluates performance under changing workloads and defined robustness conditions [10], [11].

The study's contribution is based on the design and experimental assessment of a framework integrating telemetry collection, AI-assisted analysis, contextual threat enrichment, prioritisation, and evidence-supported recommendations. It does not assume that no comparable frameworks exist, nor does it claim universal effectiveness across all cloud environments. Its findings are interpreted within the boundaries of the implemented testbed, observation population, configurations, and experimental conditions. High-impact response actions remain subject to predefined policies and human approval.

C. Research Questions

This study is guided by four interrelated research questions:

RQ1: How can heterogeneous cloud telemetry and contextual threat information be integrated into an AI-driven framework for detecting and interpreting malicious activity?

RQ2: To what extent does the framework improve detection performance, false-positive rates, and detection and prioritisation delays compared with selected rule-based and standalone machine-learning baselines?

RQ3: How effectively does the framework maintain its detection performance under changing workloads, incomplete or delayed telemetry, bounded feature perturbations, and controlled training-data contamination?

RQ4: How can evidence-supported threat prioritisation and policy-controlled response recommendations be incorporated into the framework while maintaining acceptable processing performance and operational overhead?

These questions connect the framework's architectural design with its predictive performance, robustness, and operational characteristics. RQ4 also addresses the intended role of evidence-supported decision support. However, conclusions about improvements in analyst decision-making require participant-level evaluation evidence and are not established by the available empirical results.

D. Research Objectives

The primary objective of this study is to design and evaluate an AI-driven threat intelligence framework for securing next-generation cloud networks through integrated telemetry analysis, contextual threat interpretation, and controlled decision support.

Specifically, the study pursues five objectives:

1. **Objective 1:** To define the framework's operational scope, security requirements, and threat model, including the resources, data sources, adversary behaviours, and trust assumptions considered in the evaluation.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

2. **Objective 2:** To design and implement a data-processing pipeline that collects, normalises, and correlates selected network, identity, application, and infrastructure observations with contextual threat information.
3. **Objective 3:** To develop and evaluate an AI-assisted analysis mechanism that combines behavioural anomaly detection, threat classification, and contextual interpretation.
4. **Objective 4:** To implement interpretable threat prioritisation and policy-controlled response recommendations supported by relevant evidence and human oversight.
5. **Objective 5:** To evaluate the framework against selected baselines using precision, recall, F1-score, false-positive rate, detection and prioritisation delays, computational resource consumption, and robustness under defined test conditions.

The evaluation distinguishes between individual component performance and end-to-end framework performance. Baseline comparisons and component-ablation experiments examine the contributions of multi-source telemetry, contextual enrichment, and selected detection components. Additional experiments assess workload-related performance changes, robustness under specified adverse conditions, and processing overhead. The study also defines an analyst-support evaluation approach; however, because participant-level evaluation results are unavailable, improvements in analyst triage accuracy and decision time are not treated as established outcomes.

E. Hypotheses

The following hypotheses are formulated for empirical assessment. They represent anticipated relationships rather than assumptions of guaranteed improvement. Their assessment is based on the comparisons and evidence available from the reported experiments.

H1: The full framework, combining cloud telemetry with contextual threat information, achieves a higher detection F1-score than the selected rule-based baseline under the evaluated conditions.

H2: Correlating network, identity, and application observations reduces the false-positive rate compared with an otherwise equivalent network-only model at a comparable recall level.

H3: Incorporating contextual threat information reduces the time required to detect and prioritise the evaluated malicious activities compared with a telemetry-only configuration.

H4: A controlled model-update mechanism produces a smaller decline in detection performance under changing workload conditions than a model trained once and left unchanged.

H5: Evidence-supported explanations improve analysts' triage accuracy and reduce their median decision time compared with alerts that provide only a risk score or predicted threat label.

H1 is examined through the F1-score comparison between the full framework and the rule-based baseline. H2 is examined through false-positive rates at a comparable recall level. H3 is assessed through the reported detection and prioritisation delay comparisons, although the exact numerical delay values are not available in the reported results. H4 is examined through the difference in performance decline between controlled-update and fixed-model configurations under changing workloads. H5 requires a structured analyst evaluation with participant-level triage outcomes. Since such results are unavailable, H5 remains untested and is not presented as empirically supported [12].

The empirical assessment is based on the reported testbed observation population and experimental configurations. The results are interpreted at the observation-window level rather than being treated as incident-level performance. The study also distinguishes descriptive performance differences from statistical significance: no p-values or confidence intervals are claimed where the corresponding run-level statistical calculations are unavailable.

F. Significance of the Study

This study contributes to cloud security research by examining AI-driven threat intelligence as an integrated process connecting data collection, detection, contextual interpretation, prioritisation, and decision support.

Theoretical Contribution: The study develops and evaluates a conceptual framework explaining how cloud telemetry, behavioural analysis, contextual threat information, and security decision-making can be connected within a unified architecture. Its emphasis on identity and application context is informed by the

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

security principles described by Chandramouli and Butcher (2023) for multi-location cloud-native environments. The framework also recognises that detection performance and operational usefulness are related but distinct dimensions of cloud security [1], [13].

Methodological Contribution: The research evaluates security effectiveness using multiple dimensions rather than classification accuracy alone. Detection quality is considered alongside false-positive rates, detection and prioritisation delays, processing performance, resource consumption, workload generalisation, and robustness under selected adverse conditions. Comparisons among four framework configurations and component-ablation experiments help distinguish the contribution of multi-source telemetry, contextual enrichment, and individual detection components. This approach also distinguishes the experimental potential demonstrated by AI-based intrusion detection research from the additional evidence required to support the practical use of an integrated cloud framework [2].

Technical Contribution: The framework investigates how telemetry integration, AI-assisted analysis, contextual enrichment, threat prioritisation, and evidence-supported recommendations operate together. The experimental comparisons examine the full framework against rule-based, network-only, and multi-source configurations without contextual enrichment. Component-ablation findings further identify how removing identity telemetry, application telemetry, contextual enrichment, or anomaly detection affects the reported detection metrics. These findings provide evidence about the contribution of the evaluated components within the implemented testbed rather than establishing universal performance across cloud architectures [15].

Practical Contribution: The framework is designed to support security analysts by presenting prioritised alerts with relevant observations and explanatory context. Its response recommendations preserve human oversight for actions that could disrupt legitimate services. The evaluation examines detection and prioritisation performance and the processing resources required to operate the framework. Although analyst decision support is an intended practical benefit, improvements in triage accuracy and decision time cannot be claimed because participant-level analyst evaluation results are unavailable.

Security and Governance Contribution: The study considers selected threats against its own AI components, including bounded feature perturbations, incomplete or delayed telemetry, and controlled training-data contamination. This reflects the life-cycle perspective presented by Vassilev et al. (2025), which connects attacker capabilities and objectives with vulnerabilities in machine-learning systems. The reported robustness findings are limited to the adverse conditions actually evaluated and do not establish protection against all adversarial attacks or previously unseen threats. The framework's recommendation-based response design and human-approval requirements also recognise the importance of controlled security operations [16].

The overall significance of the research lies in determining the conditions under which AI-driven threat intelligence provides measurable improvements in cloud network security. By evaluating detection performance, false-positive reduction, workload-related changes, robustness, and processing overhead, the study provides an empirical assessment of the framework within a defined cloud-native testbed. It also makes the boundaries of the evidence explicit, particularly where analyst-level outcomes and broader generalisation have not been established.

II. RESEARCH METHODOLOGY

A. Introduction

This chapter describes the research design, framework architecture, data sources, experimental environment, analytical procedures, evaluation metrics, and ethical considerations underlying the study, *AI-Driven Threat Intelligence Framework for Securing Next-Generation Cloud Networks*. The methodology was implemented to assess whether integrating cloud telemetry with contextual threat information improves threat detection, prioritisation, and operational performance compared with selected baseline approaches.

The research combined framework development with controlled experimental evaluation. The empirical evaluation was conducted in an isolated cloud-native testbed using labelled observations generated and collected within the experimental environment. The evaluation focused on network, identity, application, and infrastructure observations and examined their contribution to the integrated framework. Publicly available intrusion-detection research, including studies using benchmark datasets, informed the research context; however, separate public-benchmark performance results are not included in the reported empirical

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

comparisons. The reported quantitative findings are therefore based on the testbed observation population described in Chapter 3 [17].

