

MODULO 08 • AGOSTO 2026

AI Applicata I: Triage e Classificazione

Quando i reperti si misurano in terabyte:
dove l'automazione riduce il backlog senza sostituire il giudizio

triage • hash percettivi • anomaly detection • human-in-the-loop

Corso di Laurea Magistrale in Sicurezza Informatica

Al termine del modulo lo studente sarà in grado di

01 • PROGETTARE

Una pipeline di triage AI-assistita: prioritizzazione dei reperti, soglie operative, punti di controllo umano

02 • APPLICARE

Classificazione multimediale e hash percettivi su larga scala, comprendendone finalità e limiti

03 • VALIDARE

Gli esiti automatici con replica manuale su campione, documentando il protocollo di verifica in relazione

Due benefici distinti del triage automatico. Riduce il tempo di risposta su volumi altrimenti ingestibili; riduce inoltre l'esposizione psicologica degli operatori ai contenuti più gravi, a tutela della loro salute.

La catena del triage assistito dall'AI



Le quattro aree si susseguono nelle slide del modulo nello stesso ordine dello schema: ciascuna presuppone la precedente.

Ordinare l'enorme prima di esaminare il rilevante

Il triage classifica i reperti per probabile rilevanza **prima** dell'esame umano: filtri su hash noti (NSRL per il noto-benigno), classificatori di contenuto, prioritizzazione per tipologia e contesto. L'analista inizia dai reperti più probabilmente rilevanti.

Nel contrasto ai contenuti illeciti, piattaforme come **Griffeye Analyze** e **Magnet.AI** combinano classificatori visuali e **hash percettivi** (PhotoDNA, PDQ), che riconoscono un'immagine nota anche dopo ridimensionamenti e ricodifiche.

Hash crittografico vs percettivo

Crittografico (SHA-256): identità esatta del file. Basta un bit modificato per cambiare l'hash.

Percettivo (PhotoDNA, PDQ): somiglianza visiva. Robusto alle trasformazioni ma probabilistico: ammette collisioni e ha valore di indizio.

Il triage stabilisce l'ordine di esame e lascia invariato il perimetro della prova: ciò che il modello declassa resta nel fascicolo ed è esaminabile; l'esclusione automatica di reperti sarebbe una delega inammissibile.

Known-good e known-bad: il triage prima dell'AI

Il filtro più antico ed efficace del triage è deterministico: il confronto degli hash dei file con liste di riferimento.

- **Known-good** (es. **NSRL** del NIST): milioni di hash di file di sistema e software commerciale noto; tutto ciò che combacia è rumore di fondo, escludibile dall'esame di merito
- **Known-bad**: hash di malware e contenuti illeciti noti; ciò che combacia va in cima alla coda
- Il residuo, né buono né cattivo noto, è il perimetro su cui lavorano analista e classificatori

ESEMPIO • L'EFFETTO IMBUTO

File sull'immagine: 412.000

- match NSRL: 371.000

- duplicati interni: 14.500

= da esaminare: 26.500

Oltre il 93% del volume eliminato senza alcun modello e senza alcun rischio interpretativo: l'hash match è esatto e ripetibile, quindi facile da difendere in aula.

Ordine corretto della pipeline: prima i filtri deterministici (hash, firme di formato), poi i classificatori sul residuo. I classificatori hanno costi e tassi d'errore, quindi si applicano solo dove i filtri deterministici non bastano.

Riconoscere il tipo reale dei file

Rinominare un file è il più elementare degli occultamenti: un archivio cifrato diventa «vacanze.jpg». Il primo livello di difesa è deterministico: i **magic bytes** (firma d'intestazione: FF D8 per JPEG, 50 4B per ZIP) confrontati con l'estensione dichiarata. La **discrepanza** è di per sé un segnale di triage.

Il secondo livello è statistico: **entropia** e distribuzione dei byte distinguono testo, immagini compresse, archivi e contenuto cifrato anche su **frammenti privi di intestazione** (settori recuperati dal carving, slack space, flussi parziali). Qui i classificatori ML superano le firme: apprendono il «profilo statistico» di ogni formato e lo riconoscono in frammenti di pochi KB.

Limite da dichiarare: cifrato e compresso tendono all'entropia massima; per distinguerli servono test di uniformità e, in relazione, prudenza lessicale.

