Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

An Explainable AI-Driven Framework for Risk-Adaptive Security and Selective Auditing of Cloud Virtual Machines: Paper I—Explainable Regression-Based Selective Auditing

Virendra Singh Thakur[1], Prof. Dipti Sharma[2]

[1]M.Tech Scholar, Department of Computer Science and Engineering, Patel College of Science and Technology
[2]Assistant Professor, Department of Computer Science and Engineering, Patel College of Science and Technology

Abstract

Cloud audit pipelines that apply every extended rule to every virtual-machine event can consume review capacity on low-priority activity. Purely statistical filtering, however, can omit resources that policy requires an organization to review. This paper presents ACB-RM-SA, a policy-first framework that combines an interpretable ridge-regression score for virtual-machine sensitivity with target-specific auditor suitability and deterministic mandatory-review gates. A reproducible synthetic evaluation generated 400 virtual-machine profiles, 120 identities and 60,000 access events using random seed 42. The ridge model was trained on 280 profiles and tested on 120 profiles, yielding MAE 0.0141, RMSE 0.0178 and R^2 0.9807 against simulated assessor ratings. The high fit is expected because those ratings were produced from a declared, related linear policy with small interactions and noise; it is an internal consistency check rather than evidence of production predictive validity. At the capacity-driven threshold of 0.65, the framework selected 33,787 events (56.31%) for extended review and reduced modeled rule-evaluation work by 34.59%. In the 400-VM local batch benchmark, mean processing time decreased from 10.15 ms to 7.26 ms (28.49%). All 18,044 policy-critical events were retained by deterministic gating. This 100% figure is a policy-coverage property, not attack-detection recall. The results support interpretable prioritization while preserving explicit security obligations, but real-cloud telemetry and analyst validation remain necessary.

Keywords: Cloud Auditing, Virtual Machines, Ridge Regression, Least Privilege, Risk-Adaptive Access Control, Cloud Security

1. Introduction

Cloud platforms combine elastic infrastructure, managed services and rapidly changing application components, which creates a broad audit surface spanning identities, workloads, data and control planes [6–8]. Recent cloud-security reviews consequently emphasize that assurance must connect technical controls with observable evidence rather than treat all events as equivalent [6–8].

Exhaustive evaluation is attractive because it appears conservative: every event receives the complete set of rules. In a large environment, however, the approach duplicates expensive transformations and creates evidence that investigators may never inspect. The alternative—using a learned score as the sole authority—is also unsafe. A low predicted score must not suppress a contractual, regulatory or segregation-of-duties requirement. Current zero-trust guidance similarly emphasizes explicit policy decisions, continuous context and removal of implicit trust based on network location [1,2].

ACB-RM-SA addresses this tension by separating prioritization from authority. An explainable regression model estimates virtual-machine sensitivity. A second, transparent score evaluates whether a particular identity is suitable for a particular resource. Hard-deny and mandatory-review rules execute independently and cannot be overridden by the regression result. Only the remaining selected events enter the extended rule set. The design follows least privilege at the evidence layer: access to audit material is granted for a justified target rather than through a permanent global auditor status.

The paper makes three contributions. First, it specifies an auditable ridge-regression model whose coefficients expose the effect of compute, memory, storage, network and data-criticality factors. Second, it integrates that model with deterministic security gates and target-specific suitability. Third, it provides a fully seeded synthetic experiment, raw timing trials and hashes so that every reported number can be regenerated. The evaluation intentionally makes no claim of production attack-detection accuracy.

2. Related Work

Recent reviews of cloud-storage integrity distinguish verification models, attack assumptions, key-management choices and auditing overheads [12]. Serverless implementations illustrate how integrity checks can be operationalized with elastic cloud services [13]. These mechanisms address whether retained data or evidence remains intact; they do not, by themselves, decide which virtual-machine events should enter a capacity-limited operational audit queue.

Contemporary control catalogs and guidance provide a governance foundation for access control, logging, continuous monitoring and evidence protection [3–5]. Surveys of cloud-native services and security frameworks show that these controls must be adapted to distributed workloads and shared-responsibility boundaries [7,8]. Automated graphical security analysis is another route to consistent cloud-policy enforcement [14]. ACB-RM-SA complements these approaches by making the audit-selection decision explicit and reproducible.