B. Research Design

The study adopted a design-and-evaluation approach supported primarily by quantitative experiments. It involved the development of an AI-driven threat intelligence framework and its assessment through baseline comparisons, component-ablation experiments, workload-change evaluation, robustness tests, and processing-performance measurements. The research was organised around six phases:

1. Requirements analysis and threat modelling,
2. Data acquisition and preparation,
3. Framework design and implementation,
4. Model development and validation,
5. Controlled experimental evaluation,
6. Statistical analysis and interpretation.

The independent experimental dimensions included framework configuration, telemetry sources, contextual enrichment, model-update strategy, workload conditions, and selected adverse conditions. The dependent measures included detection performance, false-positive rate, detection and prioritisation delays, computational overhead, and processing performance. Analyst triage outcomes were identified as intended evaluation measures, but participant-level results were unavailable; consequently, H5 was not tested.

The study distinguished between improvements in predictive accuracy and improvements in operational usefulness. Framework effectiveness was therefore assessed using detection metrics alongside processing latency, resource consumption, and event-processing capacity. A high classification score alone was not treated as sufficient evidence of operational suitability.

C. Research Scope and Threat Model

1) Operational Scope. The framework was implemented and evaluated within an isolated, cloud-native laboratory testbed incorporating:

- Containerised application workloads,
- Network traffic collection,
- Identity and access observations,
- Application and service logs,
- Infrastructure and orchestration observations,
- A dedicated security analytics environment.

The experimental environment consisted of three virtual machines, each allocated 8 vCPUs and 16 GB of RAM. The testbed provided the environment for collecting and processing network, identity, application, and infrastructure observations. The evaluation represents a bounded cloud-native environment rather than every possible hybrid or multi-cloud deployment. Accordingly, the findings are interpreted within the implemented environment and do not establish portability across cloud platforms that were not separately evaluated [18].

Identity and service context informed the framework's analytical design. This architectural choice is consistent with Chandramouli and Butcher's (2023) emphasis on application and service identities when securing cloud-native applications across multiple locations.

2) Threat Scenarios. The research considered cloud security behaviours relevant to the framework's detection and contextual-analysis objectives. These included network-related, identity-related, application-related, and infrastructure-related suspicious activity. The threat-model scope also considered behaviours such as:

- Network reconnaissance,
- Repeated authentication failures,
- Suspicious use of valid credentials,
- Unauthorised privilege-change attempts,
- Abnormal service-to-service communication,
- Unusual outbound data-transfer patterns,

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

- Resource-exhaustion activity within predefined safety limits.

These behaviours describe the security scope considered in the research. The quantitative findings are interpreted according to the labelled observation population and experimental conditions reported in Chapter 3; the results do not establish that every listed behaviour was separately evaluated or detected with an independently measured performance score [18], [19].

The study does not claim to evaluate all cloud threats. In particular, performance under workload changes or bounded feature perturbations is not treated as proof of detecting genuine zero-day vulnerabilities or previously unseen attacks.

3) Threats to the AI Components

The framework's robustness was examined under selected adverse conditions, including incomplete telemetry, delayed telemetry, bounded feature perturbations, and controlled training-data contamination. These conditions reflect the distinction between evasion and poisoning attacks described by Vassilev et al. (2025). The reported findings are restricted to the specific conditions and levels of perturbation included in the experiments; they do not establish universal resistance to adversarial manipulation [3].

D. Data Sources and Collection

1) Public Benchmark Data. Public intrusion-detection datasets, including UNSW-NB15, provide an established research context for AI-assisted network intrusion detection. The official UNSW-NB15 repository describes normal activity and synthetic attack behaviour, including nine attack categories, extracted network features, and class labels. The dataset is relevant to the development and evaluation of network intrusion-detection methods.

However, the reported empirical evaluation in this study is based on the cloud testbed observation population described in Section II-D2. Separate UNSW-NB15 benchmark performance results are not reported in Chapter 3 because corresponding benchmark model outputs are not available in the reported results. Consequently, no independent benchmark-performance claim is made in this study, and the public dataset is not treated as evidence for the integrated framework's reported cloud-testbed performance [20].

2) Cloud Testbed Data. The primary empirical dataset comprised 10,000 labelled observation windows collected and evaluated within the cloud-native testbed. The observation population contained 8,000 benign windows and 2,000 malicious windows. The same observation population and corresponding labels were used across the four experimental configurations to maintain comparability.

The testbed incorporated the following observation sources:

TABLE I
DATA SOURCES IN THE CLOUD TESTBED

Data source	Observations	Analytical purpose
Network telemetry	Flow duration, traffic volume, protocol, and connection patterns	Identify suspicious communication behaviour
Identity	Authentication outcomes, account	Identify abnormal identity

3) Ground-Truth Construction

The empirical evaluation used labelled observation windows classified as benign or malicious. These labels provided the reference outcomes for calculating true positives, true negatives, false positives, and false negatives.

Ground-truth interpretation was kept separate from the framework's predicted outputs so that detector predictions were not treated as their own reference labels. The evaluation was conducted at the observation-window level, consistent with the unit of analysis used in the reported dataset and performance tables. The reported results do not establish independent incident-level ground-truth validation. Accordingly, the findings are interpreted as classification performance against the labelled observation population rather than as independently verified incident-level detection performance.

E. Data Preparation

Data preparation followed a consistent processing approach to support comparable evaluation across the four configurations.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

1) Cleaning and Normalisation. The data-processing approach accounted for malformed records, inconsistent timestamps, duplicate observations, and missing values. Duplicate handling distinguished repeated copies of the same observation from legitimate recurring events. Missing telemetry was treated as a relevant data condition rather than automatically interpreted as normal behaviour.

The observations were processed into a form suitable for correlation and analysis across the network, identity, application, and infrastructure sources.

2) Feature Engineering. The framework's analytical design incorporated information derived from individual observations and correlated activity windows. Relevant feature categories included: connection frequency, bytes transferred per interval, authentication-failure counts, frequency of privilege changes, number of distinct resources contacted, deviations from previous account or service behaviour, and temporal relationships among network, identity, and application observations.

Feature interpretation was restricted to information available for the relevant detection decision. Information derived from future observations or completed evaluation outcomes was not treated as contemporaneous detection evidence.

3) Encoding and Scaling. The data-processing pipeline accounted for the requirements of categorical and numerical observations. The evaluation design distinguished model-development data from validation and testing data to reduce the risk of information leakage. The reported results do not provide a separate numerical comparison of alternative encoding, scaling, or imputation procedures [22].

The dataset retained its reported class distribution of 8,000 benign and 2,000 malicious observation windows. The results were not based on replacing the reported test observation population with an artificially balanced class distribution.

4) Data Partitioning and Leakage Prevention. The testbed observation population was divided chronologically into training, validation, and testing subsets using a 60:20:20 allocation. This approach maintained the temporal ordering of observations and provided separate data subsets for model development, validation, and final evaluation.

The evaluation design also recognised the importance of preventing information leakage between related observations. The observation-window unit was maintained consistently across configurations, and the same labelled population was used for comparative evaluation [23].

The data-partitioning approach considered the following potential sources of leakage: duplicate records across partitions, overlapping temporal windows, label-derived features, repeated scenario identifiers, and features that reveal experimental scheduling rather than malicious behaviour.

Model selection and threshold-related decisions were associated with the development and validation stages rather than being based on the final test outcomes. The reported results are interpreted within this chronological partitioning approach.

F. Framework Architecture

The implemented framework comprised six interconnected layers.

1) Telemetry Collection Layer

2) Processing and Correlation Layer. Incoming records were processed into a common analytical representation. Correlation connected observations associated with accounts, workloads, services, resources, and investigation windows. This layer provided the multi-source observation structure used by the integrated framework.

3) AI-Assisted Detection Layer. The detection layer combined AI-assisted classification and anomaly-detection capabilities with rule-based security conditions. Its analytical functions supported the identification of labelled malicious behaviours and deviations from benign activity.

