



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

MSc Cybersecurity — Research Proposal

Cross-Cloud Forensic Readiness:

Evaluating the Reliability and Portability of AI-Based Evidence
Classification Across Microsoft Azure and Amazon Web Services

Student Name: Md Hasibul Islam

Supervisor: Suresh Shanmugasundaram

Module: Research Methods and Professional Development

Module code: ACFA7004

University of Wales Trinity Saint David (UWTSD)

Submission Date: 14th August 2026

Table of Contents

1. Introduction	3
1.1 Context and Background	3
1.2 Research Problem	4
1.3 Research Aim and Objectives	5
1.4 Significance of the Research	6
2. Literature Review	6
2.1 Deep Learning for Log and Network Anomaly Detection	7
2.2 Cross-Domain Generalisation and Domain Shift	7
2.3 Cloud Forensic Readiness	8
2.4 Theoretical and Technical Frameworks	9
2.5 Relation to the Present Study and Summary	9
3. Research Design and Methodology	10
3.1 Research Design	10
3.2 Research Methods and Materials	10
4. Expected Outcome	12
5. Timeline	13
References	14

1. Introduction

1.1 Context and Background

Digital forensics, once concerned primarily with the seizure and imaging of physical storage devices, has been reshaped by the migration of computing to the cloud. Cybersecurity as a discipline now sits at the intersection of three rapidly evolving fields — artificial intelligence, cloud computing, and digital forensic science — and it is precisely this intersection that frames the present research. Contemporary enterprises no longer operate within a single, self-contained perimeter; instead, an estimated 87% run workloads across two or more cloud providers simultaneously, a configuration commonly described as multi-cloud (Basharat, 2025). This architectural shift delivers scalability and resilience, but it fragments the evidential record: Microsoft Azure records network activity through Network Security Group and Virtual Network Flow Logs, whereas Amazon Web Services records equivalent activity through Virtual Private Cloud Flow Logs, and the two use incompatible schemas, field names, and value encodings.

In parallel, the methods used to analyse security telemetry have advanced considerably. The application of deep learning — and, more recently, Transformer-based language models — to system and network logs has produced anomaly detectors that achieve near-perfect accuracy on established benchmarks (Guo, Yuan and Wu, 2021; Chen and Liao, 2022). The latest developments extend this trajectory in two directions: evidential deep learning, which augments classification with a calibrated measure of predictive uncertainty (Guo et al., 2024), and the application of large language models directly to cloud audit trails for threat detection and attack-technique mapping. At the same time, the forensic community has moved away from a reactive posture towards forensic readiness — the proactive structuring, preservation, and legal safeguarding of digital evidence before an incident occurs — so that evidence is investigation-ready and admissible when required (Idugboe et al., 2026). The convergence of these developments creates both an opportunity and a problem: powerful AI classifiers can now triage cloud evidence at scale, yet their reliability when evidence originates from heterogeneous providers has not been established.

The scale of this challenge is not incidental but structural. Cloud logging is designed by each provider around its own service model, so semantically identical events, such as a denied inbound connection, a successful authentication, or a port probe, are

represented with different field names, value vocabularies, and levels of granularity across Azure and AWS. When an incident spans both, an investigator must reconcile these representations by hand, a process that is slow, error-prone, and difficult to defend under the evidential standards a court expects. Automating this reconciliation with machine learning is therefore attractive, and a growing literature applies classifiers to cloud and network telemetry for exactly this purpose. However, the discipline has historically prioritised a single figure of merit, namely detection accuracy, and has paid comparatively little attention to whether the resulting judgements are reliable enough to serve as evidence, or whether they remain reliable once the underlying provider changes. It is this mismatch between what the literature optimises and what forensic practice requires that the present research seeks to close.

1.2 Research Problem

The central problem this research addresses is that AI-based classifiers used to categorise and triage digital evidence in multi-cloud environments may report high accuracy on aggregated data while failing to generalise to individual providers, producing evidence whose reliability is overstated and whose admissibility is therefore compromised.