Recent mapping and design studies show that cloud-native access control combines subject, resource and environmental attributes, while risk-adaptive models add contextual estimates to the authorization decision [9–11]. A zero-trust dynamic access-control design likewise illustrates the use of changing context instead of permanent trust [10]. The present framework applies these ideas to audit prioritization but preserves a strict boundary: deterministic policy remains authoritative, and regression is used only for transparent ranking.

Explainability is especially important when learning components influence cybersecurity operations [16–18]. Interpretable-model research favors models whose behavior can be inspected directly when the

problem permits it [15], while systematic work on explanation evaluation warns that plausible examples alone are not sufficient evidence [19]. Security-ML evaluations are also vulnerable to data leakage, unrealistic labels and overextended claims [20]. These findings motivate the paper's simple coefficient-based model, held-out test split and explicit limitation of claims.

3. Problem Definition and Research Questions

Let $V = \{v_1, \ldots, v_n\}$ be cloud virtual machines and $E = \{e_1, \ldots, e_m\}$ be access or change events. Each VM has a normalized five-dimensional profile, and each event links a VM to an identity and environmental context. An exhaustive comparator applies five inexpensive global gates and nineteen extended rules to every event. The proposed method applies the same five gates globally but evaluates the nineteen extended rules only for events selected by mandatory policy or transparent risk criteria.

RQ1: How closely can the interpretable ridge model reproduce held-out simulated assessor ratings?

RQ2: At a capacity-constrained operating threshold, how much extended rule-evaluation workload is avoided?

RQ3: Does mandatory policy gating retain every policy-critical event regardless of the regression threshold?

RQ4: How does the threshold trade selected workload against rule-evaluation reduction?

The objective is not to minimize latency without constraint. The operating point must retain all policy-critical events, keep the detailed-review share below the declared capacity target of 60%, and produce reasons that an auditor can inspect.

4. Proposed ACB-RM-SA Framework

Figure 1. Policy-First ACB-RM-SA Processing Architecture

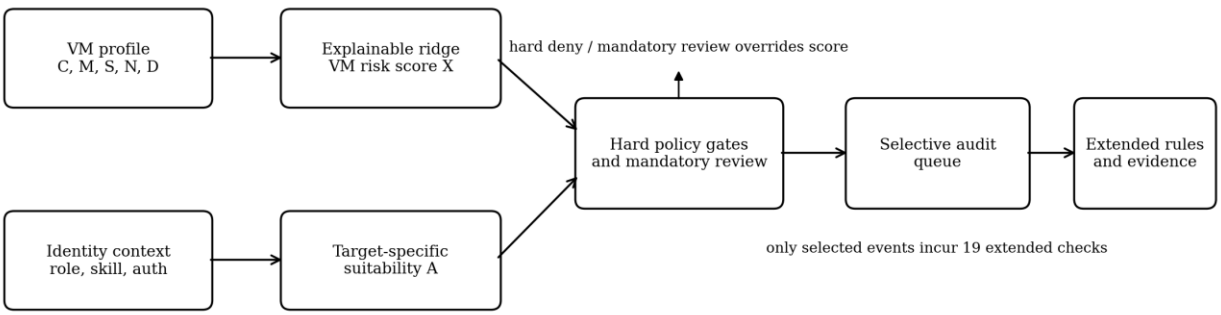


Figure 1 separates two scoring paths from the policy authority. The VM profile produces sensitivity X_i. Identity and target attributes produce suitability A_{ij}. Hard-deny and mandatory-review gates receive

both contexts and can route an event directly to restriction or extended review. The score therefore explains priority but never cancels an explicit security obligation.