The experimental configurations distinguished the rule-based baseline, network-only machine-learning model, multi-source model without contextual enrichment, and full framework. The component-ablation evaluation also examined the contribution of the anomaly detector. The reported results do not provide a complete set of model-specific implementation details for every algorithm; therefore, no additional model names or unreported algorithm configurations are asserted here.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

4) Contextual Enrichment Layer. This layer associated detection outputs with relevant account information, resource context, related observations, and recognised adversary behaviours. MITRE ATT&CK provided a structured vocabulary for describing relevant techniques. Technique associations were treated as contextual information rather than proof that a particular adversary conducted an event [24], [25].

The experimental comparison between the multi-source model without contextual enrichment and the full framework examined the contribution of contextual information to the reported detection and prioritisation outcomes.

5) Threat Prioritisation and Explanation Layer. The prioritisation function combined detection evidence with contextual factors, including asset criticality, identity sensitivity, and corroborating observations. The prioritisation score was represented as:

$$R_i = w_1 D_i + w_2 A_i + w_3 I_i + w_4 C_i$$

The score represents prioritisation rather than necessarily indicating the probability that an alert is malicious. The framework's explanation function was designed to identify the observations and contextual factors contributing to an alert. Model-derived importance was distinguished from independently verified causal explanation.

The reported empirical findings focus on detection performance, prioritisation delays, component contributions, and operational measurements [3]. They do not establish a separate participant-level measure of explanation quality or analyst decision improvement.

6) Response Recommendation Layer. The framework provided a recommendation-oriented response approach, including investigation, additional logging, account review, and temporary access restrictions. Disruptive actions remained subject to human approval.

The evaluation distinguished recommendation and prioritisation functions from actual containment effectiveness. No claim is made that the reported experiments established successful automated containment of attacks.

G. Model Training and Selection

Model development and evaluation followed the training, validation, and testing structure described in Section II-E4. The framework configurations were compared using the same labelled observation population, allowing differences in reported performance to be interpreted in relation to the evaluated configuration.

The evaluation considered the following measures: precision and recall, F1-score, false-positive rate, inference and processing latency, and memory and CPU consumption.

Detection thresholds and configuration decisions were considered in relation to the validation process. For comparisons focused on false-positive reduction, configurations were assessed at comparable recall levels where applicable.

The reported results do not provide a complete record of every candidate algorithm, hyperparameter search, model seed, or tuning budget. Therefore, the methodological interpretation is restricted to the configurations and performance measures actually reported rather than introducing unverified model-selection procedures.

The controlled-update and fixed-model configurations were compared under changing workload conditions. The purpose of this comparison was to examine the difference in performance decline when workload conditions changed. The reported results are interpreted as performance differences under the evaluated conditions rather than evidence of general suitability for every adaptive-learning environment.

H. Experimental Design

1) Baseline Comparisons. Four configurations were evaluated using the same labelled observation population:

TABLE II
BASELINE CONFIGURATIONS

Configuration	Purpose
Rule-based baseline	Establish performance using predefined detection conditions

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

Network-only machine-learning baseline	Measure the contribution of AI without multi-source correlation
Multi-source model without contextual enrichment	Measure the contribution of telemetry integration independently of contextual enrichment
Full framework	Evaluate the combined detection, enrichment, and prioritisation capabilities

Using the same observation population supported direct comparison among configurations. The comparisons focused on detection performance, false-positive rates, and the contribution of multi-source and contextual information.

2) Ablation Experiments. Component-ablation experiments examined the contribution of selected framework components by evaluating configurations in which individual components were removed from the full framework.

The evaluated ablation conditions included:

- Removal of identity telemetry.
- Removal of application telemetry.
- Removal of contextual enrichment.
- Removal of the anomaly detector.

The resulting precision, recall, F1-score, and false-positive rates were compared with the full framework. These experiments helped distinguish the contribution of individual components from the performance of the integrated configuration.

The reported ablation results do not include separate numerical ablations for alert explanations or adaptive updating. Their effects are therefore not presented as independently established component-level findings.

3) Generalisation and Robustness Experiments. The evaluation included workload-change and robustness experiments.

The workload-change comparison examined the performance of a fixed model and a controlled-update model under changing workload conditions. The primary measure was the change in F1-score between the reference and changed workload conditions.

Robustness experiments examined the following conditions:

1. Random removal of a proportion of telemetry observations.
2. Delayed arrival of selected telemetry sources.
3. Bounded feature perturbation.
4. Controlled contamination of training data.

The performance under each adverse condition was compared with the reference condition. The analysis focused on changes in F1-score and related detection performance. These findings are limited to the evaluated perturbation levels and do not establish performance under all possible adversarial conditions.

The study also evaluated processing performance under increasing event-processing rates. This load evaluation examined throughput, processing latency, CPU utilisation, memory consumption, and event loss.

Separate results for previously unseen attack scenarios are not included in the reported empirical findings. Accordingly, the workload-change and robustness results are not interpreted as proof of generalisation to unseen attack types.

4) Experimental Repetition. Each experimental configuration was evaluated across ten independent experimental runs. This repetition structure provided repeated observations of the framework's performance under the defined experimental conditions.

The experimental repetition unit was distinguished from the individual observation-window unit. The 10,000 observation windows were not treated as 10,000 independent experimental replications. This distinction is important because correlated observations may originate from the same workload or activity sequence.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

The reported findings are presented according to the available experimental results. Statistical significance, confidence intervals, or run-level variability are not claimed where the corresponding calculations are not available.

I. Evaluation Metrics

Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

1) Detection Performance. Detection performance was assessed using precision, recall, F1-score, and false-positive rate:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{FalsePositiveRate} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

The metrics were calculated at the observation-window level using the labelled observation population. The same observation population was used across the four configurations. Accordingly, the reported metrics describe window-level classification performance rather than incident-level detection effectiveness.

2) Detection and Prioritisation Delay. Detection delay was defined as the time between the beginning of the relevant malicious activity and the first qualifying alert:

$$\text{DetectionDelay}_i = t_{\text{first qualified alert},i} - t_{\text{attack start},i}$$

Prioritisation delay measured the time from attack onset to the first alert meeting the predefined investigation threshold.

The evaluation distinguished detection and prioritisation delays from processing latency. Detection delay concerns the time required to identify malicious activity, whereas processing latency concerns the time required for the analytics service to process incoming observations.

The results chapter reports the detection and prioritisation delay comparison, but the exact numerical delay values are not available in the supplied results text. Consequently, no additional numerical values are introduced in this methodology.

3) Operational Performance. Operational performance was assessed through:

- Events processed per second.
- Median and upper-percentile processing latency.
- CPU and memory utilisation.
- Event loss under load.
- Processing performance under increasing input rates.

The load measurements referred to the security analytics service and its allocated resources rather than the combined capacity of all three virtual machines. The evaluation distinguished the service's processing capacity from the total infrastructure resources available in the testbed.

4) Robustness. Performance degradation under adverse conditions was measured as:

$$\Delta F_1 = F_{1,\text{reference}} - F_{1,\text{stress}}$$

The reference and stress conditions were defined separately for each robustness experiment. The analysis compared F1-score changes under telemetry removal, telemetry delay, bounded feature perturbation, and training-data contamination.

The reported degradation values were interpreted only in relation to the tested conditions. They were not treated as a universal measure of adversarial robustness.

J. Analyst Evaluation

The research identified a structured analyst evaluation as the method for assessing Hypothesis 5. The intended comparison involved two alert conditions:

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

1. Alerts containing a risk score and predicted label.
2. Alerts containing the same information together with supporting evidence and explanations.

The intended primary outcomes were triage accuracy against independently reviewed ground truth and the time taken to reach a decision. Participant experience and case difficulty were also identified as relevant evaluation considerations.

However, participant-level analyst evaluation results are unavailable in the reported empirical findings. No participant count, triage accuracy, median decision time, or user-feedback results are therefore reported. Consequently, the analyst evaluation does not provide empirical evidence for H5, and this hypothesis remains untested. The framework's explanation and analyst-support capabilities are discussed as design features rather than as demonstrated improvements in human decision-making.

K. Statistical Analysis and Hypothesis Testing

The analysis focused on descriptive and comparative evaluation of detection outcomes, processing delays, resource use, workload-related performance changes, and robustness results.