Several challenges in the current body of work make this problem significant. First, the phenomenon of domain shift is well documented: intrusion detection models that achieve Macro-F1 scores between 0.84 and 0.96 within a single dataset degrade to between 0.69 and 0.78 when transferred to a different environment, with benign false-positive rates rising to as much as 0.30 (Future Internet, 2026). Second, recent explainability research has traced this degradation to shortcut learning, demonstrating that models rely disproportionately on coarse features such as port categories and that reducing feature granularity merely relocates the shortcut rather than eliminating it (Hakim et al., 2026). Third, existing cross-domain evaluations are methodologically confounded: they simultaneously vary the attack types, the feature representations, and the provider, so the specific contribution of the provider to performance loss cannot be isolated (TAN-IDS, 2026). Fourth, and most critically for forensic practice, no existing study examines whether the chain-of-custody metadata that accompanies cloud evidence — collection method, handler identity, retention policy, and similar fields — itself encodes provider-specific information that a classifier may exploit, thereby contaminating the evidential judgement. These gaps mean that a classifier deployed on a previously unseen provider could fail silently, undermining the integrity of any investigation that relies upon it.

From this problem, three research questions follow. First, can a unified classifier accurately categorise attack evidence when trained on data drawn from both Azure and AWS together? Second, does that classifier remain reliable when it is trained on one provider and applied to the other, that is, is the evidential judgement portable across providers? Third, does the chain-of-custody metadata itself carry provider-specific signal that the classifier exploits, and if so, does its inclusion inflate apparent reliability while undermining genuine forensic soundness? These questions move the emphasis away from raw accuracy and towards the properties, namely portability, calibration, and freedom from provenance contamination, that determine whether AI-classified evidence can be trusted in a multi-cloud investigation.

1.3 Research Aim and Objectives

The aim of this research is to evaluate the reliability, portability, and forensic soundness of AI-based evidence classification across Microsoft Azure and Amazon Web Services, and to determine whether chain-of-custody metadata compromises cross-provider evidentiary admissibility.

To achieve this aim, the research pursues six measurable objectives:

1. To construct a provider-matched multi-cloud dataset in which identical attack classes are represented on both Azure and AWS, normalised into a unified twenty-eight-field forensic schema, thereby enabling the provider effect to be isolated from other sources of variation.
2. To design and apply a hybrid feature-selection procedure using XGBoost and LightGBM with SHAP-based importance ranking, incorporating explicit safeguards against label leakage and cross-provider contamination.
3. To implement and train four classifiers — LogEDL, BERT-Log, BiCNN with Attention, and BiLSTM with Attention — selected to represent complementary architectural paradigms.
4. To evaluate in-provider reliability using not only weighted F1 and Macro-AUC but also per-class reliability, confidence calibration, and evidential uncertainty, reflecting the requirements of court-defensible evidence.
5. To quantify cross-provider portability by training on one provider and testing on the other, and to identify which specific evidence classes lose reliability under transfer.

6. To determine, through a controlled ablation, whether chain-of-custody metadata leaks provider identity into the classifier and thereby compromises forensic soundness.

1.4 Significance of the Research

This research is significant on both scientific and practical grounds. Scientifically, it offers the first isolation of the pure provider effect in cross-cloud intrusion detection by holding attack types and feature representations constant across providers — a methodological control that prior confounded studies could not achieve (Apruzzese, Pajola and Conti, 2022). It also introduces the previously unexamined question of metadata-induced contamination, contributing a new dimension to the evaluation of AI evidence classifiers. Practically, the outcomes will provide digital forensic investigators and security operations teams with evidence-based guidance on whether, and under what confidence thresholds, an AI classifier trained on one cloud provider can be trusted on another, directly informing admissibility decisions. More broadly, by framing provider-transfer testing as an evaluation standard, the research encourages the forensic and machine-learning communities to treat evidentiary reliability, rather than aggregate accuracy, as the primary criterion for deploying AI in multi-cloud investigations.