4.1 VM Sensitivity Model

For VM i, the feature vector contains normalized compute exposure C_i, memory exposure M_i, storage exposure S_i, network exposure N_i and data criticality D_i. A regularized linear estimator is used because its coefficients can be inspected directly, consistent with current guidance favoring interpretable models where the task permits [15]. The fitted value is clipped to the unit interval:

$$X_i = \text{clip}(\beta_0 + \beta_1 C_i + \beta_2 M_i + \beta_3 S_i + \beta_4 N_i + \beta_5 D_i, 0, 1). \quad (1)$$

Ridge parameter $\alpha = 0.50$ was fixed before the reported run. Coefficients are global and readable, which allows a reviewer to see that data criticality carries the largest influence in this synthetic policy. A deployment could add richer features, but any recalibration should be approved, versioned and tested in shadow mode rather than changed autonomously during enforcement.

Table 1. Model Variables and Operational Interpretation

Symbol	Range	Interpretation
C	0–1	Compute exposure and workload intensity
M	0–1	Memory exposure and sensitivity of in-memory state
S	0–1	Persistent-storage exposure and dependency
N	0–1	Network reachability and external connectivity
D	0–1	Business and information criticality
X	0–1	Predicted VM sensitivity after clipping
A	0–1	Identity suitability for a specific target

4.2 Target-Specific Auditor Suitability

Suitability is evaluated for identity j and target i rather than assigned once to an auditor. R_{ij} represents role compatibility, T_{ij} competence for the target category, K_j authentication confidence and H_j compliance history. The declared score is:

$$A_{ij} = 0.45R_{ij} + 0.25T_{ij} + 0.20K_j + 0.10H_j. \quad (2)$$

An identity is unsuitable when $A_{ij} < X_i$. The relationship does not create a permanent trust rank, and missing attributes are not interpreted as zero risk. In an operational implementation, stale critical-resource metadata or unavailable authentication evidence should trigger conservative review.

4.3 Policy-First Selection Algorithm

Hard-deny conditions include expired authorization, a simulated segregation-of-duties violation, missing strong authentication for a high-risk VM, or role incompatibility on a VM whose risk is at least 0.62. Policy-critical status additionally includes VM risk of at least 0.78, a high-criticality and novel-source conjunction, or a change magnitude of at least 0.88. These conditions are deterministic.

$$\text{Selected}(e) = \text{Critical}(e) \vee [X_i \geq \tau] \vee [(A_{ij} < X_i) \wedge (X_i \geq 0.50)] \vee \text{Context}(e). \quad (3)$$

Context(e) is true when source novelty is at least 0.58 and the event is off-hours or has data volume of at least 0.78. The operating threshold $\tau = 0.65$ is the first tested point that satisfies the predeclared capacity objective of selecting fewer than 60% of events while mandatory gates preserve policy-critical coverage. It was not chosen to maximize measured latency reduction.

Algorithm 1. Selective Audit Decision
1. Normalize VM, identity and event attributes; reject stale critical metadata.
2. Compute X_i with Equation (1) and A_{ij} with Equation (2).
3. Evaluate hard-deny and mandatory-review conditions.
4. If a hard deny is true, restrict the action and retain evidence.
5. Otherwise evaluate Equation (3); if true, enter the extended audit queue.
6. Record score, coefficient version, policy reason, selected rules and timing.

5. Reproducible Experimental Method

5.1 Synthetic Profiles and Events

The experiment uses synthetic data to avoid disclosure of organizational telemetry and to make regeneration possible. Random seed 42 produces 400 VMs distributed among web, application, database, analytics and administrative categories. Category-specific beta distributions generate the five normalized features. A latent reference policy uses intercept 0.08 and feature weights 0.14, 0.12, 0.17, 0.16 and 0.38, plus small storage–criticality and network–criticality interactions. Three independently perturbed simulated assessor ratings are averaged for each VM. No human experts participated in creating these labels.

The identity set contains 120 synthetic operations, development, security, audit and service accounts. Authentication confidence, compliance history, multi-factor status and category competence are sampled from declared distributions. The event generator creates exactly 150 events per VM, for 60,000 events. It adds role compatibility, source novelty, off-hours context, change magnitude, normalized data volume, authorization expiry and segregation signals. This is a controlled workload, not a reconstruction of a named cloud provider or real attack campaign.