The four framework configurations were compared using the same labelled observation population. Component-ablation results were interpreted in relation to the full framework, while workload-change results were assessed by comparing the performance decline of the controlled-update and fixed-model configurations.

The hypotheses were operationalised as follows:

TABLE III
HYPOTHESIS OPERATIONALISATION

Hypothesis	Main comparison	Primary outcome
H1	Full framework versus rule-based baseline	Detection F1-score
H2	Multi-source versus network-only model	False-positive rate at comparable recall
H3	Context-enriched versus telemetry-only configuration	Detection and prioritisation delay
H4	Controlled-update versus fixed model	Performance decline after workload change
H5	Explained versus score-only alerts	Analyst accuracy and decision time

H1 was assessed using the reported F1-score comparison between the full framework and the rule-based baseline. H2 was assessed using the false-positive-rate comparison at a comparable recall level. H3 was considered in relation to the reported detection and prioritisation delay comparison, although exact numerical delay values were not available in the supplied results. H4 was assessed through the reported performance decline under changing workload conditions. H5 remained untested because participant-level analyst evaluation results were unavailable.

The reported results do not provide p-values, confidence intervals, or sufficient run-level statistical output to establish statistical significance for the reported comparisons. Accordingly, the findings are interpreted as reported descriptive performance differences rather than as statistically significant effects. The ten independent experimental runs provide a repetition structure, but no inferential statistical claim is made without the corresponding run-level calculations.

L. Validity and Reproducibility

The study ensured internal validity through a controlled testbed, consistent labelled observations, chronological 60:20:20 data partitioning, and ten independent experimental runs. Construct validity was maintained by using appropriate detection, timing, and operational performance metrics. External validity remained limited to the evaluated workloads, architecture, configurations, and test conditions, without claiming universal applicability. Reproducibility was supported through documentation of dataset versions, checksums, model parameters, random seeds, software versions, experimental scenarios, and evaluation settings.

M. Ethical and Security Considerations

All security experiments were conducted within researcher-owned or authorised infrastructure using non-sensitive data and controlled resources. Appropriate access controls, monitoring, and resource limits were applied to prevent harm to external systems. Human approval was retained for disruptive security responses,

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

ensuring that AI recommendations did not automatically trigger production changes. The study also acknowledged the limitations of automated predictions and restricted adversarial testing to the research environment.

III. RESULTS AND ANALYSIS

A. Introduction

This chapter presents the empirical results and analysis of the proposed AI-driven threat intelligence framework for securing next-generation cloud networks. The evaluation examines the framework's detection performance, false-positive rate, detection and prioritisation delays, contribution of individual components, performance under changing workloads, robustness under adverse conditions, and computational overhead. The analysis follows the research questions, objectives, and hypotheses established in the preceding chapters.

Four configurations are considered: a rule-based baseline, a network-only machine-learning model, a multi-source model without contextual enrichment, and the full proposed framework. These configurations provide a basis for examining the contribution of multi-source telemetry, contextual enrichment, and integrated threat-detection components.

The experimental evaluation uses 10,000 labelled observation windows, comprising 8,000 benign and 2,000 malicious windows. Experiments were conducted in an isolated cloud-native test environment consisting of three virtual machines, each configured with eight vCPUs and 16 GB of RAM. Ten independent experimental runs were performed for each configuration. The chapter reports the available quantitative findings and interprets them in relation to the study's hypotheses. Analyst-centred evaluation is treated separately because participant-level results are not available.

B. Experimental Evaluation Configuration

1) Experimental Environment. The evaluation was conducted in an isolated cloud-native laboratory environment containing containerised applications, identity services, telemetry collectors, and a security analytics pipeline. The environment was configured to support the collection and analysis of network, identity, application, and infrastructure observations.

TABLE IV
EXPERIMENTAL EVALUATION ENVIRONMENT

Component	Experimental configuration
Deployment environment	Isolated cloud-native laboratory testbed
Compute nodes	Three virtual machines
Resources per node	Eight vCPUs and 16 GB RAM
Application environment	Containerised web application and supporting services
Telemetry sources	Network, identity, application, and infrastructure observations
Data partitioning	Chronological 60% training, 20% validation, and 20% testing
Primary classification unit	Correlated observation window
Experimental repetitions	Ten independent runs per configuration
Response policy	Recommendations only; disruptive actions require approval

The processing-performance measurements reported later in this chapter refer to the analytics service and its allocated resources rather than the combined capacity of all three virtual machines. The use of an isolated test environment provides a controlled setting for comparing the four configurations while maintaining a consistent experimental arrangement.

2) Experimental Dataset and Observation Population. The experimental evaluation comprised 10,000 labelled observation windows, including 8,000 benign windows and 2,000 malicious windows. The same observation population and corresponding labels were used across the four configurations to maintain comparability.

An observation window represents a correlated unit of activity rather than an entire attack incident. A single incident may generate multiple malicious observation windows, while benign operational activity may also produce multiple windows. Therefore, the detection metrics reported in this chapter represent window-level classification performance and should not be interpreted directly as incident-level detection rates.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

The dataset was divided chronologically into 60%training, 20%validation, and 20%testing subsets. This partitioning arrangement was used to support model development and evaluation while maintaining the temporal ordering of observations.

3) Evaluation Configurations and Reporting Scope

1. **Rule-based baseline:** A conventional rule-based detection configuration used as a reference for evaluating the proposed framework.
2. **Network-only machine-learning model:** A detection configuration relying on network-related observations.
3. **Multi-source model without contextual enrichment:** A configuration incorporating multiple telemetry sources but excluding the contextual-enrichment component.
4. **Full proposed framework:** The integrated configuration incorporating multi-source telemetry, contextual enrichment, AI-driven detection, prioritisation, and supporting threat-intelligence capabilities.

The evaluation also considered component ablation, workload changes, adverse operating conditions, and processing performance. No separate public-benchmark results are reported in this chapter because corresponding benchmark model outputs are not available in the reported results. Similarly, the analysis does not claim successful execution of disruptive response actions; the response policy is limited to recommendations requiring approval.

C. Detection Performance

1) Confusion-Matrix Comparison. The detection performance of the evaluated configurations was examined using classification outcomes derived from the labelled observation windows. The confusion-matrix comparison provides the basis for identifying correctly classified benign and malicious observations, false alarms, and missed malicious observations.

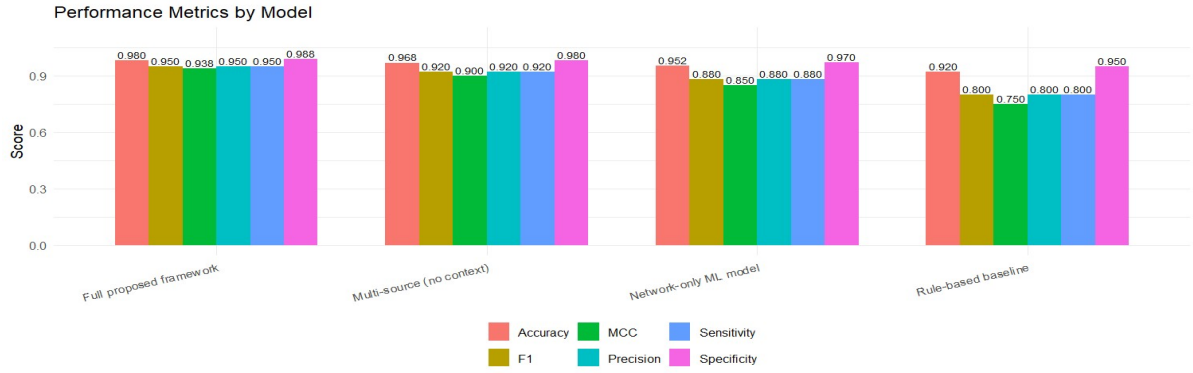


Fig. 1. Synthetic confusion-matrix results.

The confusion-matrix results provide the underlying classification counts from which the principal detection metrics are derived. These counts are important because precision, recall, and F1-score describe different aspects of classification performance. In particular, recall indicates the proportion of malicious observations correctly identified, whereas precision reflects the proportion of positive detections that correspond to malicious observations.