The wider relevance of the work follows from the trajectory of the field. As enterprises continue to distribute workloads across providers and as regulators increasingly expect demonstrable forensic readiness, the volume of evidence that must be triaged across heterogeneous clouds will only grow, and automation will become unavoidable. If AI classifiers are to be adopted responsibly in this setting, their failure modes must be understood before deployment rather than discovered during litigation. By characterising precisely when and why cross-provider classification fails, and by exposing a contamination risk that has not previously been articulated, this research contributes both a cautionary finding and a constructive evaluation methodology that practitioners and future researchers can apply to any pair of providers.

2. Literature Review

The purpose of this literature review is to examine the existing body of research at the intersection of machine-learning-based anomaly detection, cross-domain generalisation, and cloud forensic readiness, in order to identify the key contributions, theoretical foundations, and gaps that motivate the present study. The scope encompasses deep-learning and Transformer-based approaches to log and network-flow analysis, evaluations of model transferability under distributional change, and

frameworks for proactive cloud forensics, drawing primarily on literature published between 2019 and 2026. The review is organised thematically into three strands, followed by a synthesis of the theoretical and technical frameworks and a statement of how the present study relates to, and advances, this body of work.

2.1 Deep Learning for Log and Network Anomaly Detection

The first strand concerns the application of neural architectures to security telemetry. Early work established the use of recurrent networks to model normal log sequences, and the paradigm was substantially advanced by LogBERT, which framed anomaly detection as a self-supervised masked-token task over log sequences using a bidirectional Transformer encoder (Guo, Yuan and Wu, 2021). Subsequent work refined this approach: BERT-Log treated the log sequence as natural language and fine-tuned a pre-trained model for classification, reporting F1 scores above 0.99 on the HDFS and BGL benchmarks (Chen and Liao, 2022), while contrastive and dual-objective variants improved robustness and reduced dependence on log parsing. Most relevant to the present study, LogEDL augmented masked-language modelling with evidential deep learning, so that the model expresses calibrated uncertainty and can flag unfamiliar patterns through high uncertainty rather than an overconfident prediction (Guo et al., 2024). These contributions have collectively raised detection accuracy to near-ceiling levels; however, they were developed and validated almost exclusively for single-platform, binary anomaly detection on curated benchmark datasets, and none addresses multi-class attack classification across heterogeneous cloud providers.

A recurring consideration within this strand is the trade-off between semantic richness and practical deployability. Approaches that avoid explicit log parsing, learning directly from raw message content, tend to generalise better to unseen log formats but demand substantial labelled data and computation; approaches that rely on parsed templates are lighter but brittle when log structure changes. The evidential extension embodied by LogEDL is notable precisely because it addresses a different axis, namely not what the model predicts but how confident it is entitled to be, which is the property most directly relevant to whether a prediction can be presented as evidence. This observation motivates the inclusion of an uncertainty-aware model as a first-class comparator in the present study rather than as an afterthought.

2.2 Cross-Domain Generalisation and Domain Shift

The second strand interrogates whether the high accuracy reported above survives deployment in a different environment. A consistent and sobering finding has