Table 2. Experimental Configuration

Item	Declared Value
Random seed	42
Virtual machines	400
Identities	120
Events	60,000; 150 per VM
VM categories	Web, application, database, analytics, administrative
Regression split	280 train; 120 test; stratified 70/30
Ridge regularization	$\alpha = 0.50$
Operating threshold	$\tau = 0.65$

Item	Declared Value
Load levels	50, 100, 200, 300 and 400 active VMs
Timing protocol	5 warm-ups; 30 measured paired repetitions; 6 inner repeats
Rule model	5 global gates; 19 extended rules

5.2 Train/Test Protocol and Metrics

A stratified 70/30 split is created with the fixed seed. The ridge model is fitted only on the 280 training VMs using scikit-learn, and MAE, RMSE and R^2 are computed on the 120 held-out VMs. Predictions for all VMs are then used to construct the event workload. Because the target ratings are generated from a closely related policy, regression metrics measure implementation fidelity and the effect of interactions and noise. They do not establish external generalization; security-ML studies require particular care to avoid unrealistic performance interpretations [20].

Selection is evaluated through the selected-event fraction and policy-critical coverage. Policy-critical coverage is the number of policy-critical events selected for restriction or extended review divided by all policy-critical events. It must not be interpreted as precision, recall or detection rate for attacks because the synthetic generator does not provide validated attack labels.

5.3 Work and Timing Benchmark

The exhaustive comparator and ACB-RM-SA execute identical global gates and identical nonlinear extended-rule transformations. Their only intended difference is scope: the comparator applies nineteen extended rules to every event, while ACB-RM-SA applies them to selected events. Modeled work units are therefore:

$$W_{baseline} = 24m; \quad W_{proposed} = 5m + 19s, \quad (4)$$

where m is the number of events and s is the selected-event count. Wall-clock time uses a local in-memory NumPy/Pandas batch kernel. Each load level receives five warm-ups, then thirty repetitions with alternating method order; each recorded value averages six inner executions. The reported error bars are sample standard deviations, and the paired mean difference uses a two-sided 95% Student-t interval with 29 degrees of freedom. Offline generation, model fitting, network transfer, persistent logging and analyst time are excluded.

6. Results

6.1 Regression Fidelity

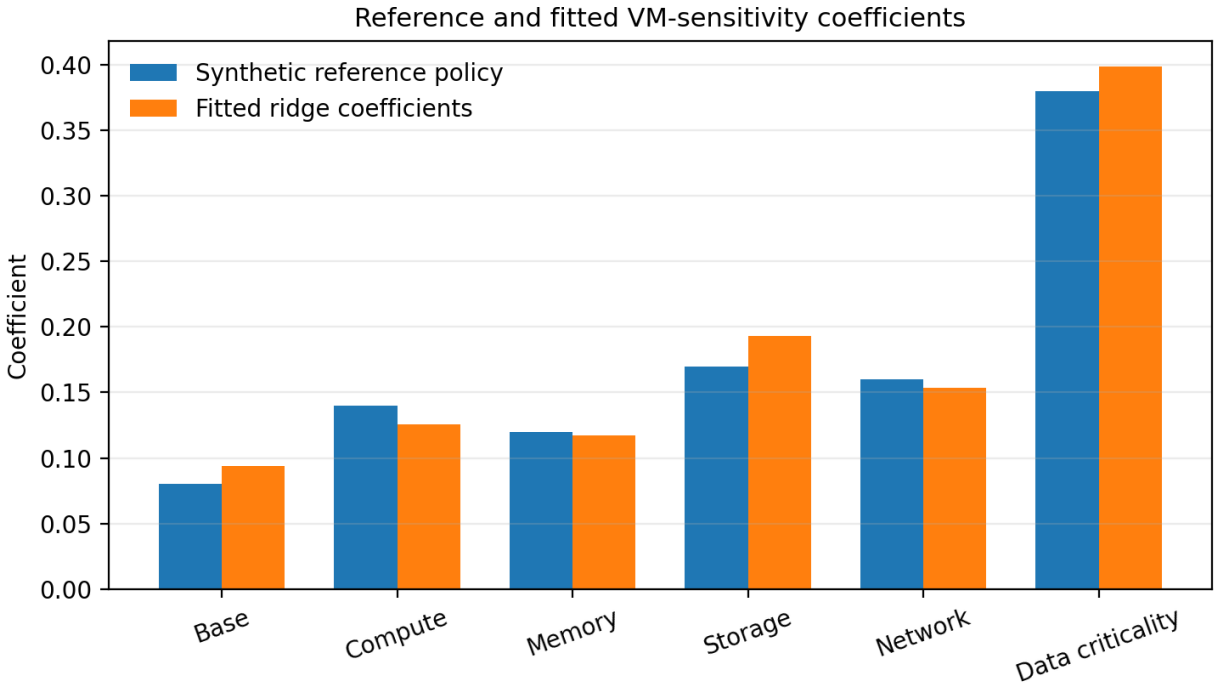
On the held-out 120 VMs, the model obtained MAE 0.0141, RMSE 0.0178 and R^2 0.9807. The fitted coefficients remain close to the declared reference policy, as shown in Table 3 and Figure 2. Data criticality is the strongest factor. The small differences arise from the two declared interactions, random perturbations and ridge shrinkage. The result answers RQ1 only for the synthetic generator.

