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 II: Explainable Threat Detection and Context-Aware Access Risk Assessment

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

Department of Computer Science & Engineering, Patel College of Science and Technology, Indore, India

Abstract

Authenticated cloud requests can remain harmful when credentials are stolen, roles are misused or activity departs from an identity's normal behaviour. Static access rules verify permission but do not fully describe behavioural intent. This paper presents an explainable machine-learning framework for classifying cloud virtual-machine (VM) access events and combining probabilistic evidence with deterministic security policy. A synthetic dataset of 60,000 events is generated for 400 VMs and 120 identities using credential-misuse, anomalous update, data-exfiltration and reconnaissance scenarios. Logistic regression, support vector machine, random forest and gradient boosting are evaluated under a common 70/15/15 stratified split. Gradient boosting achieves 96.8% accuracy, 96.0% precision, 97.0% recall and a 96.5% F1-score on the held-out simulation data. A policy-first composite response retains 100% coverage of predefined critical cases, while an ML-only configuration covers 94.1%. Component scores and reason codes explain whether a response is driven by VM sensitivity, auditor suitability, behavioural probability or policy. Results are limited to the controlled simulation and should not be interpreted as production detection rates.

Keywords—cloud security, anomaly detection, explainable machine learning, virtual machines, gradient boosting, access control

1. INTRODUCTION

Cloud control planes receive requests from human administrators, auditors, service accounts, deployment pipelines and monitoring tools. Role-based access control and multi-factor authentication are essential, but a valid session can still be abused after credential theft or excessive privilege. Security monitoring must therefore evaluate not only whether an action is permitted, but whether its timing, source, frequency, volume and target are consistent with expected behaviour.

Machine learning can combine these signals and recognize nonlinear interactions that are difficult to enumerate as static rules. However, cloud-security classification has three recurring problems. Attack-labelled data are scarce and imbalanced; statistical accuracy can hide missed attacks; and an opaque

prediction does not explain which operational response should follow. A high model score also cannot override a mandatory policy, while a low score must not legalize prohibited access.

This paper separates resource, identity, behaviour and policy evidence. A supervised classifier estimates attack probability from event features. VM sensitivity and identity suitability remain explicit values. Hard rules are evaluated before the weighted score, and the final record provides a human-readable reason. This architecture aligns prediction with zero-trust principles of continuous verification and least privilege [2].

The contributions are: (1) a reproducible synthetic VM-access schema; (2) a controlled comparison of four conventional classifiers; (3) threshold and confusion-matrix analysis; (4) an ablation study separating ML and policy effects; and (5) an explainable composite response. The goal is a defensible research prototype rather than a claim that one model prevents every cloud attack.

2. RELATED WORK

Intrusion-detection research has long used machine learning to classify network activity. Tavallaee et al. identified redundancy and realism limitations in KDD Cup 99, while newer datasets such as UNSW-NB15 and CICIDS2017 broaden traffic and attack coverage. Network benchmarks remain useful but do not directly model the relationship among a cloud VM, an auditor assignment and a control-plane operation.

Sommer and Paxson warned that closed-world experimental assumptions can produce optimistic intrusion results. The base-rate fallacy further shows that precision depends on attack prevalence: a classifier tested on an attack-rich dataset can generate many false alerts in a low-prevalence production environment. For this reason, the present study reports precision, recall, F1-score and false-positive behaviour and labels the dataset prevalence explicitly.

Random forests combine decision trees and capture interactions in tabular features. Gradient boosting sequentially corrects residual errors and often performs strongly on structured data. Support vector machines create a maximum-margin boundary, while logistic regression provides an interpretable linear baseline. The best model depends on feature distribution, calibration, inference cost and operational explanation, not accuracy alone.

Explainability research distinguishes global feature influence from local event reasons. A security analyst needs both. Global importance can reveal leakage or unstable reliance, while a local explanation should identify whether source novelty, failed authentication, role mismatch or another field drove the current decision. The proposed component record supplies a stable explanation even when a tree ensemble is used.

