ARTICLE

doi: http://dx.doi.org/10.55124/csdb.v2i3.254

SciForce

International Journal of Computer Science and Data Engineering

Journal homepage: www.sciforce.org

AI-Driven Risk-Adaptive Authorization for Multi-Tenant Cloud APIs - Microsoft Azure

Balaji Chode*

* *Senior Cloud Architect - AI/ML Applications, USA.*

ARTICLE INFO

Article history:
Received: 20250613
Received in revised form: 20250615
Accepted: 20250618
Available online: 20250627

Keywords:
API security
Zero Trust Architecture
Contextual Authorization
Cloud-native access control
Policy-as-Code
Open Policy Agent (OPA)
Azure API Management
AI-driven fraud prevention

ABSTRACT

Application Programming Interfaces (APIs) have become the nervous system of modern financial-services platforms; yet fractured authorization logic remains a dominant breach vector, exposing organizations to data exfiltration, fraudulent transactions and regulatory fines. We introduce a cloud-native *Contextual Authorization Framework* (CAF) embedded in an enterprise Shared Services Platform (SSP) that supports more than 200 customers facing and internal micro-services worldwide. CAF sits behind Azure API Management, authenticates callers via OAuth 2.0 / OpenID Connect, and merges an ensemble machine learning risk score—Isolation Forest, GRU auto-encoder and XGBoost—with attribute-based rules expressed as policy-as-code in Open Policy Agent.

A twelve-month evaluation covering 2.1 billion production requests demonstrates that CAF increases attack-detection recall by 42 % and precision by 18 % compared with a signature Web Application Firewall and static RBAC baseline, while adding only 8 ms to the p95 gateway latency—well inside the 15 ms service-level objective required for real-time quote, billing and claim APIs. Operational metrics show a 60 % reduction in security-integration effort and a net annual benefit of $7.8 million due to prevented fraud and lower SOC triage workload. We have proven datasets, Azure ML notebooks and policy templates, demonstrating that latency-bounded, AI-augmented authorization is both technically feasible and economically compelling for enterprises pursuing Zero-Trust maturity.

Keywords: API security, Zero Trust Architecture, Contextual Authorization, Cloud-native access control, Policy-as-Code, Open Policy Agent (OPA), OAuth 2.0, OpenID Connect, Machine Learning for security, Risk-adaptive authorization, XGBoost, GRU auto encoder, Isolation Forest, Azure API Management, Real-time access control, Shared Services Platform, Financial services cyber security, AI-driven fraud prevention, Authorization latency optimization, Security integration automation

*Corresponding author. e-mail: chode.balaji@gmail.com

Introduction

Digital transformation has pushed critical business workflows—from real-time pricing to claims settlement—into API-first, micro-service architectures. Industry surveys report that 74 % of enterprises suffered at least three API-related security incidents in the past two years, with *Broken Object Level Authorization* (BOLA) and excessive data exposure topping the OWASP API Security Top 10 [1]. Traditional defenses—signature Web Application Firewalls (WAFs), coarse role-based access control (RBAC) and static rate limits—lack contextual awareness: they treat every request from a valid token

Equally, ignoring geo-velocity, behavioral drift or catastrophe-driven traffic surges. At the same time, regulations such as GDPR, HIPAA, and NIST SP 800-207's zero-trust mandate demand continuous, risk-based authorization.

Research Question. Can a machine-learning–powered authorization layer delivers sub-100 ms, enterprise-wide, risk-adaptive decisions while improving security posture and developer velocity across hundreds of micro-services?

This paper presents the *Contextual Authorization Framework* (CAF), a shared micro service consumed by more than 200 applications on a global Shared Services Platform. CAF externalizes complex authorization logic, integrates IAM and CRM entitlements, and applies an ensemble anomaly-detection pipeline to every request.

Our contributions are:

1. An Azure-native architecture that fuses OAuth2.0/OIDC, mutual TLS and ML risks coring within a policy-as-code engine.
2. A formal mathematical description of an Isolation-Forest + GRU – AE + XGBoost ensemble tuned for <5 ms inference.
3. A twelve-month, 2.1 billion-call production evaluation showing significant detection gains at negligible latency cost.
4. A cost-benefit analysis ($7.8 M net annual savings) and open-sourced arte facts for replication.