Table 3. Reference and Fitted VM-Sensitivity Coefficients

Term	Reference Policy	Fitted Ridge

Term	Reference Policy	Fitted Ridge
Intercept	0.0800	0.0940
Compute	0.1400	0.1256
Memory	0.1200	0.1173
Storage	0.1700	0.1934
Network	0.1600	0.1538
Data Criticality	0.3800	0.3985

Figure 2. Reference and Fitted VM-Sensitivity Coefficients



6.2 Audit Scope and Mandatory Coverage

At $\tau = 0.65$, ACB-RM-SA selected 33,787 of 60,000 events (56.31%). The generator produced 18,044 policy-critical events and 9,497 hard-deny events. All policy-critical events were selected because Critical(e) is an explicit disjunct in Equation (3). This answers RQ3 as a property of the policy implementation, not as empirical evidence that all malicious behavior would be detected.

The operating point meets the capacity constraint because the detailed queue remains below 60%. An event can still be selected below the VM threshold when suitability is insufficient or when contextual novelty is combined with off-hours or high-volume behavior. This prevents the threshold from becoming a single point of failure.

6.3 Modeled Work and Measured Batch Time

Table 4. Batch Timing by Active VM Count (Mean ± SD, n = 30)

VMs	Events	Selected (%)	Baseline (ms)	ACB-RM-SA (ms)	Reduction (%)
50	7,500	59.36	1.27 ± 0.19	1.01 ± 0.18	20.72
100	15,000	56.25	2.19 ± 0.06	1.80 ± 0.07	17.90
200	30,000	56.74	4.62 ± 0.27	3.48 ± 0.46	24.77
300	45,000	55.51	7.35 ± 0.64	5.71 ± 0.43	22.36
400	60,000	56.31	10.15 ± 1.11	7.26 ± 0.36	28.49

At 400 active VMs, the exhaustive path required 1,440,000 modeled work units, while ACB-RM-SA required 941,953, a reduction of 34.59%. Mean local batch time decreased from 10.15 ms to 7.26 ms. The paired mean difference was 2.891 ms with a 95% confidence interval of 2.497–3.285 ms. Across the five evaluated loads, the observed mean reduction ranged from 17.90% to 28.49%. This supports RQ2 for the declared in-memory kernel.