emerged: models that perform almost perfectly within a dataset degrade sharply across datasets. Apruzzese, Pajola and Conti (2022) formalised cross-evaluation through the XeNIDS framework, showing that single-dataset evaluation conceals substantial risk while cross-evaluation can also extend a model's effective detection surface. Layeghy and Portmann (2023) combined explainability with cross-domain evaluation in DI-NIDS, projecting traffic into a domain-invariant space using adversarial training. More recent work has quantified the problem with precision: a 2026 study across four benchmarks reported Macro-F1 falling from the 0.84–0.96 range to 0.69–0.78 under transfer (Future Internet, 2026), and the TAN-IDS framework argued that observed degradation is frequently attributable not only to genuine domain shift but also to uncontrolled feature incompatibilities (TAN-IDS, 2026). Critically, Hakim et al. (2026) used explainability analysis and targeted ablation to show that cross-domain failure stems from reliance on coarse port-category shortcuts, and that coarsening feature resolution relocates rather than removes the shortcut. Heterogeneous ensemble and domain-adaptation approaches have been proposed to narrow the gap (Connection Science, 2025), yet all of this work is conducted on unrelated public datasets that differ in attacks, features, and source simultaneously, leaving the isolated effect of the cloud provider unmeasured.

2.3 Cloud Forensic Readiness

The third strand addresses the forensic dimension. Foundational work established forensic readiness as a proactive organisational capability (Alenezi, Atlam and Wills, 2019) and formalised cloud forensic readiness as a service through finite-state modelling (Kebande and Choo, 2022). System-level contributions such as the Cloud Evidence Tracing System demonstrated integrated forensic investigation across large-scale public cloud platforms (Wu et al., 2022). Most recently, Idugboe et al. (2026) proposed a machine-learning-driven intrusion detection framework with forensic readiness for cloud-native environments, evaluating Logistic Regression, Decision Tree, Random Forest, XGBoost, and Autoencoder models and reporting XGBoost as the strongest performer. These frameworks rightly emphasise chain-of-custody, evidence preservation, and admissibility; however, they evaluate detection on single platforms and by conventional accuracy metrics, and none tests whether an AI classifier's judgements remain reliable — or its custody metadata uncontaminated — when evidence crosses providers.

A further observation across the forensic strand is that admissibility depends on more than correctness. For evidence to be relied upon, its provenance must be documented,

its handling auditable, and the confidence attached to any automated judgement defensible. The chain-of-custody fields emphasised by these frameworks exist precisely to satisfy the first two requirements; yet the same fields, if consumed as input features by a classifier, could allow the model to infer the provider from how the evidence was collected rather than from what the evidence shows. No reviewed study considers this possibility, which sits at the boundary between the forensic and machine-learning literatures and which the present research is designed to investigate directly.

2.4 Theoretical and Technical Frameworks

The research reviewed rests on several foundations that also underpin this study. The Transformer architecture and its self-attention mechanism (Vaswani et al., 2017) form the basis of the language-model detectors; evidential deep learning provides principled uncertainty quantification (Sensoy, Kaplan and Kandemir, 2018); and domain-adversarial training offers a route to distributional alignment (Ganin et al., 2016). For feature selection, gradient-boosted decision trees — specifically XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) — are widely used, and SHAP provides a theoretically grounded, model-agnostic method of feature attribution (Lundberg and Lee, 2017) that is particularly suited to the explainability demands of forensic evidence. Design Science Research (Hevner et al., 2004; Peffers et al., 2007) supplies the overarching methodological frame for building and evaluating the artefact.

2.5 Relation to the Present Study and Summary

This study is positioned at the confluence of the three strands and advances each. Against the anomaly-detection literature, it repurposes state-of-the-art architectures — including the uncertainty-aware LogEDL — for multi-class evidence classification rather than single-platform binary detection. Against the cross-domain literature, it removes the central confound by holding attacks and features constant across a provider-matched dataset, thereby isolating the provider effect for the first time and responding directly to the shortcut-learning findings of Hakim et al. (2026). Against the forensic-readiness literature, it reframes evaluation around evidentiary reliability, calibration, and admissibility, and it introduces the novel investigation of chain-of-custody metadata contamination. In summary, while existing research has independently achieved accurate detection, documented domain shift, and articulated forensic-readiness principles, no study has connected them to ask whether AI evidence classification is reliable and portable across cloud providers — the gap this proposal addresses.

