

AI-Driven Alert Triage in Security Operations Centers

UC3M



Marcelino Lázaro

Francisco-J. González-Serrano

UFV



Lorena Álvarez-Pérez

GMV



Aitor Gutiérrez-López

J. L. Álvarez-Aldana

Marta Gil-López

Andrés Izquierdo-Núñez

Jairo Montero-Santos

AIAI 2026

IFIP International Conference on Artificial Intelligence Applications & Innovations
Chania, Crete - 16-19 July 2026



Funding & Acknowledgments: This work is part of a project financed by “European Union [NextGeneration-EU](#), the [Recovery Plan, Transformation and Resilience](#), through [INCIBE](#)” under the Strategic Innovative Public Procurement Initiative (IECPI). The results correspond to a prototype in a pre-commercial stage. This work has also been partially funded by PID2021-125652OB-I00 (ML-SAM), funded by MCIN/AEI/10.13039/501100011033 and co-financed by ERDF “A way of making Europe”.

A SOC monitors, detects, and responds to cybersecurity threats for multiple clients. GMV-CERT has a two Level analyst structure: L1 triage, L2 deep analysis

1. CYBERSECURITY ALERTS

The SOC receives a high volume of alerts from multiple security sources detecting different types of incidents.



Malware

e.g., virus, ransomware, trojans



Intrusion & Attack

e.g., brute force, exploitation, lateral movement



Phishing & Social Engineering

e.g., phishing emails, credential theft



Network Anomalies

e.g., C2 traffic, scanning, DDoS



Policy / Data / User Risks

e.g., data exfiltration, privilege misuse, risky behavior



High volume, noisy, and imbalanced stream (many benign vs. few critical)

2. GMV-CERT: SERVICE TO MULTIPLE CLIENTS

The SOC monitors and responds to a large number of clients (projects) across different industries and geographies.



Client / Project 1



Client / Project 2

...



Client / Project N



Each client uses its own security tools and nomenclature. Alerts and information are not standardized.

3. TWO-LEVEL ANALYST STRUCTURE

The SOC operates with two levels of analysts working together to monitor, analyze and respond to alerts.



L1 ANALYSTS (8x5)

First line analysts
Initial triage and analysis



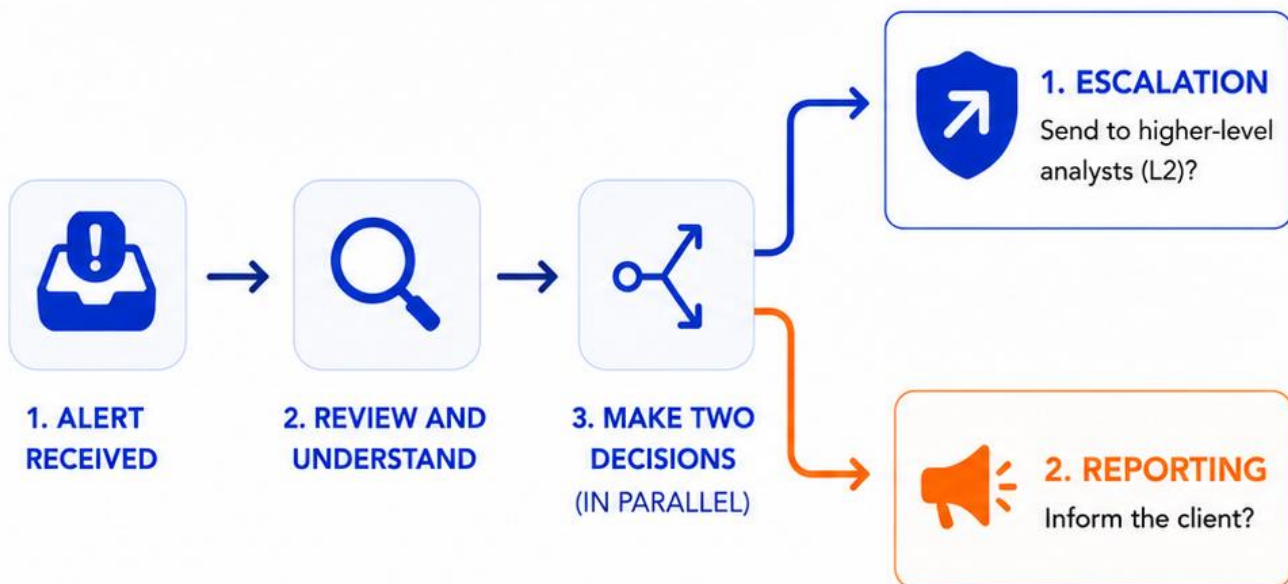
L2 ANALYSTS (24x7)