Figure 3. Measured Batch Processing Time at Five Load Levels

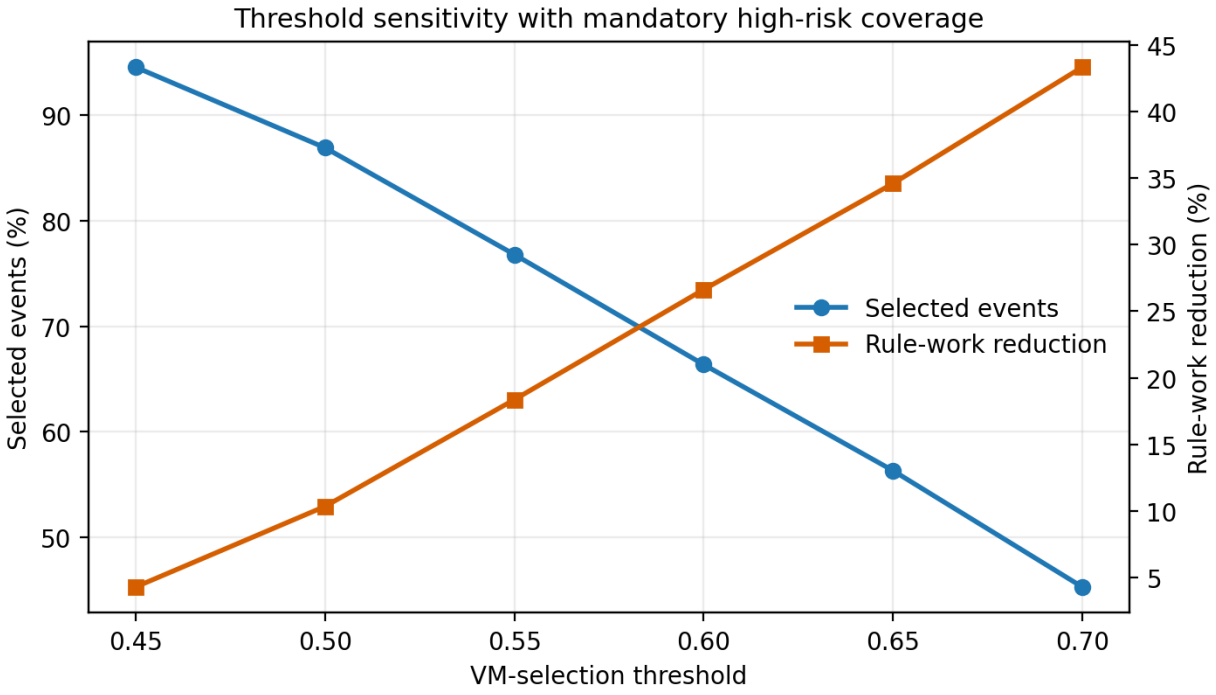
Timing does not scale perfectly because operating-system scheduling, cache state and runtime noise affect millisecond-level measurements. More importantly, these values exclude network calls, database writes and analyst activity. They should be used to reproduce the relative scope effect, not as a service-level prediction for a production cloud.

6.4 Threshold Sensitivity

Table 5. Threshold Sensitivity at 400 VMs

Threshold	Selected Events (%)	Work Reduction (%)	Policy-Critical Coverage (%)
0.45	94.57	4.30	100.00
0.50	86.93	10.35	100.00
0.55	76.81	18.36	100.00
0.60	66.37	26.63	100.00
0.65	56.31	34.59	100.00
0.70	45.28	43.32	100.00

Figure 4. Threshold Trade-Off Under Mandatory Policy Gating



Raising τ from 0.45 to 0.70 reduces the selected share from 94.57% to 45.28% and increases modeled work reduction from 4.30% to 43.32%. Policy-critical coverage remains 100% at every tested point because it is not threshold-dependent. The sensitivity curve answers RQ4 and makes the governance trade-off visible: a higher threshold frees capacity but increases reliance on contextual and mandatory gates for events below the threshold.

7. Discussion

The results demonstrate the computational consequence of separating cheap global gates from an extended evidence path. At the selected threshold, roughly 44% of events avoid nineteen extended

transformations. The framework remains explainable at three levels: global ridge coefficients show how VM features influence X; suitability weights show why a specific identity is or is not appropriate for a target; and the selection record stores the exact policy or contextual reason that admitted an event to review.

Governance is more important than the numerical threshold. Organizations should define mandatory review from law, contract, control objectives and threat models before tuning capacity. Regression coefficients and threshold versions should be protected configuration items. A reviewer should be able to reconstruct the input profile, coefficient set, hard-rule result and evidence path for each decision. This approach is compatible with current zero-trust and risk-adaptive access-control principles [1,2,9–11], but it does not replace identity proofing, secure logging or integrity controls.

The method also narrows evidence exposure. A target-specific suitability check avoids treating membership in an auditor group as unrestricted access to every VM record. Least privilege is strengthened when the evidence system releases only the resources and fields required for a justified assignment. Recent cloud-integrity and operational auditing mechanisms [12,13] can be layered under the selection mechanism to protect retained evidence.