3. Research Design and Methodology

3.1 Research Design

This research adopts a Design Science Research (DSR) approach, as formalised by Hevner et al. (2004) and operationalised by Peffers et al. (2007). DSR is appropriate because the study does not merely observe an existing phenomenon but designs, constructs, and evaluates a novel artefact — an evaluation framework for cross-provider forensic evidence classification — to solve a demonstrated real-world problem. The DSR cycle maps directly onto the work: the problem is identified from the literature and forensic practice; the artefact, comprising the unified schema, the feature-selection procedure, the model suite, and the reliability-evaluation protocol, is designed and built; and the artefact is then rigorously evaluated against defined criteria of reliability and cross-provider portability. This design-and-evaluate orientation suits a problem whose answer is a validated method and a set of empirical findings rather than a purely theoretical proof.

An alternative would have been a purely exploratory or observational design, but such an approach would neither produce the reusable, evaluable artefact that the research problem demands, nor impose the controlled conditions needed to attribute performance changes to the provider rather than to incidental dataset differences. DSR is also well suited to computing research because it explicitly treats the construction of the artefact as a form of knowledge contribution, while retaining the rigour of systematic evaluation. Because the evaluation criteria, in-provider reliability, cross-provider portability, and freedom from metadata contamination, are defined in advance, the artefact's success or failure can be judged objectively against them rather than by post hoc rationalisation.

3.2 Research Methods and Materials

Methods

The research employs a quantitative methodology, which is appropriate because the research questions concern measurable classification performance, statistical reliability, and the magnitude of performance change under provider transfer — all of which require numerical experimentation and comparison rather than interpretive analysis. The research strategy combines controlled experimentation with software development: the artefact is implemented in software, and its behaviour is measured under systematically varied conditions, allowing causal attribution of performance changes to specific manipulations.

The system design follows a five-stage pipeline. Raw Azure and AWS flow logs are first normalised into a unified twenty-eight-field forensic schema that captures identity, temporal, network, and chain-of-custody attributes. Fifty-five features are then engineered from this schema. A hybrid feature-selection stage applies XGBoost and LightGBM, ranks features by SHAP importance, and — to prevent leakage — is fitted on the training partition only, with the selection restricted to the source provider in cross-provider experiments. The selected features are passed to four classifiers (LogEDL, BERT-Log, BiCNN with Attention, and BiLSTM with Attention), each representing a distinct paradigm: uncertainty-aware, semantic and contextual, local-pattern, and temporal-sequential respectively. Finally, the evaluation stage measures performance under the in-provider and cross-provider conditions.

The unified schema is central to the design because it is what makes the provider effect measurable. It records, for every event, a consistent set of identity attributes (such as the acting principal and resource), temporal attributes (timestamp and derived cyclical features), network attributes (address octets and port-role indicators), and chain-of-custody attributes (collection method, handler identity, retention policy, legal basis, and integrity hashing). Because Azure and AWS events are mapped into precisely the same fields, the classifier sees an identical structure regardless of origin, and any residual difference in behaviour can therefore be attributed to the provider itself rather than to schema mismatch. The engineered features deliberately include both behavioural signals, which should transfer across providers, and provider-identifying signals, whose influence the ablation is designed to expose. The four classifiers are configured comparably, trained under the same optimiser, batch size, and epoch budget, and the best model on the validation set is retained, so that the comparison reflects architectural differences rather than differences in training effort.

For data analysis and evaluation, models are assessed using weighted F1 and Macro-AUC for overall performance, supplemented by per-class precision, recall, and F1 to reveal class-specific behaviour. Reliability is assessed through confidence calibration and, for LogEDL, evidential uncertainty, reflecting the admissibility requirement that a classifier's confidence be trustworthy. The core experiment measures cross-provider portability by training on one provider and testing on the other. Two ablation axes are then analysed: a leakage axis (training with versus without label-derived event flags) to confirm that any observed domain shift is genuine rather than an artefact, and a custody axis (training with versus without chain-of-custody fields) to detect provider-

identity contamination. A stratified 70/15/15 train–validation–test split is used throughout, with fixed random seeds to ensure reproducibility.