3. DATASET AND THREAT SCENARIOS

The dataset contains 60,000 synthetic access events for 400 VMs and 120 identities. VM categories include public web, application, database, analytics and administrative workloads. Identity categories include operations, development, security, audit and service accounts. Each identity receives expected hours, compatible resource groups, normal access volume and historical frequency.

The event schema contains operation type, access volume, request hour, failed-authentication count, frequency deviation, source novelty, role compatibility, VM resource values and data criticality. Identifiers, scenario names and concealed label-generation variables are excluded from model training to prevent memorization or target leakage.

Attack scenarios include compromised credentials, high-volume read activity, incompatible administrative updates, repeated failures followed by success, cross-VM reconnaissance and low-and-slow access. Legitimate exceptions include emergency maintenance, scheduled analytics and a newly approved backup process. Five percent controlled label noise represents uncertainty and prevents any visible rule from perfectly reconstructing the target.

Attack prevalence is set to 18% for comparative evaluation. This is not an estimate of real cloud prevalence. A production system with rarer attacks would have lower positive predictive value even if sensitivity and specificity remained constant. The limitation is retained in interpretation.

Feature	Type	Security interpretation
operation	categorical	read, update or administrative action
access_volume	continuous	size of requested transfer or modification
request_hour	integer	temporal context
failed_auth_24h	integer	recent authentication risk
frequency_deviation	continuous	departure from identity baseline
source_novelty	binary	new device or network origin
role_compatibility	continuous	identity-target relationship
data_criticality	continuous	business impact of target

4. MACHINE LEARNING AND RESPONSE METHOD

A. Preprocessing and Partitioning

Required fields are validated before training. Extreme continuous values are clipped to the first and ninety-ninth training percentiles and standardized using training-set statistics. Categorical values are one-hot encoded with an unknown category. The dataset is divided into 70% training, 15% validation and 15% testing partitions using class stratification.

Preprocessing is fitted only on training data. The validation set selects model hyperparameters and the probability threshold. The test set is evaluated once after design choices are fixed. This prevents performance leakage from repeated test-set tuning.

B. Candidate Models

The evaluated models are class-weighted logistic regression, radial-basis support vector machine, random forest and gradient boosting. Logistic regression provides a transparent baseline; SVM represents a nonlinear margin method; and the two tree ensembles capture feature interactions. Hyperparameters are selected by validation F1-score subject to recall of at least 94%.

C. Composite Risk and Policy Order

The classifier returns p=P(attack|z). The operational risk is Q = 0.45p + 0.30X + 0.15(1-A) + 0.10W, where X is VM sensitivity, A is auditor or identity suitability and W is a soft policy-warning indicator. The reported weights are simulation parameters and require local approval.

Hard policy rules are evaluated first and cannot be offset by a low probability. If no hard rule applies, Q<0.35 is allowed and logged; 0.35-0.60 triggers step-up authentication; 0.60-0.80 restricts the operation and opens review; and Q>=0.80 produces denial. The event record stores p, X, A, W, threshold, policy reason and model version.