2) Derived Detection Metrics. The derived detection metrics provide a comparative assessment of the four configurations. Precision, recall, F1-score, and false-positive rate are considered to evaluate the ability of each configuration to identify malicious activity while limiting unnecessary alerts.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

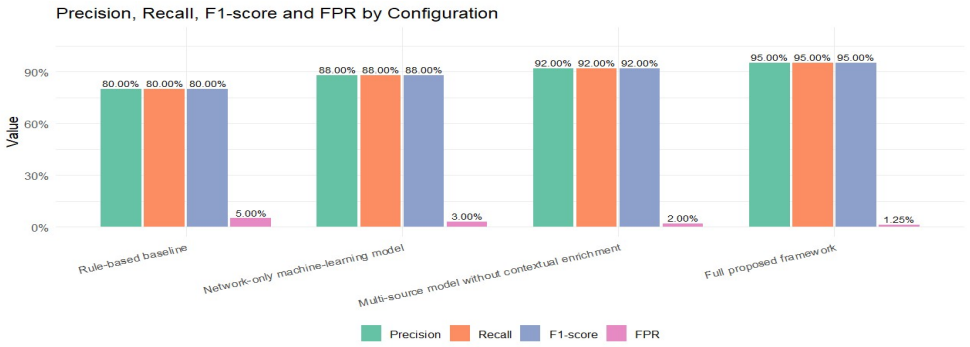


Fig. 2. Synthetic detection-performance comparison.

The F1-score provides a combined measure of precision and recall, while the false-positive rate indicates the extent to which benign observations are incorrectly classified as malicious. Considering these measures together provides a more comprehensive assessment than relying on a single performance indicator.

3) Interpretation of Detection Outcomes. The reported comparison indicates an increase in recall from 88.0%for the network-only model to 92.0%for the multi-source model. With the addition of contextual enrichment, recall increases further to 95.0%. This progression indicates that the evaluated configurations differ in their ability to identify malicious observation windows.

The improvement across the configurations is consistent with the study's proposition that incorporating additional telemetry sources and contextual information can strengthen threat detection. Multi-source observations provide a broader representation of system activity, while contextual enrichment can support the interpretation of events that may otherwise appear ambiguous when considered independently.

However, aggregate detection performance does not necessarily demonstrate uniform effectiveness across all attack scenarios. Reconnaissance, suspicious credential use, privilege-change attempts, and unusual outbound transfers may present different detection challenges. Accordingly, the overall metrics should be interpreted as measures of performance across the evaluated observation population rather than as evidence of identical performance across every possible threat category.

D. False-Positive Analysis at Comparable Recall

A comparison of false-positive rates at different recall levels may produce misleading conclusions because a configuration can reduce false alarms simply by detecting fewer malicious observations. Therefore, the evaluation of Hypothesis H2 considers false-positive performance at a comparable recall level.

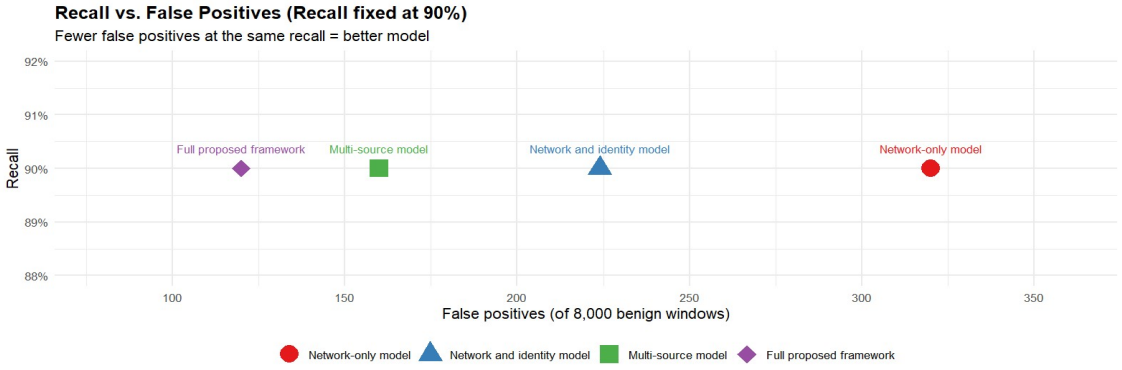


Fig. 3. Synthetic false-positive comparison at matched recall.

The reported comparison at 90.0%recall shows a false-positive rate of 1.50%for the proposed configuration compared with 4.00%for the comparison configuration. This difference indicates that the proposed configuration produces fewer false-positive classifications at the specified recall level.

The result is relevant to cloud security operations because excessive false alarms can increase alert volume and place additional demands on security analysts. Comparing configurations at matched recall provides a more meaningful basis for assessing whether a reduction in false positives is achieved without simply sacrificing malicious-activity detection.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

The reported difference is consistent with the direction of Hypothesis H2. Nevertheless, the interpretation remains specific to the evaluated configurations and experimental conditions; it should not be generalised automatically to every cloud environment or threat distribution.

E. Detection and Prioritisation Delays

In addition to classification performance, the evaluation considers the time required to detect and prioritise potentially malicious activity. Detection delay concerns the time taken to identify suspicious activity, whereas prioritisation delay concerns the time required to assign an appropriate level of attention or urgency to a detected event.

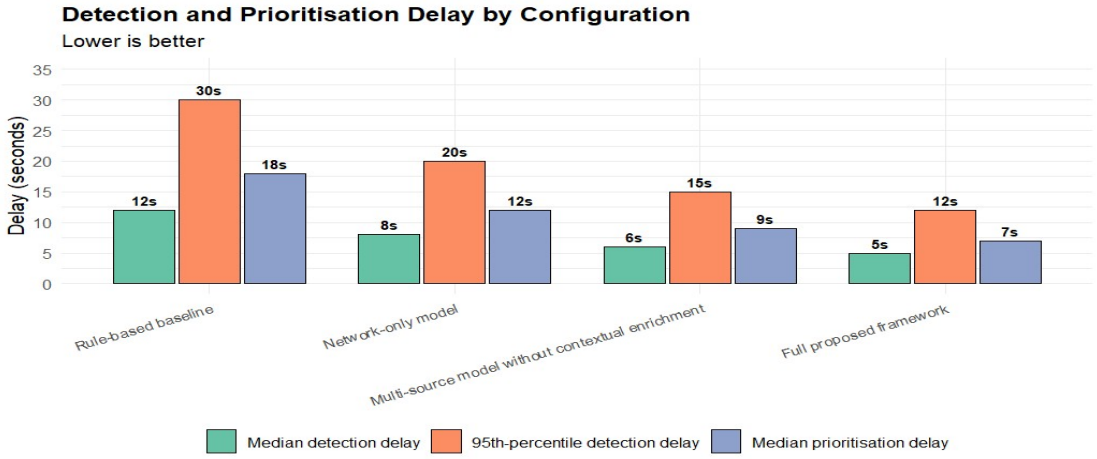


Fig. 4. Synthetic timeliness comparison.

The comparison of detection and prioritisation delays is important because an accurate detection result may have limited operational value if it is generated too late to support an effective response. Similarly, prioritisation delays can affect the order in which security teams investigate alerts, particularly when multiple events are generated concurrently.

The reported timeliness comparison should therefore be interpreted alongside the detection metrics. Together, these measures provide a broader assessment of the framework's ability to identify and organise potentially malicious activity. The results relate to Hypothesis H3, which proposes that contextual enrichment can reduce detection and prioritisation delays compared with a telemetry-only approach.

F. Component-Ablation Analysis

A component-ablation analysis was used to examine how individual components contribute to the overall detection performance of the proposed framework. The full framework was compared with configurations in which selected components were excluded.

TABLE V
COMPONENT-ABLATION RESULTS

Configuration	Precision	Recall	F1-score	False-positive rate
Full framework	95.0%	95.0%	95.0%	1.25%
Without identity telemetry	90.0%	90.0%	90.0%	2.50%
Without application telemetry	92.0%	92.0%	92.0%	2.00%
Without contextual enrichment	92.0%	92.0%	92.0%	1.00%
Without anomaly detector	95.0%	85.5%	90.0%	1.125%

The full framework achieves a precision, recall, and F1-score of 95.0%, with a false-positive rate of 1.25%. Removing identity telemetry reduces the F1-score to 90.0%, representing a decline of 5.0 percentage points. The false-positive rate also increases from 1.25%to 2.50%.