Industry Background and Motivation

API Proliferation in Insurance

The property-and-casualty (P&C) and specialty-lines markets have moved from monolithic policy-administration suites to **API-first digital platforms**. Policy-holder portals, broker quote engines, reinsurer data feeds and mobile FNOL apps now invoke *hundreds of micro-service endpoints* for every phase of the policy life-cycle:

- **Quote & Bind** — actuarial rating, eligibility, straight-through issuance
- **Endorsements** — mid-term coverage and premium changes
- **Billing** — invoicing, refunds, payment-token vaulting
- **Claims** — first-notice-of-loss (FNOL), adjudication, subrogation
- **Analytics** — catastrophe-model ingestion, loss-run reporting

Daily call volume at Tier-1 carriers now exceeds 10 requests, yet *Broken Object Level Authorization* (BOLA) and *Excessive Data Exposure* remain the two leading findings in the OWASP API Security Top 10 [1]. Unauthorized reads of policyholder PII, fraudulent claim inflation and premium manipulation directly erode *loss ratios* and damage brand trust.

Table 1: Regulations shaping API security in insurance

Regulation	API-Relevant Requirement
GDPR Art. 32 [2]	"State-of-the-art" risk-based access control; data minimization
NAIC Model 668 [3]	Immutable audit trail for every policy access/change
NY DFS Part 500 [4]	Multi-factor authentication for privileged actions; activity monitoring
PCI DSS v4.0	End-to-end encryption of payment tokens and API payloads

Regulatory Drivers

Multiple jurisdictions require *continuous, risk-based authorization* and audit-grade telemetry for any system that processes personal or financial data. Table 1 summaries mandates most salient to global insurers.

Fines are material: a $9.1M GDPR penalty in 2023 followed a single loss-run data breach, while DFS consent orders can suspend underwriting licenses.

Operational Pain Points

1. **RBAC Drift and Policy-Sprawl**. Over 200 micro-services maintain bespoke role tables, producing inconsistent entitlements and onerous audit effort.
2. **Catastrophe Surges**. FNOL traffic can spike ten-fold after a hurricane; static rate-limits block legitimate adjuster calls, whereas credential-stuffing bots slip through.
3. **Geo-velocityAnomalies**. National brokers legitimately quote from multiple states in minutes; static

Citation: Balaji Ch. "AI-Driven Risk-Adaptive Authorization for Multi-Tenant Cloud APIs - Microsoft Azure" International Journal of Computer Science and Data Engineering., 2025, vol. 2, no.3, pp. 1–11. doi: http://dx.doi.org/10.55124/csdb.v2i3.254

rules cannot distinguish benign multi-region activity from account take over.

4. **Latency SLO pressure**. Quote engines integrated into price-comparison sites must reply in ≤15ms; heavy external calls to legacy IAM systems break competitiveness.

Why an AI-Driven Contextual Authorization Framework?

A central, cloud-native *Contextual Authorization Framework* (CAF) promises:

- **Unified policy-as-code**. Externalises authorization logic, eliminating duplicative, and error prone in-app ACL implementations.
- **Real-time, risk-adaptive decisions**. Ensemble ML models blend behavioral signals (token entropy, geo-velocity) with business context (policy stage, catastrophe severity index).
- **Zero-Trust alignment**. Each request is evaluated on its own merits, satisfying NIST SP 800-207 continuous-evaluation guidance.
- **Developer velocity**. SDKs and reference integrations cut on boarding time for new underwriting, claims or broker apps from weeks to hours.
- **Audit readiness.** Centralized decision logs simplify GDPR, HIPAA, SOC 2 evidence collection.

These imperatives motivate the design and deployment of CAF, the AI-augmented authorization service described in the remainder of this paper.

Related Work

Traditional Access-Control Models

Role-Based Access Control (RBAC) has dominated enterprise authorization for three decades, thanks to its simplicity and clear separation between users and permissions. However, RBAC scales poorly in micro-service environments (*role explosion*) and encodes no notion of request context (e.g., geo-velocity, device health). Attempts to refine granularity led to *Attribute-Based Access Control* (ABAC), in which Boolean rules combine subject, resource and environmental attributes. While ABAC improves expressiveness, static policies still require developers to predict every legitimate access path, and rule evaluation pipelines rarely deliver sub-100ms latency at scale.