Advanced analysts
Deeper investigation and response



L1 ANALYSTS

First line triage



1. ESCALATION

Send to higher-level analysts (L2)?

Decision Error	Business impact
 Miss (not escalate)	High impact A complex alert is not escalated to L2.
 False escalate	Lower impact An alert is escalated unnecessarily.



2. REPORTING

Inform the client?

Decision Error	Business impact
 False report	High impact An unnecessary report is sent to the client.
 Miss (not report)	Lower impact A relevant alert is not reported to the client.



DATASET (GMV-CERT)



170,653
raw SOC alerts



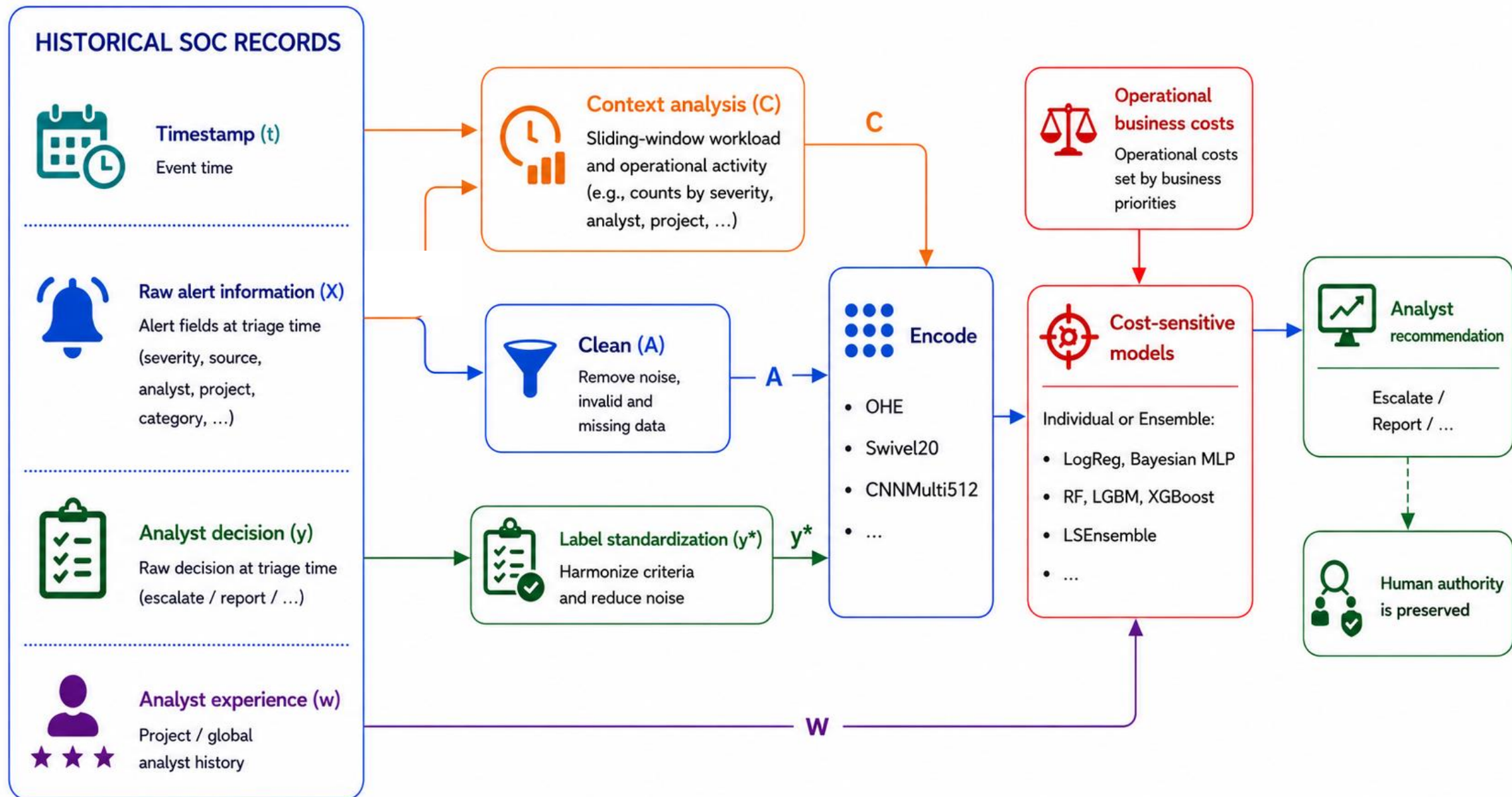
25
monitored
projects



27
real-time
features



MIXED TYPES
categorical, numerical,
textual and binary
information



RAW ALERT RECORD

(examples)

-  Title / Description with prefixes
(ABC_XX0001) Windows events
– Dangerous command execution
-  Inconsistent categories
United States / united states /
USA / U.S.
-  Synonyms and variations
blocked / dropped / DENIED /
auditoria correcta / rooté
-  Missing values
NaN / None / – / (blank)
-  Encoding issues
Ã© / Ã± / Ã¡ / Ä· / Ä³
-  Irrelevant alerts
Mailbox ... / CERT-EU ...
-  Business rules (out of scope)
Events before project start date
Dropped projects ...

CLEAN (preprocessing)

Make raw alerts consistent, reliable and useful

-  **1. Clean text**
Remove prefixes and extra
spaces, standardize formatting.
-  **2. Normalize values**
Map synonyms and variants to a
canonical representation.
-  **3. Handle missing values**
Fill NaN / None / blanks with
standard placeholders.
-  **4. Fix encoding issues**
Detect character encoding
(e.g., UTF-8) and fix.
-  **5. Filter data & apply rules**
Remove irrelevant alerts and
apply business rules.