Materials

The research data comprise 248,828 network security event records — 146,595 from Azure and 102,233 from AWS — spanning five traffic classes: normal, DoS, DDoS, port scanning, and SSH brute force. Data are collected in researcher-controlled cloud environments, and attack traffic is generated using the standard tools nmap (port scanning), Hydra (SSH brute force), and hping3 (DoS and DDoS). This self-generated, provider-matched dataset is appropriate because no public benchmark provides identical attack classes normalised across both Azure and AWS, and such matching is essential to isolate the provider effect that is central to the research questions. With respect to ethical considerations, all attacks are conducted exclusively on the researcher's own isolated virtual machines within private accounts; no third-party systems are targeted and no personal data are collected, and the study will proceed only following institutional ethics approval (see Appendix A). The software stack comprises Python with PyTorch for model implementation, together with scikit-learn, XGBoost, LightGBM, and the SHAP library for feature selection and interpretation, and standard scientific-computing libraries for analysis. Hardware requirements are modest and reproducible: model training is performed on Kaggle-hosted NVIDIA T4 GPUs, supported by a Linux virtual private server for data management, with Azure and AWS free-tier sandboxes used for data collection.

4. Expected Outcome

The research is expected to yield several outcomes. In terms of anticipated results, it is hypothesised that all four classifiers will achieve high in-provider performance, with weighted F1 above 0.99, but will exhibit a severe drop under cross-provider transfer, confirming that aggregate accuracy overstates evidentiary reliability. It is further anticipated that architectures exploiting relational or uncertainty information will degrade less abruptly than purely sequential ones, and that removing chain-of-custody metadata will measurably alter cross-provider behaviour, demonstrating contamination. The implications are both theoretical and practical: theoretically, the work will establish provider-transfer testing and metadata-contamination analysis as necessary components of evaluating AI evidence classifiers; practically, it will offer investigators concrete, confidence-calibrated guidance on the admissibility of cross-provider AI classifications. The principal limitations are that the dataset is synthetic and researcher-generated rather than a public benchmark, that only two providers are

examined (excluding Google Cloud), that the analysis operates at the network-flow layer rather than performing full incident reconstruction, and that a random rather than temporal split is used; a temporal-split evaluation is identified as a valuable extension for future work.

Taken together, these outcomes are expected to make three distinct contributions. The first is methodological: a reproducible protocol for isolating the provider effect using a matched dataset, which other researchers can apply to further providers such as Google Cloud. The second is empirical: a quantified account of how much cross-provider degradation is genuine and how much is attributable to leakage or metadata contamination, addressing directly the confounding that the reviewed literature leaves unresolved. The third is practical: a set of confidence-calibrated guidelines indicating the conditions under which an AI classification may responsibly be treated as admissible evidence. Even where the results confirm rather than overturn expectations, establishing these properties with controlled evidence is itself a contribution, because the current absence of such evidence is precisely what allows overstated reliability claims to persist.

5. Timeline

The project is planned across approximately eight months, structured into seven phases with defined milestones, as summarised below.

Phase	Timeframe	Major tasks and milestones
Phase 1	Week 1	Finalise literature review and proposal; obtain institutional ethics approval; provision Azure and AWS sandbox environments.
Phase 2	Week 2	Collect matched attack and benign traffic; map raw logs into the unified 28-field forensic schema. Milestone: validated dataset.
Phase 3	Week 3	Engineer features; run SHAP-based hybrid XGBoost/LightGBM selection with leakage safeguards. Milestone: selected feature sets.
Phase 4	Week 4	Implement and train LogEDL, BERT-Log, BiCNN+Attention and BiLSTM+Attention. Milestone: four trained baseline models.
Phase 5	Week 5	Conduct cross-provider transfer and the two ablation experiments (leakage and custody axes). Milestone: complete results set.
Phase 6	Week 6	Analyse results; compute calibration, uncertainty and per-class reliability. Milestone: validated findings.
Phase 7	Week 6	Write, revise and submit the dissertation. Milestone: final submission.