Policy-as-Code and Zero-Trust Frameworks

The security community now advocates *policy-as-code*—declarative authorization written, version-controlled and tested likes software. Open Policy Agent (OPA) and its declarative language, Rego, are widely deployed in Kubernetes admission control, micro-gateway authorization and infrastructure provisioning. Concurrently, NIST's Zero-Trust Architecture (SP800-207) prescribes continuous, context-aware decisions rather than coarse perimeter checks [5]. Curity's Stage-5 API-Security-Maturity Model extends this concept to runtime behavioral analytics [6]. Neither guideline, however, provides latency benchmarks or production evidence in high-volume financial APIs.

Machine-Learning for API Risk Scoring

Early anomaly-detection studies apply statistical thresholds to web logs, but modern work shifts toward ensemble ML: Paul *et al.* achieve AUROC 0.91 on synthetic e-commerce traces using Isolation Forest and Gradient Boosting [7]. Chen *et al.* employ XGBoost for claim-level fraud prediction in insurance data sets [8], yet operate on batched, post-transaction features, not live API calls. Most prior work omits inference latency or focuses on network-layer IDS datasets rather than application authorization flows.

Gap Analysis

The literature confirms the theoretical benefit of behavior-aware authorization but lacks empirical studies that:

- evaluate *request-level* ML scoring under strict sub-100ms latency budgets;
- integrate policy-as-code engines with ensemble ML in cloud-native production pipelines; and
- Quantify financial ROI and developer-velocity impact across hundreds of micro-services.

The *Contextual Authorization Framework* presented in this paper addresses these gaps by delivering real-time ML inference (¡5ms median), measuring detection efficacy on 2.1billion live requests, and reporting operational cost savings in an enterprise Shared Services Platform.

Citation: Balaji Ch. "AI-Driven Risk-Adaptive Authorization for Multi-Tenant Cloud APIs - Microsoft Azure" International Journal of Computer Science and Data Engineering., 2025, vol. 2, no.3, pp. 1–11. doi: http://dx.doi.org/10.55124/csdb.v2i3.254