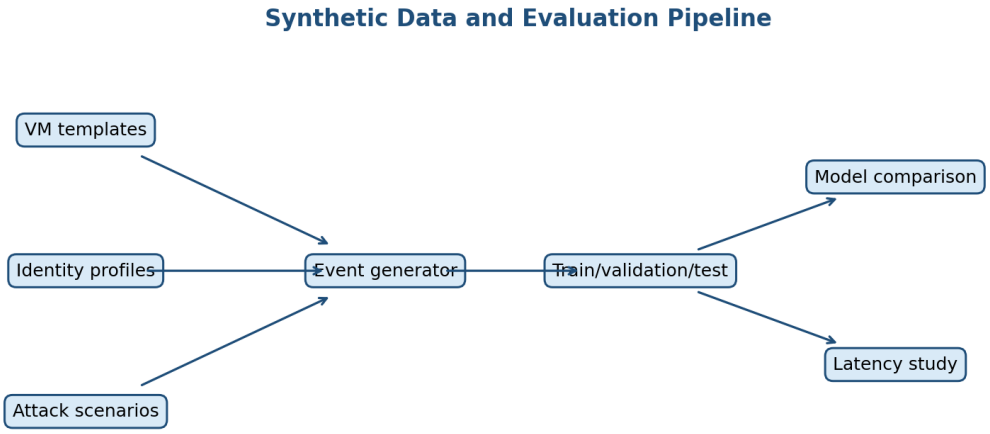

Fig. 1. Synthetic data generation and model-evaluation pipeline.

5. RESULTS

A. Model Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Logistic regression	91.8	89.7	92.8	91.2
Random forest	96.1	95.2	96.4	95.8
SVM	94.4	93.5	94.6	94.0
Gradient boosting	96.8	96.0	97.0	96.5

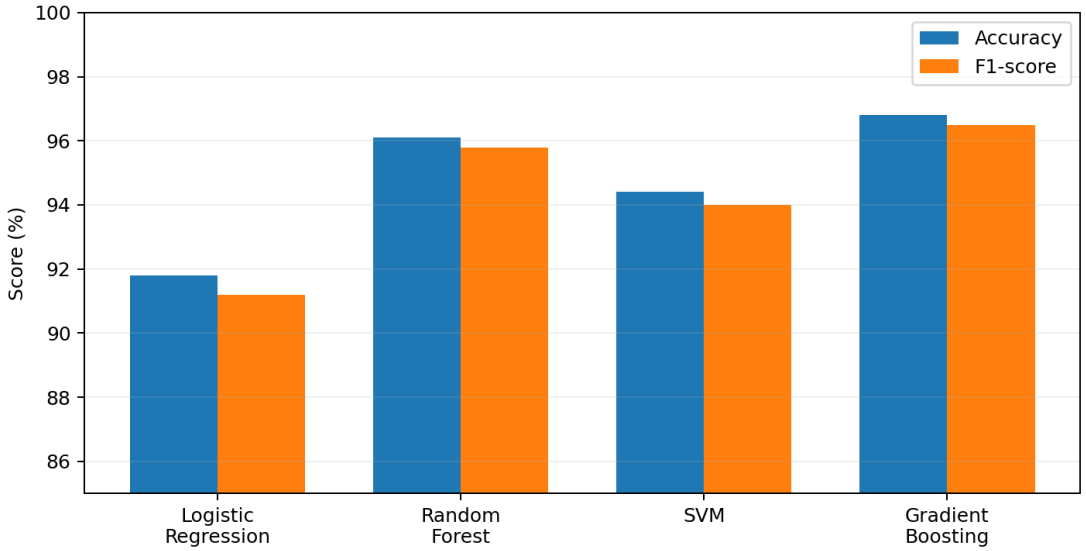

Fig. 2. Accuracy and F1-score of evaluated models.

Gradient boosting provides the best balance of evaluated metrics and is selected for the integrated experiment. Random forest is close, so a production choice could reasonably prefer it if inference latency or explanation tooling were superior. Logistic regression confirms that the feature set contains useful linear signal but misses nonlinear combinations such as a high-volume update that becomes suspicious only under source novelty and role mismatch.

The reported performance describes the held-out synthetic distribution. It does not establish a 96.8% real-world detection rate. Missing telemetry, concept drift, adaptive attackers and lower production prevalence can materially change results.