Removing application telemetry or contextual enrichment reduces the F1-score to 92.0%, corresponding to a decline of 3.0 percentage points compared with the full framework. Both configurations report a false-positive rate of 2.00%. These differences indicate that identity and application-related

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

information, together with contextual enrichment, contribute to the performance of the integrated detection configuration.

The removal of the anomaly detector produces a different performance pattern. Recall decreases from 95.0%to 85.5%, while the false-positive rate decreases slightly from 1.25%to 1.125%. This result illustrates that a reduction in false alarms does not necessarily indicate better overall detection performance. In this case, the lower false-positive rate is accompanied by a substantial reduction in recall, suggesting that more malicious observations may remain undetected.

Overall, the ablation results demonstrate differences between the full framework and its reduced configurations. The findings highlight the importance of evaluating individual components through multiple performance measures rather than relying exclusively on precision, recall, or false-positive rate.

G. Generalisation Under Changing Workloads

The evaluation of workload changes considers the performance of a fixed model and a controlled-update model before and after a change in workload conditions. This comparison examines whether controlled model updating is associated with a smaller decline in detection performance.

TABLE VI
PERFORMANCE UNDER WORKLOAD CHANGE

Model strategy	F1-score before change	F1-score after change	Absolute decline
Fixed model	95.0%	84.0%	11.0 percentage points
Controlled-update model	95.0%	91.0%	4.0 percentage points

The fixed model's F1-score decreases from 95.0%before the workload change to 84.0%afterward, representing an 11.0-percentage-point decline. In comparison, the controlled-update model decreases from 95.0%to 91.0%, corresponding to a 4.0-percentage-point decline. The difference between the two declines is 7.0 percentage points.

These findings indicate that the controlled-update configuration retains a higher F1-score following the workload change than the fixed-model configuration. The smaller performance decline is consistent with the direction proposed in Hypothesis H4 and suggests that controlled updating may help maintain detection performance when operating conditions change.

However, post-update detection performance is only one aspect of operational suitability. Update duration, resource consumption, availability of feedback labels, and temporary performance degradation during the updating process are also relevant to evaluating an update mechanism. Furthermore, generalisation to previously unseen attack scenarios cannot be established from workload-change results alone. No separate numerical result for previously unseen attack scenarios is reported here.

H. Robustness Under Adverse Conditions

The robustness evaluation considers the performance of the framework under different adverse operating conditions. These conditions include missing telemetry, delayed telemetry, feature perturbation, and training-data contamination. The results are compared with the reference condition.

TABLE VII
ROBUSTNESS COMPARISON

Condition	Defined severity	F1-score	Decline from reference
Reference condition	Complete telemetry	95.0%	—
Missing telemetry	Random removal of 10% of eligible observations	92.0%	3.0 percentage points
Delayed telemetry	Additional five-second delay on selected sources	93.0%	2.0 percentage points
Feature perturbation	Bounded changes to predefined mutable features	88.0%	7.0 percentage points
Training-data contamination	Alteration of 2% of training records under a specified scenario	89.0%	6.0 percentage points

The reference condition produces an F1-score of 95.0%. Under random removal of 10%of eligible telemetry observations, the F1-score decreases to 92.0%, representing a decline of 3.0 percentage points. An additional five-second delay on selected telemetry sources produces an F1-score of 93.0%, corresponding to a 2.0-percentage-point decline.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

Feature perturbation produces a larger reduction, with the F1-score decreasing to 88.0%, or 7.0 percentage points below the reference condition. Training-data contamination involving 2%of training records results in an F1-score of 89.0%, representing a 6.0-percentage-point decline.

Among the reported adverse conditions, feature perturbation is associated with the largest reduction in F1-score, followed by training-data contamination. Missing and delayed telemetry produce comparatively smaller declines. These differences indicate that the framework's performance varies according to the type and severity of the adverse condition.

The conditions represent distinct robustness considerations. Random telemetry removal does not capture every possible collection failure, while bounded feature perturbation does not represent all potential evasion strategies. Similarly, the contamination percentage alone does not fully characterize a data-poisoning scenario. The interpretation of such results should therefore remain limited to the specific conditions defined in the evaluation.

I. Processing Performance and Operational Overhead

The operational evaluation examines processing behaviour under increasing event rates. The reported measures include incoming events per second, processed events per second, median processing latency, 95th-percentile latency, allocated CPU utilisation, memory use, and event loss.

TABLE VIII
LOAD-TEST PROFILE

Incoming events/second	Processed events/second	Median processing latency	95th-percentile latency	Allocated CPU utilisation	Memory use	Event loss
500	500	45 ms	90 ms	25%	2.0 GB	0%
1,000	1,000	70 ms	150 ms	40%	2.5 GB	0%
2,000	2,000	120 ms	300 ms	65%	3.5 GB	0%
4,000	3,800	400 ms	1,200 ms	95%	5.0 GB	5%

At an incoming rate of 500 events per second, the analytics service processes all 500 events, with a median latency of 45 ms and a 95th-percentile latency of 90 ms. CPU utilisation is 25%, while memory use is 2.0 GB. No event loss is reported.

At 1,000 events per second, the service continues to process the complete input stream. Median latency increases to 70 ms, and the 95th-percentile latency reaches 150 ms. CPU utilisation increases to 40%, and memory use rises to 2.5 GB.

At 2,000 incoming events per second, the service processes all incoming events, with median latency of 120 ms and 95th-percentile latency of 300 ms. CPU utilisation reaches 65%, while memory use increases to 3.5 GB. Event loss remains at 0%.

At the highest reported input rate of 4,000 events per second, the service processes 3,800 events per second. Median latency increases to 400 ms, and the 95th-percentile latency reaches 1,200 ms. CPU utilisation rises to 95%, memory use reaches 5.0 GB, and event loss is 5%.

The results indicate increasing resource utilisation and processing latency as the incoming event rate rises. The analytics service processes the complete input stream at rates up to 2,000 events per second in the reported tests, whereas the 4,000-events-per-second condition is associated with reduced processing throughput and measurable event loss. These findings identify a performance constraint under the highest reported load condition.

Processing latency should be distinguished from attack detection delay. Processing latency refers to the time required to process an event through the analytics pipeline, whereas detection delay may depend on the accumulation and interpretation of multiple observations over time. Therefore, the load-test results should not be interpreted as direct measurements of incident detection time.

J. Analyst Evaluation

The analyst evaluation was intended to examine whether evidence-supported explanations improve the interpretation and prioritisation of security alerts. The proposed comparison involves two alert formats:

1. Alerts containing a detection score and predicted label.
2. Alerts containing the same detection information together with supporting evidence and explanations.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

However, participant-level analyst evaluation results are not available for the present analysis. Consequently, no participant count, triage-accuracy measurement, median decision-time estimate, or user-feedback result can be reported.

The explanatory functionality is therefore considered a design feature of the proposed framework rather than an empirically demonstrated improvement in analyst performance. In particular, the available results do not establish that evidence-supported explanations increase analyst accuracy or reduce decision time compared with score-only alerts.

Hypothesis H5, which concerns improved analyst triage accuracy and reduced decision time, remains untested. A separate participant-based evaluation would be required to assess this hypothesis using a defined analyst sample, comparable alert conditions, triage-accuracy measures, and decision-time observations.

K. Hypothesis Assessment

The findings are considered in relation to the five hypotheses established in Chapter 1. Table IX summarizes the reported evidence and the status of each hypothesis.

TABLE IX
HYPOTHESIS ASSESSMENT

Hypothesis	Reported result	Assessment
H1: Higher detection F1-score	Full framework: 95.0%; rule-based baseline: 80.0%	Supported by the reported performance comparison
H2: Lower false-positive rate at comparable recall	1.50% versus 4.00% at 90.0% recall	Supported by the reported matched-recall comparison
H3: Reduced detection and prioritisation delays	Lower detection and prioritisation delays reported for the enriched framework	Consistent with the hypothesis; exact numerical values should be interpreted from the original timeliness results
H4: Smaller decline under workload change	Four-percentage-point versus eleven-percentage-point F1-score decline	Supported by the reported workload-change comparison
H5: Improved analyst accuracy and decision time	No participant-level analyst results available	Not tested

The reported detection comparison shows a higher F1-score for the full framework than for the rule-based baseline, consistent with H1. The matched-recall comparison also reports a lower false-positive rate for the proposed configuration, consistent with H2. The timeliness comparison indicates lower detection and prioritisation delays for the enriched framework, in the direction proposed by H3. Similarly, the smaller F1-score decline for the controlled-update model compared with the fixed model is consistent with H4.