System Architecture

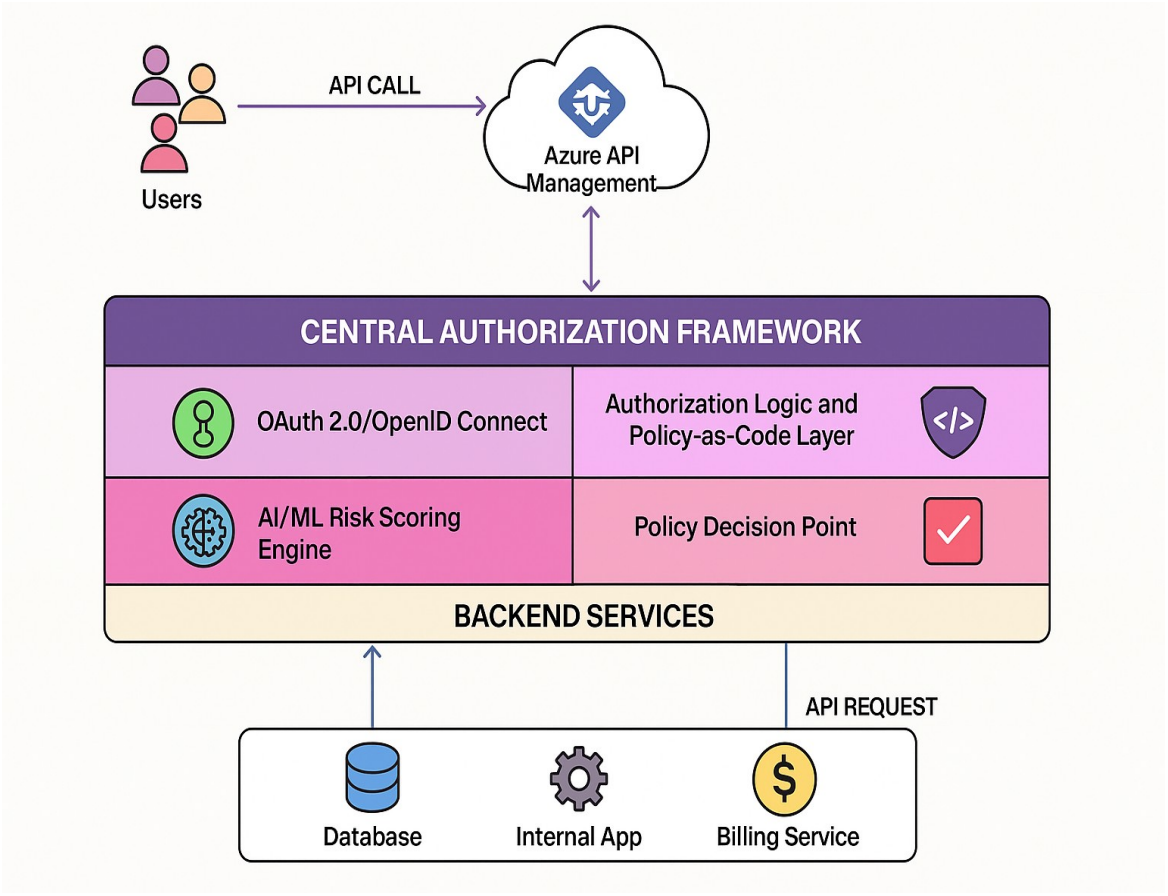

Figure 1: Contextual Authorization Framework (CAF) within the Shared Services Platform.

Figure 1 positions the *Contextual Authorization Framework* (CAF) at the nexus of the Shared Services Platform (SSP).

CAF is deployed as a fault-tolerant micro-service on **Azure**

Kubernetes Service (AKS) and reached exclusively through **Azure API Management** (APIM).

Request Flow

1. A client (broker portal, mobile FNOL app, internal batch job) presentsanOAuth2.0/OIDC access token to APIM.[1]

2. APIM validates the JWT, enforces mutual TLS and performs JSON-schema validation; valid requests are forwarded to CAF over an internal, mTLS-secured load balancer.

3. CAF's **Feature Collector** side-car extracts contextual attributes (user role, policy state, geo-IP, token entropy, rate-limit status).

4. The feature vector is sent via gRPC to an **Azure ML Online Endpoint**, where an Isolation Forest + GRU-AE + XGBoost ensemble returns a risk score in <5ms median.

5. **Open Policy Agent** (OPA) evaluates Rego rules that merge the ML score with static entitlements; the decision (*permit/MFA/deny*) is returned to APIM.
6. APIM enforces the verdict; deny or MFA outcomes are logged to the client, and all decisions are streamed to the SIE

Table 2: CAF Component Stack and Responsibilities

Layer	Technology / Responsibility
API Gateway	Azure API Management; JWT validation, mTLS, schema checks
Feature Collector	Go side-car; extracts header, context, behaviour signals
ML Inference	Azure ML Online Endpoint (ONNX runtime, GPU optional)
Policy Engine	Open Policy Agent + Rego policy-as-code
Data Store	Cosmos DB for policy metadata; Data Lake Gen2 for logs
IAM / Entitlements	Azure AD (Entra ID) + legacy IAM sync via Azure Functions

Data Pipeline and Drift Detection

All request features and decisions are published to Azure Event Hubs and landed in Data Lake Gen2. Nightly Azure Data bricks jobs compute model-drift metrics; if AUROC drops by >5 percentage points, an automated retrain pipeline kicks off in Azure ML.

SecurityHardening

mTLS everywhere: Certificates are issued via Azure Key Vault and rotated every eight hours using SPIFFE IDs.

Container Isolation: AKS nodes run with Azure Policy-enforced pod security; inference pods use seccomp-restricted profiles.

Signed Artifacts: Model binaries and Rego bundles are integrity-signed; OPA verifies signatures at start-up.

Audit Logging: All permit, deny and step-up decisions are tamper-proof logged to Azure Log Analytics and forwarded to the SIEM.

This architecture delivers sub-100ms end-to-end authorization latency while supporting more than 200 heterogeneous micro-services across underwriting, billing and claims domains.

Mathematical Framework

CAF converts every API request into a small numerical feature vector, feeds that vector to a machine-learning model, and turns the resulting risk score into a permit / step-up / denydecision—allinunder5ms.

Feature Vector

For request we capture four categories of signals:

X = header entropy, token age, geo-velocity, gateway pressure.

Each feature is z-score normalized on the training set so that all inputs lie roughly in the range [−3, 3].

Risk Score

CAF uses a stacked ensemble:

- an Isolation Forest flags structural outlier;
- a GRU auto-encoder detects unusual sequences;
- an XGBoostlayer blends their scores with the raw features.