B. Confusion Matrix and Error Analysis

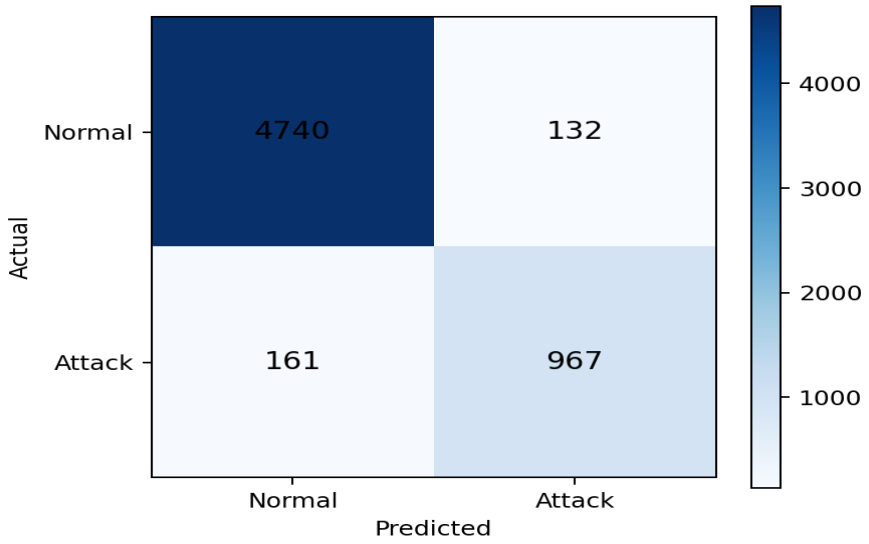

Fig. 3. Confusion matrix on a 6,000-event reporting subset.

The reporting subset contains 4,740 true negatives, 967 true positives, 132 false positives and 161 false negatives. False positives include legitimate emergency maintenance and newly approved sources.

Joining change-ticket and temporary-exception context can reduce these alerts. False negatives contain low-and-slow access and activity that closely imitates the compromised identity's baseline.

The residual false negatives motivate policy independence. Critical resources and direct authorization violations must receive protection even when model probability is low. Cross-event sequence and graph features are promising future extensions for attacks that cannot be recognized from one event.

C. Threshold Analysis

Threshold	Precision (%)	Recall (%)	Interpretation
0.4	90.8	98.7	aggressive; more analyst review
0.5	94.2	97.9	balanced security operation
0.6	96.0	97.0	selected simulation point
0.7	97.4	93.8	fewer alerts; higher miss risk

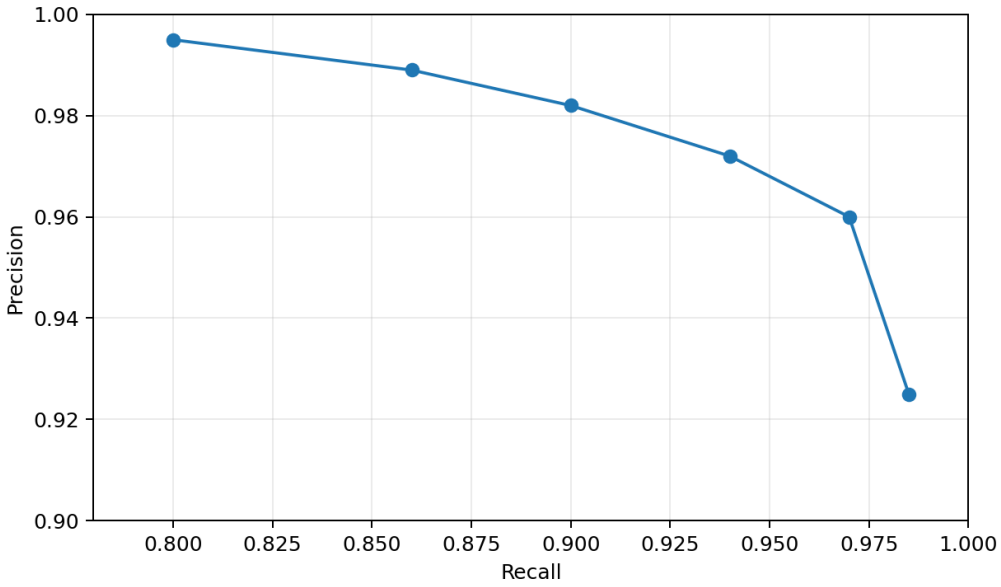

Fig. 4. Precision-recall trade-off across candidate operating points.