8. Threats to Validity and Limitations

Construct validity is limited by synthetic labels. The simulated assessor ratings are derived from a policy that closely resembles the fitted model, which explains the high R^2. Real assessors may disagree, apply nonlinear reasoning or use unavailable contextual knowledge. The study therefore reports regression fidelity rather than risk-prediction accuracy. No classification or attack-detection accuracy is claimed.

Internal validity is limited by a single deterministic generator and one seed. The fixed seed is valuable for exact reproduction but does not measure variation across organizations. Although timing order alternates and thirty paired repetitions are retained, local CPU scheduling remains a nuisance factor. The workload formula isolates rule scope, but each real rule can have a different cost.

External validity is limited because the benchmark contains no live OpenStack, AWS, Azure or Google Cloud telemetry, no network round trips and no persistent evidence store. The 100% policy-critical coverage result is guaranteed by code and does not measure previously unknown threats. Adversaries could manipulate stale or weak resource attributes, so deployment must protect metadata provenance and fail conservatively. Trustworthy-AI and adversarial-ML guidance similarly requires lifecycle risk management rather than reliance on a single performance score [21,22].

Future work should run multi-seed experiments, obtain authorized and de-identified operational traces, compare coefficients with independent analyst ratings, report tail latency, and measure investigation outcomes. A shadow deployment should compare selected and non-selected samples before the framework is permitted to affect enforcement.

9. Conclusion

ACB-RM-SA combines explainable VM-sensitivity regression, target-specific auditor suitability and deterministic policy gates. In the reproducible synthetic workload, the capacity-driven operating point selected 56.31% of events, reduced modeled extended-rule work by 34.59% and lowered mean local batch time at all tested loads. Mandatory gates retained every policy-critical event by construction. The contribution is a transparent and testable prioritization architecture, not proof of complete cloud security. Production use requires locally approved policy, protected telemetry, independent validation and human escalation.

Data Availability and Reproducibility

The accompanying reproducibility package contains the Python experiment, generated VM, identity and event tables, fitted coefficients, threshold sweep, raw paired timing trials, summary JSON and figures. Running run_experiment.py with the documented environment and seed regenerates the artifacts. Full SHA-256 hashes are recorded in summary.json and the package manifest. Software versions were Python 3.12.13, NumPy 2.3.5, pandas 2.2.3 and scikit-learn 1.8.0. Versioned source and dependency records also support the traceability practices recommended for secure software development [23].

Ethics and Privacy Statement

This study used only procedurally generated synthetic profiles and events. It involved no human participants, personal records, production credentials or customer cloud telemetry.

Author Contributions

Virendra Singh Thakur: conceptualization, methodology, software, formal analysis, visualization and original draft. Dipti Sharma: supervision, methodology review, validation, and review and editing of the manuscript. Both authors are responsible for approving the submitted version.

References

1. Cybersecurity and Infrastructure Security Agency, Zero Trust Maturity Model, Version 2.0, April 2023. https://www.cisa.gov/resources-tools/resources/zero-trust-maturity-model.
2. 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, September 2023. https://doi.org/10.6028/NIST.SP.800-207A.
3. Cloud Security Alliance, Cloud Controls Matrix and CAIQ, Version 4.1, January 2026. https://cloudsecurityalliance.org/artifacts/cloud-controls-matrix-v4-1.
4. Cloud Security Alliance, Security Guidance for Critical Areas of Focus in Cloud Computing, Version 5, July 2024. https://cloudsecurityalliance.org/artifacts/security-guidance-v5.
5. C. Pascoe, S. Quinn and K. Scarfone, The NIST Cybersecurity Framework (CSF) 2.0, NIST CSWP 29, February 2024. https://doi.org/10.6028/NIST.CSWP.29.
6. F. K. Parast, C. Sindhav, S. Nikam, H. I. Yekta, K. B. Kent and S. Hakak, “Cloud computing security: A survey of service-based models,” Computers & Security, vol. 114, Article 102580, 2022. https://doi.org/10.1016/j.cose.2021.102580.