CLEANED ALERT RECORD

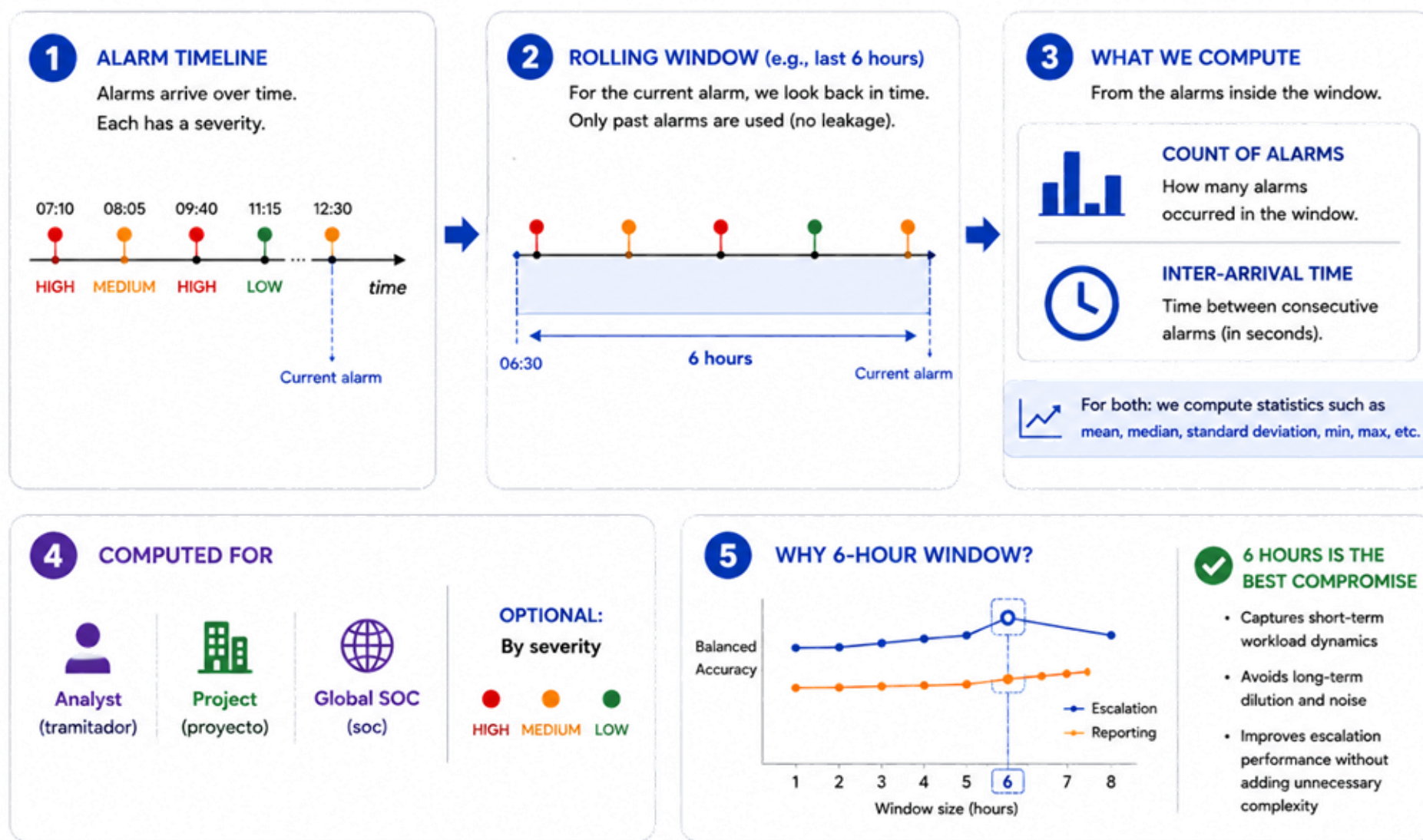
(examples)

-  Title / Description
Windows events –
Dangerous command execution
-  Country (normalized)
United States
-  Action (standardized)
audit success / deny / block
-  Missing values handled
(standard placeholders)
-  Correct encoding
á / é / í / ó / ú / ñ
-  Irrelevant alerts removed
(e.g., Mailbox, CERT-EU)
-  Business rules applied
Only events within scope
kept

- Identical observable alert patterns can appear with different labels.
 - Analysts may decide using additional operational information not present in the alerts available to the model.
 - Therefore, this is not necessarily an SOC operational inconsistency, but a data-level ambiguity from the model's point of view.
- Label standardization** reduces this ambiguity, making the learning target more consistent while preserving all the original samples.



- Motivation:
 - During validation with GMV-CERT, we observed that **escalation** decisions depend not only on alert content, but also on the recent operational workload around each alert.
- Contribution:
 - We extend alert-only modelling (A) with operational context (C), derived from recent alert activity at the analyst, project and SOC levels, and optionally segmented by severity.
- Window selection:
 - A **6-hour** rolling window provided the best trade-off between short-term workload information and excessive temporal dilution.
- Result:
 - Context improves **escalation**, where workload and time pressure matter, but has little effect on **reporting**.



Encoded form (example)



1 MODELS EVALUATED

We evaluate individual and ensemble-based classifiers.

INDIVIDUAL CLASSIFIERS



Logistic Regression

Linear probabilistic model.



Bayesian Binary MLP

Optimizes Bayes cost via Parzen output-density estimation.