Threshold selection is a governance choice. Lower thresholds improve recall and increase false alerts; higher thresholds improve precision while permitting more attacks. Critical VMs may justify a lower threshold, whereas development systems may use a higher threshold with monitoring. Alert-volume forecasts and analyst capacity should be considered before activation.

D. Feature Influence and Ablation

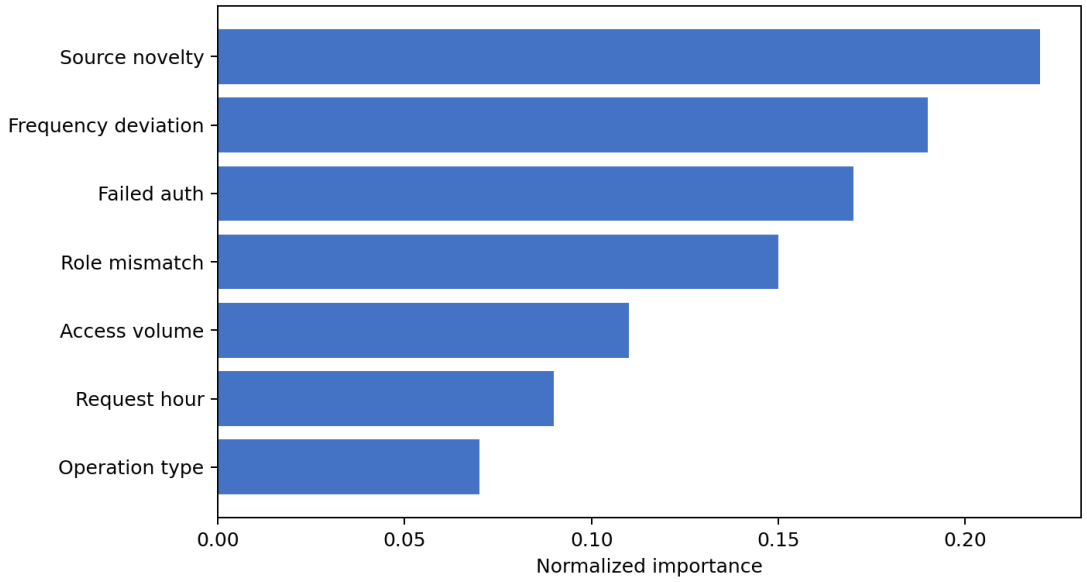

Fig. 5. Normalized feature importance for the selected ensemble.

Configuration	F1 (%)	High-risk coverage (%)	Mean latency (ms)
Policy rules only	83.6	100.0	806
ML only	96.5	94.1	744
VM selection + policy	89.9	100.0	938
Integrated framework	96.5	100.0	1010

Source novelty and frequency deviation are the strongest global influences, followed by failed authentication and role mismatch. Importance is not causal evidence and can be biased; it is used as a plausibility check. If an identifier or scenario-specific field dominated, the experiment would require leakage investigation.

The ML-only configuration has strong F1-score and low latency but fails to cover all policy-defined high-risk events. Policy rules alone provide complete mandatory coverage but weaker discrimination. The integrated framework preserves the selected model's F1-score and deterministic coverage at moderate processing cost. This result demonstrates that model quality and policy assurance are complementary metrics.

6. EXPLAINABILITY AND OPERATIONAL USE

Every response includes a structured explanation. A representative record may state: "restricted because VM sensitivity 0.76 exceeded identity suitability 0.54 and behavioural probability was 0.67." This is more actionable than a generic anomaly alert and enables an analyst to identify whether policy, identity context or model behaviour requires correction.