7. T. Theodoropoulos et al., “Security in cloud-native services: A survey,” Journal of Cybersecurity and Privacy, vol. 3, no. 4, pp. 758–793, 2023. https://doi.org/10.3390/jcp3040034.
8. M. Chauhan and S. Shiaeles, “An analysis of cloud security frameworks, problems and proposed solutions,” Network, vol. 3, no. 3, pp. 422–450, 2023. https://doi.org/10.3390/network3030018.
9. M. S. Rahaman, S. N. Tisha, E. Song and T. Cerny, “Access control design practice and solutions in cloud-native architecture: A systematic mapping study,” Sensors, vol. 23, no. 7, Article 3413, 2023. https://doi.org/10.3390/s23073413.
10. R. Wang, C. Li, K. Zhang and B. Tu, “Zero-trust based dynamic access control for cloud computing,” Cybersecurity, vol. 8, Article 12, 2025. https://doi.org/10.1186/s42400-024-00320-x.
11. N. Alharbe, A. Aljohani, M. A. Rakrouki and M. Khayyat, “An access control model based on system security risk for dynamic sensitive data storage in the cloud,” Applied Sciences, vol. 13, no. 5, Article 3187, 2023. https://doi.org/10.3390/app13053187.
12. P. Goswami, N. Faujdar, S. Debnath, A. K. Khan and G. Singh, “Investigation on storage level data integrity strategies in cloud computing: Classification, security obstructions, challenges and vulnerability,” Journal of Cloud Computing, vol. 13, Article 45, 2024. https://doi.org/10.1186/s13677-024-00605-z.
13. F. Chen, J. Cai, T. Xiang et al., “Practical cloud storage auditing using serverless computing,” Science China Information Sciences, vol. 67, Article 132102, 2024. https://doi.org/10.1007/s11432-022-3597-3.
14. S. An, A. Leung, J. B. Hong, T. Eom and J. S. Park, “Toward automated security analysis and enforcement for cloud computing using graphical models for security,” IEEE Access, vol. 10, pp. 75117–75134, 2022. https://doi.org/10.1109/ACCESS.2022.3190545.
15. C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova and C. Zhong, “Interpretable machine learning: Fundamental principles and 10 grand challenges,” Statistics Surveys, vol. 16, pp. 1–85, 2022. https://doi.org/10.1214/21-SS133.
16. N. Capuano, G. Fenza, V. Loia and C. Stanzione, “Explainable artificial intelligence in cybersecurity: A survey,” IEEE Access, vol. 10, pp. 93575–93600, 2022. https://doi.org/10.1109/ACCESS.2022.3204171.
17. F. Charmet, H. C. Tanuwidjaja, S. Ayoubi et al., “Explainable artificial intelligence for cybersecurity: A literature survey,” Annals of Telecommunications, vol. 77, pp. 789–812, 2022. https://doi.org/10.1007/s12243-022-00926-7.
18. G. Rjoub, J. Bentahar, O. Abdel Wahab, R. Mizouni, A. Song, R. S. Cohen, H. Otrok and A. Mourad, “A survey on explainable artificial intelligence for cybersecurity,” IEEE Transactions on Network and Service Management, vol. 20, no. 4, pp. 5115–5140, 2023. https://doi.org/10.1109/TNSM.2023.3282740.
19. M. Nauta, J. Trienes, S. Pathak et al., “From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI,” ACM Computing Surveys, vol. 55, no. 13s, Article 295, pp. 1–42, 2023. https://doi.org/10.1145/3583558.
20. D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro and K. Rieck, “Dos and don'ts of machine learning in computer security,” in 31st USENIX Security Symposium, pp. 3971–3988, 2022. https://www.usenix.org/conference/usenixsecurity22/presentation/arp.

21. E. Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, January 2023. https://doi.org/10.6028/NIST.AI.100-1.
22. 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 100-2 E2025, March 2025. https://doi.org/10.6028/NIST.AI.100-2e2025.
23. K. Scarfone, M. Souppaya and D. Dodson, Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities, NIST Special Publication 800-218, February 2022. https://doi.org/10.6028/NIST.SP.800-218.