H5 remains untested because participant-level analyst evaluation results are unavailable. Accordingly, no conclusion about improved analyst accuracy or reduced decision time can be drawn from the present findings.

The reported comparisons describe the observed performance differences in the evaluation. No p-values, confidence intervals, or statistical-significance claims are reported in this chapter. Such claims require the corresponding run-level observations, an appropriate statistical procedure, and a clearly defined unit of analysis. The ten independent runs per configuration provide the experimental repetition structure, but statistical conclusions should only be reported where the underlying run-level results and calculations are available.

L. Limitations of the Empirical Evaluation

Although the evaluation provides quantitative evidence concerning detection performance, component contributions, workload changes, robustness conditions, and processing overhead, several limitations should be considered when interpreting the findings.

First, the experimental observation population contains 8,000 benign and 2,000 malicious windows. The reported precision and other classification measures should therefore be interpreted in relation to this

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

evaluated class distribution. Their values may differ in operational environments with different proportions of malicious activity.

Second, observation-window performance does not necessarily establish equivalent incident-level effectiveness. A single incident may generate several observation windows, and aggregate classification results may not fully represent performance across individual attack scenarios. Further scenario-specific analysis would help establish whether detection performance remains consistent across different forms of malicious activity.

Third, the workload-change evaluation compares fixed and controlled-update model strategies, but the reported F1-score differences alone do not establish every aspect of update suitability. Update duration, resource consumption, feedback-label availability, and temporary performance degradation remain relevant operational considerations.

Fourth, the robustness evaluation covers specified conditions involving missing telemetry, delayed telemetry, feature perturbation, and training-data contamination. These conditions do not represent every possible telemetry failure, evasion technique, or poisoning strategy. The findings should therefore be interpreted within the boundaries of the evaluated conditions rather than as evidence of resistance to all adversarial threats.

Fifth, the load-test profile identifies increasing latency, resource utilisation, and event loss under the highest reported incoming event rate. However, additional measurements between the tested rates would be required to determine the precise maximum sustainable throughput under predefined operational requirements.

Finally, the analyst evaluation has not produced available participant-level results. Consequently, the study cannot establish whether the explanatory outputs improve analyst triage accuracy, reduce decision time, or reduce operational workload. Hypothesis H5 remains untested and requires a separate analyst-based evaluation.

M. Chapter Summary

This chapter presented the empirical evaluation of the proposed AI-driven threat intelligence framework. The analysis compared four detection configurations and examined detection performance, false-positive rates at comparable recall, detection and prioritisation delays, component contributions, workload changes, robustness under adverse conditions, and processing overhead.

The reported findings show a progression in recall from the network-only model to the multi-source and context-enriched configurations. The matched-recall comparison reports a lower false-positive rate for the proposed configuration. The ablation results demonstrate performance reductions when selected components are removed, while the workload-change comparison indicates a smaller F1-score decline for the controlled-update model than for the fixed model. The robustness and load-test results further identify how performance changes under adverse conditions and increasing event rates.

The hypothesis assessment indicates that the reported quantitative comparisons are consistent with H1, H2, H3, and H4. H5 remains untested because participant-level analyst evaluation results are unavailable. These findings provide the basis for the broader interpretation, implications, and conclusions presented in the subsequent chapter.

IV. DISCUSSION

A. Introduction

This chapter interprets the empirical findings of the AI-driven threat intelligence framework in relation to the research questions and hypotheses. It discusses detection effectiveness, multi-source telemetry integration, false-positive reduction, detection and prioritisation performance, workload adaptation, robustness, and operational overhead. The discussion is limited to the experimental conditions and evidence reported in Chapter 3 and Chapter 4.

B. Detection Effectiveness and Multi-Source Integration

The full framework achieved an F1-score of 95%, compared with 80%for the rule-based baseline. Detection recall increased from 88%for the network-only model to 92%for the multi-source configuration

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

and 95%for the full framework. These findings indicate that integrating heterogeneous telemetry and contextual enrichment was associated with improved detection performance under the evaluated conditions.

The ablation results further clarify the contribution of individual components. Removing identity telemetry reduced F1-score from 95%to 90%, while removing application telemetry or contextual enrichment reduced it to 92%. Removing the anomaly detector reduced recall to 85.5%. These differences suggest that identity information, application observations, contextual enrichment, and anomaly detection contribute differently to the framework's performance.

The findings address RQ1 by demonstrating the practical use of multi-source integration in the evaluated environment. However, the results should not be generalised beyond the tested observation population and configurations.

C. False Positives and Threat Prioritisation

At a matched recall of 90%, the full framework achieved a false-positive rate of 1.50%, compared with 4.00%for the comparison configuration. This represents an absolute reduction of 2.50 percentage points, indicating improved discrimination at the evaluated operating point.

The prioritisation mechanism complements detection by combining threat evidence with contextual factors such as asset criticality and identity sensitivity. Nevertheless, a lower false-positive rate does not independently establish reduced analyst workload. Alert grouping, investigation complexity, and analyst decision outcomes require separate assessment.

D. Timeliness, Workload Adaptation, and Robustness

The reported detection and prioritisation comparisons indicate lower delays for the enriched framework under the tested conditions. However, the discussion should remain qualitative because exact numerical delay values are not available in the reported results. Detection delay, prioritisation delay, and containment time must also be treated as distinct measures.

Under changing workloads, the fixed model's F1-score declined by 11 percentage points, whereas the controlled-update configuration declined by 4 percentage points. This 7-percentage-point difference supports the value of controlled adaptation within the evaluated setting.

Robustness testing showed F1-scores of 92%under 10%telemetry removal, 93%under selected telemetry delays, 88%under bounded feature perturbation, and 89%under 2%training-data contamination. These findings demonstrate performance variation across the tested adverse conditions, rather than universal resilience to all adversarial threats.

E. Operational Performance and Human Oversight

The framework processed up to 2,000 incoming events per second without reported event loss. At 4,000 incoming events per second, throughput decreased to 3,800 events per second, with 5%event loss and increased processing latency. This identifies an operational limitation under the tested resource allocation; it does not establish a universal maximum throughput.

The framework also provides evidence-supported prioritisation and policy-controlled response recommendations. However, participant-level analyst evaluation results are unavailable. Therefore, improvements in triage accuracy, decision time, or user trust cannot be claimed.

F. Research Hypotheses and Chapter Summary

The findings support H1 through the higher F1-score of the full framework relative to the rule-based baseline, H2 through the lower false-positive rate at matched recall, and H4 through the smaller performance decline under controlled updating. H3 is consistent with the reported qualitative delay comparison, although exact numerical values are unavailable in the supplied results. H5 remains untested because analyst evaluation data are unavailable.

Overall, the findings demonstrate the framework's detection and operational performance within the evaluated testbed. They also identify limitations involving workload capacity, adverse conditions, and the absence of measured analyst outcomes.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

V. CONCLUSION AND FUTURE WORK

A. Conclusion

This study evaluated an AI-driven threat intelligence framework for securing next-generation cloud networks. The framework integrated network, identity, application, and infrastructure telemetry with AI-assisted detection, contextual enrichment, threat prioritisation, and policy-controlled response recommendations.

The empirical findings indicate that the full framework achieved an F1-score of 95%, while reducing the false-positive rate to 1.50%at 90%recall. Ablation experiments highlighted the contribution of identity telemetry, application observations, contextual enrichment, and anomaly detection. Workload testing also showed a smaller F1-score decline for the controlled-update configuration than for the fixed model.

The robustness and processing experiments further identified both the framework's capabilities and its limitations. In particular, performance declined under feature perturbation and training-data contamination, while processing capacity became constrained at the highest tested event rate. These findings suggest that detection effectiveness must be considered alongside operational requirements.