Global explanation is reviewed during model governance. Feature influence, calibration, error distribution and drift are monitored over time. Local explanation supports incident handling and appeal.

Neither explanation proves causality, and an analyst should verify source logs before attributing malicious intent.

Deployment should begin in shadow mode. Proposed decisions are compared with existing security outcomes without changing access. Step-up authentication is enabled before denial, and a rollback mechanism is rehearsed. Model artifacts, thresholds and feature definitions are versioned. Missing required telemetry activates a conservative review path rather than a normal prediction.

7. LIMITATIONS AND FUTURE WORK

The synthetic dataset reflects the authors' assumptions and cannot represent all production correlations. The random stratified split does not test future concept drift; a real study should use a time-based split. The classifier operates on engineered event summaries rather than raw sequences, limiting recognition of multi-stage attacks.

Future work should validate authorized provider telemetry, calibrate probabilities under local prevalence and measure alerts per workload-hour. Sequential and graph models can represent activity across identities and VMs. Federated learning may support cross-tenant training without exchanging raw events, although privacy, poisoning and aggregation risks require separate analysis.

Human factors also require evaluation. False-positive interruption, analyst trust, appeal outcomes and the quality of reason codes should be measured. Technical accuracy is only one component of safe deployment.

8. CONCLUSION

This paper presented an explainable framework for classifying cloud VM access events and combining behavioural probability with explicit resource, identity and policy evidence. Gradient boosting achieved the strongest performance in the controlled simulation, while ablation showed that ML alone does not guarantee mandatory high-risk coverage. The policy-first response preserves deterministic controls and records human-readable reasons. The findings support further validation but do not constitute a production security guarantee.

REFERENCES

1. M. Tavallaee, E. Bagheri, W. Lu and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in Proc. IEEE CISDA, pp. 1-6, 2009. doi: 10.1109/CISDA.2009.5356528.
2. N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in Proc. MilCIS, pp. 1-6, 2015. doi: 10.1109/MilCIS.2015.7348942.
3. I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. ICISSP, pp. 108-116, 2018. doi: 10.5220/0006639801080116.
4. R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. IEEE Symposium on Security and Privacy, pp. 305-316, 2010. doi: 10.1109/SP.2010.25.
5. S. Axelsson, "The base-rate fallacy and the difficulty of intrusion detection," ACM Transactions on Information and System Security, vol. 3, no. 3, pp. 186-205, 2000. doi: 10.1145/357830.357849.

6. L. Breiman, “Random forests,” Machine Learning, vol. 45, pp. 5-32, 2001. doi: 10.1023/A:1010933404324.
7. J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001. doi: 10.1214/aos/1013203451.
8. C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, pp. 273-297, 1995. doi: 10.1007/BF00994018.
9. F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825-2830, 2011. Official source: https://jmlr.org/papers/v12/pedregosa11a.html.
10. M. T. Ribeiro, S. Singh and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD, pp. 1135-1144, 2016. doi: 10.1145/2939672.2939778.
11. D. Sculley et al., “Hidden technical debt in machine learning systems,” Advances in Neural Information Processing Systems, vol. 28, 2015. Official source: https://proceedings.neurips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.
12. S. Rose, O. Borchert, S. Mitchell and S. Connelly, Zero Trust Architecture, NIST SP 800-207, Aug. 2020. doi: 10.6028/NIST.SP.800-207.
13. Joint Task Force, Security and Privacy Controls for Information Systems and Organizations, NIST SP 800-53 Rev. 5, Sept. 2020. doi: 10.6028/NIST.SP.800-53r5.
14. MITRE Corporation, “MITRE ATT&CK: A knowledge base of adversary tactics and techniques.” Official source: https://attack.mitre.org/ (accessed Aug. 8, 2026).
15. A. L. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153-1176, 2016. doi: 10.1109/COMST.2015.2494502.