ESEMPIO • L'ARCHIVIO TRAVESTITO

Il triage segnala 14 file con estensione .jpg la cui firma è 50 4B (ZIP) e l'entropia compatibile con contenuto cifrato interno.

Sono archivi protetti rinominati, raccolti in una cartella di sistema. La segnalazione automatica richiede pochi millisecondi; spetta poi all'analista decidere come procedere, anche con gli strumenti giuridici previsti per ottenere le chiavi.

PhotoDNA e PDQ: riconoscere immagini note

L'hash crittografico identifica i byte: basta una ricompressione e il match sparisce. L'hash percettivo codifica invece **caratteristiche visive** dell'immagine (distribuzione di luminanza su una griglia, nel caso di PDQ), producendo impronte che restano **vicine** per immagini visivamente simili.

- Il confronto restituisce una **distanza**: sotto soglia, probabile stessa immagine
- Robusto a ridimensionamento, ricompressione, lievi ritagli
- Fragile a rotazioni forti, specchiature, alterazioni mirate

ESEMPIO • LO STESSO SCATTO, TRE FILE

Una foto nota circola come originale, come copia ricompressa dal messaging e come miniatura ridotta: tre SHA-256 completamente diversi, tre impronte PDQ a distanza minima tra loro.

Il filtro crittografico avrebbe visto tre file ignoti; quello percettivo li aggancia tutti e tre alla stessa immagine di riferimento.

Statuto probatorio: il match percettivo è probabilistico e ammette collisioni. In relazione vale come **indizio di somiglianza** che prioritizza la verifica visiva umana; l'identità del contenuto la stabilisce l'esame del reperto.

Il costo umano della revisione

Il problema

Fra gli analisti esposti ogni giorno a materiale traumatico la letteratura documenta **stress traumatico secondario e burnout**. La loro tutela è un dovere dell'organizzazione.

Il contributo dell'AI

Hash percettivi e classificatori identificano il **materiale già noto** senza visione umana e prioritizzano il nuovo, mostrato con anteprime sfocate. L'operatore esamina così meno contenuti, in dosi e condizioni controllate.

Le misure organizzative

Rotazione degli incarichi, limiti di esposizione giornaliera, supporto psicologico strutturato e **non stigmatizzato**, debriefing periodici.

NOTA • La verifica umana resta obbligatoria per ogni elemento che entra nel fascicolo e la visione finale spetta all'operatore; l'AI riduce l'esposizione al materiale *non rilevante* e a quello *già catalogato*.

Trovare ciò che la query non sa cercare

Su log e super-timeline (moduli 3 e 5), i modelli non supervisionati segnalano deviazioni dalla baseline: accessi fuori orario, volumi anomali, sequenze di eventi mai osservate, beaconing periodico.

L'anomaly detection segnala anche eventi che l'analista non avrebbe pensato di cercare con una query manuale; le due tecniche sono **complementari**.

Il problema dei falsi positivi

- Un evento anomalo può essere lecito: manutenzioni, picchi legittimi, errori di configurazione
- Base rate fallacy: su milioni di eventi, anche l'1% di falsi positivi sommerge l'analista
- La baseline appresa può incorporare l'attacco, se già in corso durante l'addestramento

Ogni segnalazione è un'ipotesi: la conferma passa dagli artefatti originali.

L'esempio dei dieci milioni di eventi

Un detector «eccellente» su un fenomeno raro produce comunque una valanga di falsi allarmi. Ipotesi: 10.000.000 di eventi, 100 realmente malevoli, detector con recall 99% e tasso di falsi positivi 0,1%.

Veri positivi: $100 \times 99\% = 99$

Falsi positivi: $9.999.900 \times 0,1\% \approx 10.000$

Allarmi totali ≈ 10.099

$P(\text{malevolo} \mid \text{allarme}) = 99/10.099 \approx 1\%$

Su cento allarmi, **novantanove sono falsi** anche con un detector «al 99%»: l'effetto deriva dalla bassa prevalenza del fenomeno.

Per la relazione: dichiarare la prevalenza attesa del fenomeno insieme alle metriche, perché senza base rate precision e recall non si lasciano interpretare.