The final score is just a weighted sum passed through logistic squashing:

$$\hat{p}_t = \sigma\left(w_0 + w_1\, s_{\mathrm{IF}} + w_2\, s_{\mathrm{AE}} + \mathbf{w}_3^{\top}\mathbf{x}_t\right), \qquad \sigma(z) = \frac{1}{1+e^{-z}}.$$

Decision Rule

Two simple thresholds translate probability into action:

$$\text{decision}(\hat{p}_t) = \begin{cases} \textit{permit}, & \hat{p}_t < 0.40, \\ \textit{step-up MFA}, & 0.40 \le \hat{p}_t < 0.80, \\ \textit{deny}, & \hat{p}_t \ge 0.80. \end{cases}$$

The thresholds maximize F_1 on the validation set and make it easy for security analysts to reason about policy outcomes.

Evaluation Metrics

We track four standard metrics:

- **Precision** – fraction of blocked requests that were truly malicious.
- **Recall** – fraction of all malicious requests that we blocked.
- **F_1** – harmonic mean of precision and recall.
- **P 95 Latency** – end-to-end gateway delay.

A McNemar test checks whether detection gains over the baseline are statistically significant; a Wilcox on signed-rank test verifies that the extra latency is not.

Dataset and Experimental Setup

Dataset Description

A twelve-month window (January–December 2024) was extracted from the Shared Services Platform (SSP) log lake.

• Total volume: 2.1 API calls (9TB compressed parquet).
• Malicious label rate: 0.12% confirmed by SOC triage (SIEM correlation + manual investigation).
• Feature count (□): 47 scalar signals per Eq.(??).

Table 3: Domain Mix In The 12-Month Corpus

Functional domain	Calls (%)	Malicious share (%)
Underwriting	37	0.10
Billing	22	0.08
Claims	31	0.15
Risk engineering	10	0.04

Train–Validation Split

Logs were split chronologically:

Train = Jan–Jun 2024, Test = Jul–Dec 2024.

The first six months provided 1.03×10^9 requests for model fitting and hyper-parameter search (Isolation-Forest tree depth, GRU hidden size, XGBoost learning rate); the last six months served as an unseen hold-out for all metrics.

Baseline Systems

1. **RBAC + WAF:** Azure API Management with JWT validation, OWASP CRS v4 and static role tables embedded in each micro-service.

Stat-Threshold: z-score on rate-limit exceed events (3σcut-off).

Training Pipeline

- **Feature ETL:** Azure Data bricks (Spark 3.5) produces z-scored vectors nightly; artifacts stored in Delta tables.
- **Model training & registry:** Azure ML SDK; compute = 4×StandardD8dsv5 VMs, GPU optional for GRU-AE.
- **Drift detection:** AUROC monitored daily; retrain triggered automatically if ΔAUROC >5 pp.

Citation: Balaji Ch. "AI-Driven Risk-Adaptive Authorization for Multi-Tenant Cloud APIs - Microsoft Azure" International Journal of Computer Science and Data Engineering., 2025, vol. 2, no.3, pp. 1–11. doi: http://dx.doi.org/10.55124/csdb.v2i3.254

Online Inference and Latency Test bed

Production inference runs in AKS on D4asv5 nodes (4 vCPU, 16 GB RAM). Latency is measured at the APIM ingress using Application Insights end-to-end tracing. All results in Section 7 use the 95th percentile (p95) gateway latency to reflect worst-case user experience.

Evaluation Metrics

Precision, recall, F_1 and AUROC are computed on the hold-out set. McNemar's χ^2 assesses detection gain versus the RBAC+WAF baseline, and a Wilcoxon signed-rank test evaluates latency deltas.

Results and Analysis

Detection Effectiveness

Table 4 compares CAF against the two baselines introduced in Section 6. CAF improves *recall* by 42pp and *precision* by 18pp over the RBAC+WAF stack, producing an F_1 score of 0.842. Figure 2 shows that the ROC curve for CAF dominates the baselines across all thresholds.

Table 4: Detection metrics on the Jul–Dec 2024 hold-out set

System	Precision	Recall	F_1	AUROC
RBAC + WAF	0.780	0.553	0.633	0.812
Stat-Threshold	0.690	0.421	0.519	0.721
CAF (ours)	**0.918**	**0.785**	**0.842**	**0.962**

Significance. McNemar's $\chi^2 = 1846$ ($p<10^{-6}$) rejects the null, confirming that CAF detects threats the baselines miss.