ENSEMBLE-BASED METHODS



Random Forest

Bagging-based tree ensemble: trains many trees in parallel and combines their votes.



LightGBM

Fast gradient boosting trees: builds trees sequentially to correct previous errors.



XGBoost

Regularized gradient boosting trees: adds regularization to control complexity and improve robustness.



LSEnsemble

Class switching + Bayesian threshold correction to control the operating point.

2 COST-SENSITIVE LEARNING

Training incorporates operational costs and analyst experience.



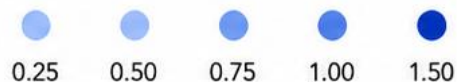
Class weights

Set from operational business costs. Reflect the relative cost of false alarms and misses.



Sample weights

Analyst range experience



Bayes cost

$$C_B = \bar{c}_{FA} P_{FA} + \bar{c}_M P_M$$

false alarms + misses, weighted by cost

P_{FA} = probability of false alarm

P_M = probability of miss

$\bar{c}_{FA}, \bar{c}_M$ = normalized costs

$$(\bar{c}_{FA} + \bar{c}_M = 1)$$



Bayes Benefit

$$\bar{B}_B = 1 - C_B$$

higher is better

3 EXPERIMENTAL SETUP

DATASET (GMV-CERT)



Real-world SOC alerts
16 months of data

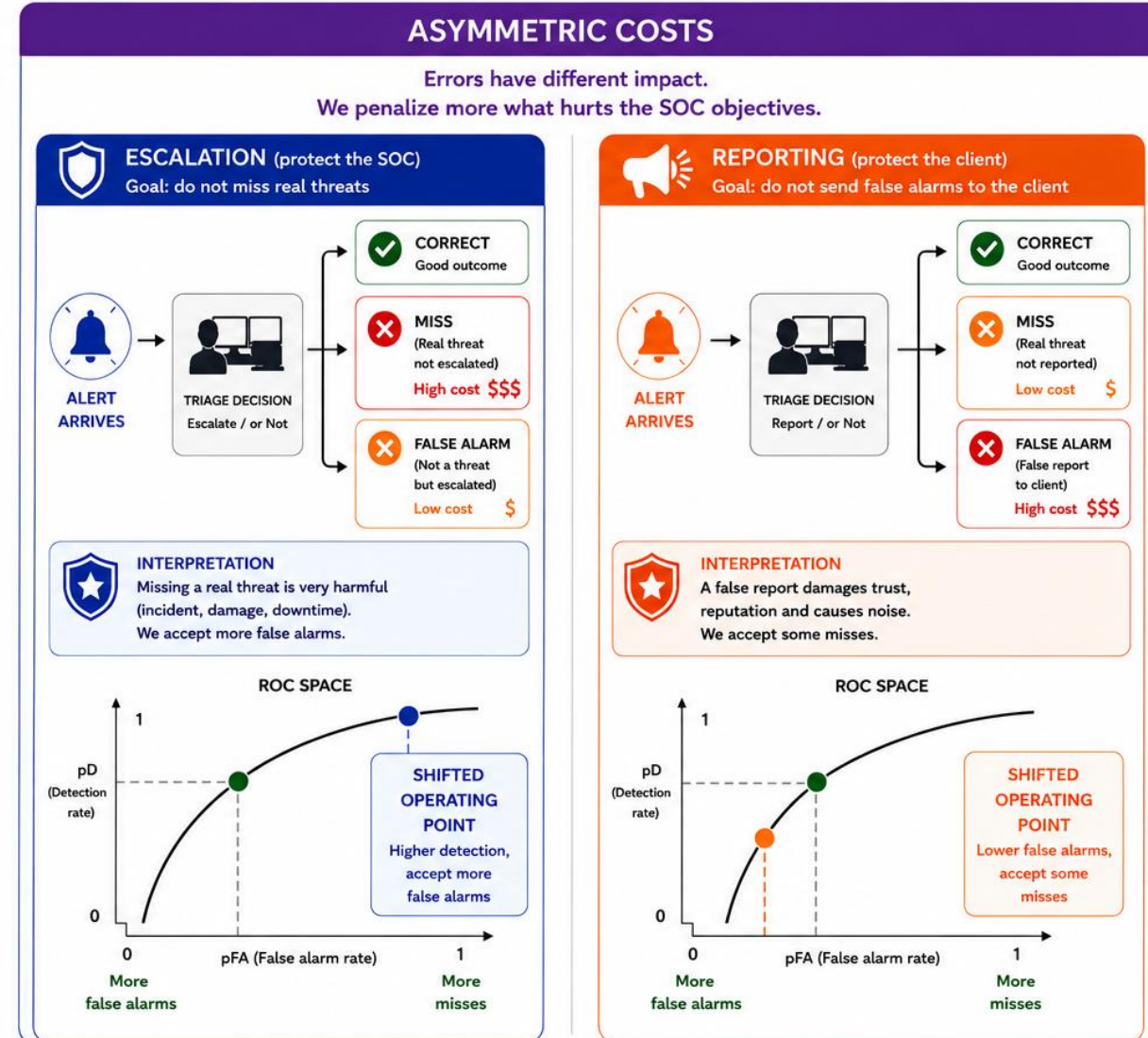
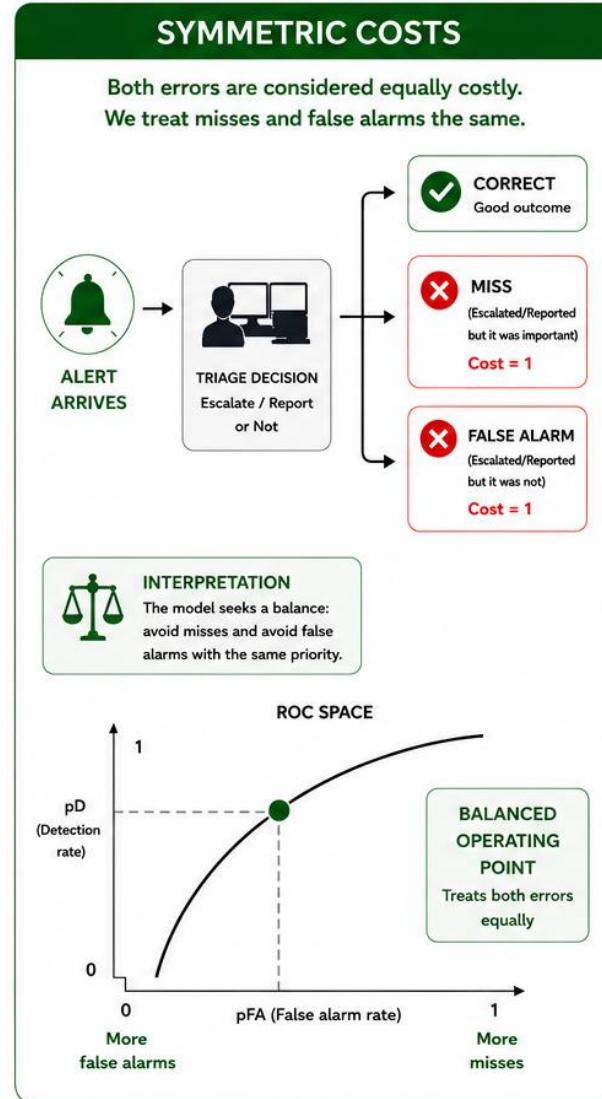


HYPERPARAMETER VALIDATION



Grid search validation on the last 20% of the training period using Bayes Benefit as metric.

- With symmetric costs, misses and false alarms are penalized equally.
- With asymmetric costs, the penalty reflects the business objective:
 - Escalation:** missing a real incident is more costly
 - prioritize detection (e.g. > 95 %).
 - Reporting:** sending a false report to the client is more costly
 - prioritize precision and reduce false alarms (e.g. < 5 %).