Conseguenze operative

- Gli allarmi si **aggregano** (per host, per destinazione) prima di investigare: un pattern su molti allarmi vale più del singolo
- Si stratifica: regole deterministiche a valle del detector tagliano i falsi noti
- Il carico atteso per gli analisti va **calcolato prima** di adottare lo strumento

La soglia come politica dell'errore

Il classificatore produce un punteggio continuo; la soglia lo trasforma in decisione. Spostarla scambia un tipo di errore con l'altro:

Soglia	Effetto	Quando ha senso
Bassa (es. 0,3)	Recall alto, molti falsi allarmi	Non si può perdere nulla: reati gravi, triage iniziale ampio
Media (es. 0,5)	Compromesso di default	Nessuna ragione dichiarata per deviare
Alta (es. 0,8)	Pochi allarmi, molto affidabili	Capacità di verifica limitata; il declassato resta comunque esaminabile

ESEMPIO • LA MOTIVAZIONE A VERBALE

«Soglia operativa fissata a 0,35: il quesito riguarda reati contro la persona e la priorità dichiarata dal committente è la completezza (recall stimato 97% sul set di validazione), a fronte di un carico di verifica manuale preventivato in N ore. I reperti sotto soglia restano nel perimetro d'esame con priorità inferiore.»

La motivazione indica il valore scelto, la ragione della scelta e la sorte dei reperti declassati.

In relazione si dichiara sempre la **soglia usata**, anche quando è quella di default, insieme alla curva precision/recall da cui discende.

Concept drift e manutenzione del modello

Un classificatore apprende il mondo com'era nel suo training set, mentre formati, app, famiglie di malware e tecniche di occultamento continuano a cambiare. Le prestazioni misurate alla validazione **decadono silenziosamente**: il modello emette score con la stessa sicurezza su dati che non riconosce più.

Si distingue il **data drift** (cambia la distribuzione degli input: nuovi tipi di file nei sequestri) dal **concept drift** proprio (cambia la relazione fra input ed etichetta: il malware adotta tecniche prima tipiche del software legittimo). Contromisure: monitoraggio di input e score, set di controllo etichettati periodici, **riaddestramento programmato**, data di validazione dichiarata in relazione.

Implicazione forense: le metriche di un modello hanno una data di scadenza. «Accuratezza 0,94» senza data è un'affermazione incompleta.

ESEMPIO • IL DETECTOR DEL 2023 NEL 2026

Un laboratorio usa dal 2023 un classificatore di immagini sintetiche. Le metriche di allora: 0,96 di accuratezza.

Riesaminato nel 2026 su un set di controllo con generatori recenti, scende a 0,71: tre generazioni di modelli generativi dopo, i «difetti tipici» che aveva imparato non esistono più. Il controesame su un caso del 2026: «le prestazioni che lei cita, su quali immagini sono state misurate, e quando?»

La replica manuale su campione come garanzia di metodo

1. Campionare

Estrarre un campione casuale e stratificato degli esiti automatici, includendo deliberatamente i reperti declassati, dove si nascondono i falsi negativi.

2. Replicare

Esaminare il campione con il metodo manuale dei moduli 1-3, in cieco rispetto all'esito del modello, e misurare la concordanza.

3. Documentare

Riportare in relazione strumento, versione, soglie, campione, concordanza e discordanze esaminate: il protocollo è parte della prova.

Perché in cieco. L'automation bias, la tendenza a fidarsi dell'esito della macchina, è documentato e sistematico: una validazione condotta conoscendo già il verdetto del modello tende a confermarlo.

Validare una pipeline di triage

Laboratorio

- Triage AI-assistito di un'immagine forense pubblica
- Filtri NSRL e classificazione
- Replica in cieco su campione e concordanza

DELIVERABLE: rapporto di validazione della pipeline

Letture obbligatorie

- Letteratura DFRWS / FSI Digital Investigation sul triage
- Documentazione Griffeye Analyze (schede tecniche)
- Nota tecnica sugli hash percettivi (PhotoDNA, PDQ)

Verifica

Si valuta il protocollo di validazione: campionamento, cecità, analisi delle discordanze.

Prossimo modulo: il modulo base della prima specializzazione scelta (M4, M5, M6, M9 o M10).

Un'automazione ben progettata concentra il lavoro del perito sulla decisione e sulla sua motivazione.



Grazie per l'attenzione

Vincenzo Calabrò

Domande, materiali del corso e contatti
sul sito del docente



vincenzocalabro.it