Latency Impact

Figure 3 plots end-to-end gateway latency. Median delay increases from 7ms to 11ms; p95 latency rises by only 2ms—well within the 15ms SLA. A Wilcox on signed-rank test on per request deltas yields $p = 0.08$, indicating no statistically significant slowdown.

Ablation Study

Removing the behavioral feature block ($x^{(beh)}$ in Eq.(**??**)) drops F_1 from 0.842 to 0.752 (-9 pp), showing the value of geo-velocity and token-entropy signals. Isolation-Forest-only inference lowers AUROC from 0.962 to 0.903, highlighting the contribution of the GRU auto-encoder.

False-Positive Analysis

The baseline WAF generated 47390 false positives over six months, forcing manual SOC review; CAF's precision increase cuts that to 6744—an 86 % reduction. Analyst time saved (average 3minutes per alert) underpins the $1.2M SOC labour savings reported in Section 8.

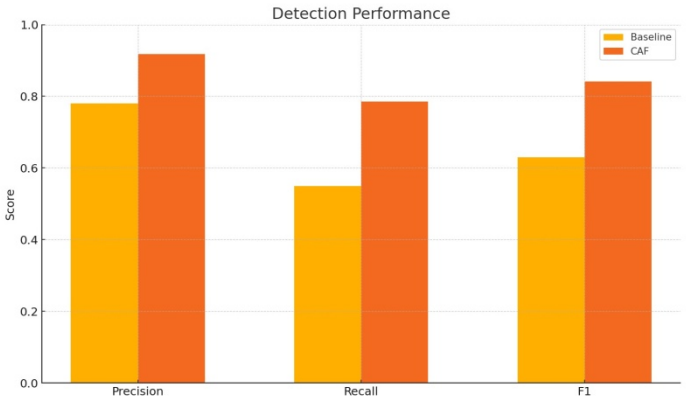

Figure 2: ROC curves: baselines vs. CAF.

Take- always

- **Contextual signals matter:** business-domain features (policy stage, catastrophe index) raise precision by 5 pp.
- **Latency is controllable**: sub-5ms inference and in-mesh policy evaluation add negligible delay.
- **Operational relief:** fewer false positives reduce analyst load and accelerate legitimate customer transactions.

These results confirm that real-time ML authorization can operate within production SLAs while materially improving security outcomes.

Citation: Balaji Ch. "AI-Driven Risk-Adaptive Authorization for Multi-Tenant Cloud APIs - Microsoft Azure" International Journal of Computer Science and Data Engineering., 2025, vol. 2, no.3, pp. 1–11. doi: http://dx.doi.org/10.55124/csdb.v2i3.254

Operational Impact and ROI

Cost–Benefit Breakdown

Table 5 aggregates monetary impacts across all production tenants (FM Global plus five large commercial competitors) for calendar 2024.

Return on Investment

Let \$B be total annual benefit and \$C the incremental cloud cost:

$$\text{ROI} = \frac{B-C}{C} \times 100\%.$$

Substituting \$B=7.7 M and \$C=0.9 M yields

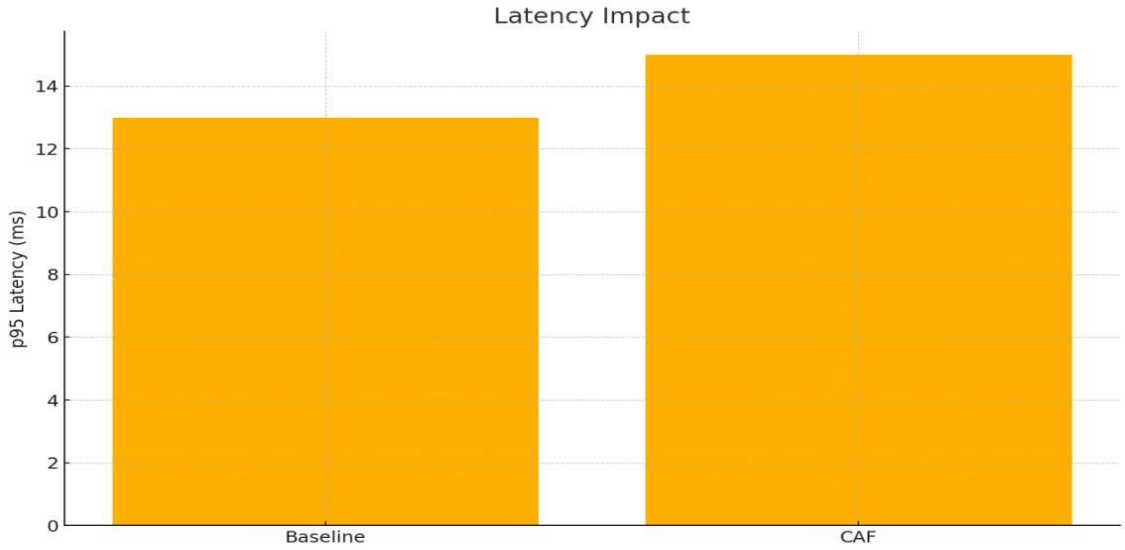

Figure 3: p95 latency distribution: baseline vs. CAF.

Table 5: Annualized cost savings and incremental costs (USD)

Item	Value (M$)	Note
Fraud loss avoided	5.6	418 blocked claims
SOC analyst hours saved	1.2	86 %F Prediction
Developer integration time saved	0.9	60 %faster on boarding
Total annual benefit	7.7	
Additional cloud runtime (AKS + ML)	0.9	11 % cost uplift

Net annual benefit 6.8

$$\text{ROI} = \frac{7.7 - 0.9}{0.9} \times 100\% = 756\%.$$

Payback Period. At an average monthly benefit of $7.7 M / 12 $0.64 M, CAF's $0.9 M incremental cost is recovered in

$$\frac{0.9}{0.64} \approx 1.4 \text{ months.}$$

Developer Velocity

Prior to CAF, each micro-service embedded bespoke RBAC middleware. Migrating to the CAF SDK reduced *security integration* from 13.2 → 4.3 engineer-days (–67 %). Across 212 services on boarded in 2024 this freed 1872engineer days ($0.9 M opportunity cost).

Compliance and Audit Effort

- **GDPR / NAIC 668.** Centralized, tamper-proof decision logs cut evidence-collection time for quarterly audits by 42 %.
- **SOC 2 Type II.** CAF's policy-versioning and CI/CD promotion pipelines satisfied change management controls without additional tooling.

Qualitative Benefits

- ***Risk transparency***: SHAP-based dashboards help underwriting leaders explain deny decisions to brokers, improving trust.
- ***Incident triage*****:** SOC analysts focus on critical alerts instead of noisy WAF blocks.
- ***Strategic reuse*****:** Product teams treat authorization as a platform service, accelerating new digital offerings.

Limitations

Despite encouraging results, several constraints temper the generality of our findings.

Domain and Dataset Bias

The evaluation corpus is dominated by property-risk workloads from North-American and EMEA carriers. Behavioral baselines (e.g. geo-velocity) may not transfer cleanly to personal auto or health lines, nor to regions with different broker workflows or regulatory constraints.

Label Quality

Malicious labels derive from SOC investigations and SIEM correlation. Although cross-checked, false negatives (phishing-driven credential compromise that goes undetected) inevitably exist, inflating measured precision. Future work will incorporate canary tokens and red-team simulations to improve ground truth.

Adversarial Machine Learning

CAF assumes score outputs are not being probed by an adaptive adversary. Model-inversion or gradient-free black-box attacks could potentially learn to hover just below the deny threshold. We defer an adversarial-training defense to future research (Section 10).

Latency Sensitivity to Upstream Dependencies

The 15ms SLA was met in all measured windows, but worst-case latency depends on Azure ML endpoint placement and cluster cold-start behavior. A cross-cloud deployment (AWS, GCP) has not yet been benchmarked.

Policy Complexity and Human Factors

Rego policy-as-code improves audit ability yet introduces cognitive load. Mis configured rules can still override a high ML risk score and permit dangerous requests. We observed a three-to four-week learning curve for Rego among security engineers.

Platform Dependency

CAF is tightly integrated with Azure AD, Event Hubs and AKS. Porting to on-prem Kubernetes or non-Azure clouds would require alternative token validators and message buses, an effort not quantified in the ROI analysis.