References

- Alenezi, A., Atlam, H.F. and Wills, G.B. (2019) 'Experts reviews of a cloud forensic readiness framework for organizations', *Journal of Cloud Computing*, 8(1), pp. 1–14.
- Apruzzese, G., Pajola, L. and Conti, M. (2022) 'The cross-evaluation of machine learning-based network intrusion detection systems', *IEEE Transactions on Network and Service Management*, 19(4), pp. 5152–5169.
- Basharat, M. (2025) 'Digital forensics in 2026: techniques for cloud-native investigations'. Available at: medium.com (Accessed: July 2026).
- Chen, S. and Liao, H. (2022) 'BERT-Log: anomaly detection for system logs based on pre-trained language model', *Applied Artificial Intelligence*, 36(1), 2145642.
- Chen, T. and Guestrin, C. (2016) 'XGBoost: a scalable tree boosting system', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.
- 'Cross-dataset temporal and semantic generalization of intrusion detection models for the future internet' (2026) *Future Internet*, 18(4), 194.
- 'Enhancing generalization of cross-domain intrusion detection: a heterogeneous deep stacked ensemble approach' (2025) *Connection Science*, 37(1), 2599708.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. and Lempitsky, V. (2016) 'Domain-adversarial training of neural networks', *Journal of Machine Learning Research*, 17(59), pp. 1–35.
- Guo, H., Yuan, S. and Wu, X. (2021) 'LogBERT: log anomaly detection via BERT', *International Joint Conference on Neural Networks (IJCNN)*. arXiv:2103.04475.
- Guo, Y., Wu, Y., Tang, Y., Guo, X. and Qi, L. (2024) 'LogEDL: log anomaly detection via evidential deep learning', *Applied Sciences*, 14(16), 7055.
- Hakim, M.A., et al. (2026) 'Cross-domain generalization failure in lightweight intrusion detection models for IIoT networks'. arXiv:2607.00553.
- Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004) 'Design science in information systems research', *MIS Quarterly*, 28(1), pp. 75–105.
- Idugboe, F., Okonkwo, H.O., Ejiyere, B.O. and Ugo, E.I. (2026) 'Machine learning-driven intrusion detection with forensic readiness in cloud-native IoT environments', *American Journal of Electrical and Computer Engineering*, 10(1), pp. 1–18.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T. (2017) 'LightGBM: a highly efficient gradient boosting decision tree', *Advances in Neural Information Processing Systems*, 30.
- Kebande, V.R. and Choo, K.-K.R. (2022) 'Finite state machine for cloud forensic readiness as a service (CFRaaS) events', *Security and Privacy*, 5(1), e182.
- Layeghy, S. and Portmann, M. (2023) 'DI-NIDS: domain invariant network intrusion detection system', *Knowledge-Based Systems*, 273, 110626.

- Lundberg, S.M. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems*, 30.
- Peffers, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S. (2007) 'A design science research methodology for information systems research', *Journal of Management Information Systems*, 24(3), pp. 45–77.
- Sensoy, M., Kaplan, L. and Kandemir, M. (2018) 'Evidential deep learning to quantify classification uncertainty', *Advances in Neural Information Processing Systems*, 31.
- 'A transfer-aware, deployment-oriented evaluation framework for NetFlow-based intrusion detection systems (TAN-IDS)' (2026) *PLOS One*, 21(4), e0346801.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, 30.
- Wu, S., Sun, W., Ding, Z. and Liu, S. (2022) 'Cloud evidence tracing system: an integrated forensics investigation system for large-scale public cloud platform', *Forensic Science International: Digital Investigation*, 41, 301391.