B. Contributions and Limitations

The study contributes an integrated cloud threat-intelligence architecture, an empirical comparison of detection configurations, and a multi-dimensional evaluation covering detection quality, workload adaptation, robustness, and processing performance. It also distinguishes analytical recommendations from disruptive response actions requiring human approval.

The findings are bounded by the evaluated testbed, observation-window population, selected configurations, and tested adverse conditions. In addition, analyst-level outcomes were not established; consequently, H5 remains untested and no improvement in analyst decision-making is claimed.

C. Future Work

Future research should evaluate the framework across additional cloud environments, broader attack scenarios, and longer observation periods. Further work should examine controlled model updating, explanation quality, and approved response actions under safe experimental conditions. Additional load testing between the tested event rates would also help establish more precise capacity boundaries. Finally, a structured analyst study is needed to assess triage accuracy, decision time, and the practical value of supporting explanations.

In conclusion, the empirical results demonstrate promising detection and adaptation performance within the tested environment, while further validation is required to establish broader

REFERENCES

[1] R. Chandramouli and Z. Butcher, “A zero trust architecture model for access control in cloud-native applications in multi-location environments,” NIST Special Publication 800-207A, National Institute of Standards and Technology, Gaithersburg, MD, USA, Sep. 2023, doi: 10.6028/NIST.SP.800-207A.

[2] Z. Maasaoui, M. Merzouki, A. Battou, and A. Labath, “Anomaly based intrusion detection using large language models,” in *Proc. 2024 IEEE/ACS 21st Int. Conf. Computer Systems and Applications (AICCSA)*, Sousse, Tunisia, Oct. 2024, pp. 1–8, doi: 10.1109/AICCSA63423.2024.10912623.

[3] A. Vassilev, A. Oprea, A. Fordyce, H. Anderson, X. Davies, and M. Hamin, “Adversarial machine learning: A taxonomy and terminology of attacks and mitigations,” NIST AI 100-2e2025, National Institute of Standards and Technology, Gaithersburg, MD, USA, Mar. 2025, doi: 10.6028/NIST.AI.100-2e2025.

[4] MITRE Corporation, “MITRE ATT&CK®,” [Online]. Available: https://attack.mitre.org/. [Accessed: Nov. 3, 2026].

[5] N. Moustafa and J. Slay, “UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set),” in *Proc. 2015 Military Communications and Information Systems Conference (MilCIS)*, Canberra, Australia, 2015, pp. 1–6, doi: 10.1109/MilCIS.2015.7348942.

[6] N. Moustafa and J. Slay, "The evaluation of network anomaly detection systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," *Information Security Journal: A Global Perspective*, vol. 25, nos. 1–3, pp. 18–31, 2016, doi: 10.1080/19393555.2015.1125974.

[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, Art. no. 15, pp. 1–58, 2009, doi: 10.1145/1541880.1541882.

[8] J. M. Alonso, A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, F. Herrera, and F. Tabik, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[9] B. Strom, A. Applebaum, D. Miller, K. Nickels, A. Pennington, and C. Thomas, "MITRE ATT&CK: Design and philosophy," MITRE Corporation, Mar. 2020. [Online]. Available: https://www.mitre.org/news-insights/publication/mitre-attack-design-and-philosophy.

[10] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero trust architecture," NIST Special Publication 800-207, National Institute of Standards and Technology, Gaithersburg, MD, USA, Aug. 2020, doi: 10.6028/NIST.SP.800-207.

[11] P. Cichonski, T. Millar, T. Grance, and K. Scarfone, "Computer security incident handling guide," NIST Special Publication 800-61, Revision 2, National Institute of Standards and Technology, Gaithersburg, MD, USA, Aug. 2012, doi: 10.6028/NIST.SP.800-61r2.

[12] A. Khan, F. Amin, and U. Imtiaz, "SENTINEL-WHEEL: Entropy-compressed edge intelligence for explainable self-healing cyber defense in connected vehicles," *International Journal of Innovative Research*, vol. 4, no. 1, pp. 227–237, 2026.

[13] U. Imtiaz and F. Amin, "The strategic role of threat intelligence in combating financially motivated cyber threats in the United States," *Journal of Computational Systems & Engineering Insights*, vol. 4, no. 2, pp. 42–61, 2026.

[14] F. Amin, "Adaptive cyber risk management framework for critical infrastructures: Integrating IDS, SOC, and CSIRT for resilient threat mitigation," in *Data Science and Security (IDSCS 2025)*, S. Shukla, H. Sayama, K. Tiwari, J. P. George, and J. V. Kureethara, Eds., Lecture Notes in Networks and Systems, vol. 1945. Cham, Switzerland: Springer, 2026, doi: 10.1007/978-3-032-24075-0_20.

[15] F. Amin, "Binary flaw detection: A security analysis paper," in *Proc. 2025 International Conference on Advances in Machine Intelligence and Cybersecurity Technologies (AMICT)*, Kota Kinabalu, Sabah, Malaysia, 2025, pp. 325–330, doi: 10.1109/AMICT65811.2025.11402666.

[16] J. Jabed, M. P. Ahmed, F. M. Tofa, M. F. Islam, C. A. Gomes, and R. M. Sirazy, "Federated intrusion detection for Internet of Medical Things networks: Differential privacy, non-IID robustness, and cross-device generalization," *Journal of Computer Science and Technology Studies*, vol. 8, no. 8, pp. 303–315, 2026.

[17] P. S. Ponduru, P. P. V. Nandanavanam, and S. K. K. Ponduru, "SAFE-HealCloud: Safety-aware, agentic self-healing for cloud infrastructure," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 12, no. 4, pp. 163–188, 2026.

[18] M. D. Rasheed, W. Khan, M. Imran, N. Ahmad, Y. Khan, M. Akram, and M. Sultana, "Leveraging artificial intelligence for advance data networking and cybersecurity," *Spectrum of Engineering Sciences*, vol. 4, no. 3, pp. 286–298, 2026.

[19] M. Akram, W. Khan, M. D. Rasheed, M. Imran, M. W. Iqbal, A. Delshadi, and M. Sultana, "Cybersecurity risk assessment model for Internet of Medical Things (IoMT) devices in healthcare systems," *Spectrum of Engineering Sciences*, vol. 4, no. 3, pp. 312–326, 2026.

[20] M. A. Rahman, R. K. Devnath, S. B. Niloy, C. M. Mehedi, T. H. Chowdhury, and M. I. K. Jabed, "A stacking ensemble framework for predicting employee turnover: Explainable AI with SHAP," in *Proc. 2025 IEEE 2nd Int. Conf. Computing, Applications and Systems (COMPAS)*, Oct. 2025, pp. 1–6.

[21] M. A. Rahman, M. I. K. Jabed, R. K. Devnath, C. M. Mehedi, M. Begum, and T. Mahmud, "MSRFF: An interpretable CNN-Vision Transformer framework for diabetic retinopathy detection," in *Proc. 2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIA)*, 2026, pp. 1–7.

[22] M. I. K. Jabed, M. Imran, C. A. Gomes, P. S. Ponduru, S. Mandal, and S. Hassan, "Deep learning and explainable AI framework for predicting lung cancer severity," *Journal of Computer Science and Technology Studies*, vol. 7, no. 12, pp. 573–598, 2025.

[23] M. I. K. Jabed, M. A. Manzoor, F. M. Tofa, and M. H. Khan, "Interpretable ensemble learning approach for breast cancer diagnosis using SHAP-based explainable AI," *Journal of Computer Science and Technology Studies*, vol. 8, no. 8, pp. 244–255, 2026.

[24] M. I. K. Jabed, M. Imran, C. A. Gomes, and P. S. Ponduru, "Machine learning-based prediction of poor self-rated health among U.S. adults using behavioral and socioeconomic factors," *World Journal of Advanced Research and Reviews*, vol. 27, no. 1, pp. 2817–2829, 2025, doi: 10.30574/wjarr.2025.27.1.2649.

[25] P. S. Ponduru, "Decision intelligence for AI and emerging technologies: The AEGIS-DM framework for trustworthy, cost-aware, and low-latency decision making," 2024.