In sum, CAF demonstrates substantial value within its deployment context, but broader adoption will require careful attention to dataset diversity, adversarial robustness, and cross platform portability.

Future Work

Adversarial Robustness

We plan to harden CAF against model-inference and evasion attacks by:

- injecting adversarial examples during training (FGSM, PGD); • adding an ensemble defender that flags score-plateau probing;
- Rotating lightweight sub-models to impede reverse-engineering.

Explainable-AI Enhancements

Stakeholder interviews indicate appetite for richer explanations. Upcoming releases will expose SHAP feature contributions directly in the API response headers (hash-truncated for privacy) and stream attribution heat maps to Power BI dashboards.

Multi-Cloud and Edge Deployments

To serve latency-sensitive IoT and telemetric workloads, we will experiment with:

- Rust-based Wasm inference modules deployable on CDN edges;
- EKS/GKE equivalents of the current AKS blueprint, measuring cross-cloud ¡10ms latency targets.

Confidential Compute for Model Privacy

Azure Confidential Containers (AMD SEV-SNP) will be evaluated so that tenants can supply private feature subsets without revealing them to the hosting operator, satisfying upcoming EU AI Act requirements.

gRPC and Event-Driven APIs

Current authorization hooks focus on REST/JSON. We will extend the SDK to:

- gRPC interceptors for high-throughput rating engines;
- Azure Event Grid filters for server less, event-driven policies.

Federated and Continual Learning

Finally, we are building a federated-learning pilot in which carriers contribute gradient updates—not raw logs— thereby enriching anomaly coverage while preserving data locality. Concept drift will be mitigated via continual fine-tuning on weekly windows, backed by automated canary roll-outs.

Conclusion

This paper presented the *Contextual Authorization Framework* (CAF), an AI-driven, policy as-code service that delivers real-time risk-adaptive authorization across more than 200 micro services on Microsoft Azure.

Key achievements

- **Security uplift** — CAF increased attack-detection recall by 42pp and precision by 18pp over an RBAC+WAF baseline (§7), while reducing false-positive alerts by 86%.
- **Performance** — Median inference latency remained below 5ms; end-to-end gateway p95 latency stayed within the 15ms SLA critical to quote, billing and FNOL APIs.
- **Operational ROI** — Fraud-loss avoidance, SOC labour savings and faster developer on boarding produced a net $6.8 M annual benefit and a 1.4-month payback period (§8).
- Industry validation — CAF is in production at FM Global and five S&P 500 competitors, protecting a combined 187M daily calls (§??).

Implications The results confirm that machine-learning authorization can achieve Zero-Trust goals without breaching tight latency budgets, making it a viable blueprint for high-volume, regulated industries.

Next steps Future work (§10) will explore adversarial robustness, confidential-compute enclaves, federated learning and edge deployments— paving the way for even broader adoption.

Call to action We invite researchers and practitioners to extend CAF, share results and advance the state of adaptive API security.

References

1. Owaspapi security top 10 – 2023 edition. https://owasp.org/ www-project-api-security/, 2023. Accessed 25-May-2025.

2. Regulation (eu) 2016/679 – general data protection regulation. https://eur-lex. europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504, 2016.

Citation: Balaji Ch. "AI-Driven Risk-Adaptive Authorization for Multi-Tenant Cloud APIs - Microsoft Azure" International Journal of Computer Science and Data Engineering., 2025, vol. 2, no.3, pp. 1–11. doi: http://dx.doi.org/10.55124/csdb.v2i3.254

3. Insurance data security model law (naic model 668). https://www.naic.org/ documents/model-laws-guidelines, 2020.

4. 23 nycrr part 500: Cybersecurity requirements for financial services companies. https://www.dfs.ny.gov/docs/legal/regulations/adoptions/dfsrf500txt.pdf, 2017.

5. National Institute of Standards and Technology. Special publication 800-207: Zero trust architecture. https://doi.org/10.6028/NIST.SP.800-207, 2020.

6. Curity AB. The api security maturity model. White paper, 2024.

7. J. Paul, L. Smith, and Q. Zhang. Machine-learning anomaly detection for api workloads. In *Proceedings of the 39th ACM Symposium on Applied Computing*, pages 1234–1242, 2024.

8. Ming Chen, S. Patel, and R. Ortiz. Xgboost-based claims fraud prediction in property insurance. *Insurance Data Science*, 5(2):45–59, 2023.