- **Escalation** benefits from context:
 - the best results use intrinsic + context features.
- **Reporting** is mainly driven by intrinsic alert features:
 - context adds little value.
- Asymmetric costs change the operating point toward the SOC needs,
 - especially by increasing detection when misses are costly.
- LSEnsemble and LightGBM are the strongest models overall.

1: RESULTS WITH SYMMETRIC COSTS (Equal misclassification costs)

Standardized labels • Metric: Bayes Benefit (BB) (In this case, BB = Balanced Accuracy)

 ESCALATION TASK Should the alert require higher-level analyst intervention?				
Feature set	Best model	BB _A (↑)	P _{FA} (↓)	P _D (↑)
 Intrinsic features	LSEnsemble	0.8751	0.0919	0.8421
 Context features	LSEnsemble	0.7509	0.3404	0.8421
 Combined features	LightGBM	0.8966 ★	0.0450	0.8421
 KEY TAKEAWAY (Escalation): Context helps escalation. Best overall result: LightGBM with A+C. LSEnsemble is strongest with A alone and C alone.				

 REPORTING TASK Should the alert be reported to the client?				
Feature set	Best model	BB _A (↑)	P _{FA} (↓)	P _D (↑)
 Intrinsic features	LightGBM	0.8751 ★	0.0777	0.8279
 Context features	LightGBM	0.5212	0.2049	0.2473
 Combined features	LightGBM	0.8705	0.0780	0.8190
 KEY TAKEAWAY (Reporting): Intrinsic alert features carry almost all the information. Context alone is weak; adding context does not improve over A alone.				

2: RESULTS WITH ASYMMETRIC COSTS (Different operating points for SOC needs)

Metric: Bayes Benefit (BB)

(Higher BB is better)

 ESCALATION TASK Cost setup: miss cost more (misses more serious) Metric: Bayes Benefit (BB)				
Feature set	Best model	BB (↑)	P _{FA} (↓)	P _D (↑)
 Intrinsic features	LSEnsemble	0.9063	0.2695	0.9649
 Context features	LightGBM	0.8819	0.4723	1.0000
 Combined features	LightGBM	0.9162 ★	0.3351	1.0000
 KEY TAKEAWAY (Escalation): Asymmetric costs push the operating point toward high detection (P _D). Best overall: LightGBM with A+C. LSEnsemble remains the best option with A alone.				

 REPORTING TASK Cost setup: false alarm cost more (false reports more harmful) Metric: Bayes Benefit (BB)				
Feature set	Best model	BB (↑)	P _{FA} (↓)	P _D (↑)
 Intrinsic features	LSEnsemble	0.8934 ★	0.0476	0.7754
 Context features	XGBoost	0.6731	0.0035	0.0263
 Combined features	LSEnsemble	0.8919	0.0381	0.7517
 KEY TAKEAWAY (Reporting): Asymmetric costs reduce false alarms (P _{FA}). Best overall: LSEnsemble with A. A+C gives a very similar result; context alone is weak.				



Best score for that task/cost block



Intrinsic features (Alert content)



Context features (Analyst workload, time, etc.)



Combined features (A + C together)

BB: Bayes Benefit (Higher is better)

P_{FA}: Lower is better

P_D: Higher is better

SOC alert triage involves two related decisions: **escalate** to L2, **report** to the client, **both**, or **neither**.

- **Escalation** benefits from operational context (**A+C**).
 - Analyst workload and alert pressure affect escalation decisions.
- **Reporting** is mainly driven by intrinsic alert information (**A**).
 - Contextual features alone are weak, and adding context does not improve the best models.
- **Label standardization** improves consistency.
 - It reduces contradictory labels for repeated alert patterns while preserving the original samples.
- **Cost-sensitive learning** moves the operating point according to SOC priorities:
 - higher detection for escalation, fewer false reports for client reporting.
- Overall, **LightGBM** provides the strongest absolute performance in most scenarios, while **LSEnsemble** achieves the best operating point for reporting under asymmetric costs.

Thank you

fran@ing.uc3m.es