Module 2: Anomaly Detection and Attack Prediction with Machine Learning

Adarsh KUMAR

Universitat Politècnica de Catalunya
Department of Computer Science

Project Coordinator:
Prof. Ilker Demirkol

MERiT Project
September 3, 2025

Co-funded by
the European Union

Outline of

Co-funded by
the European Union

Anomaly Detection and Attack Prediction

Key Points

- Anomaly detection identifies unusual patterns that may indicate intrusions, malware, or insider threats.
- Unlike signature-based methods, it can detect zero-day and previously unseen attacks.
- Techniques fall into three main categories:
  - Statistical methods
  - Machine learning-based methods
  - Hybrid approaches

MERIT

Co-funded by the European Union

Statistical Anomaly Detection

Core Idea

Builds mathematical models to define "normal" behavior and flags deviations.

Key Techniques

- Z-score / Standard deviation
- Gaussian distribution modeling
- Time-series analysis

Common Use Cases

- Sudden spike in failed logins
- Unusual data transfer volumes
- Rare login times

MERIT

Co-funded by the European Union

Statistical Detection: Pros and Cons

- **Advantages:**
  - Works with small datasets
  - No dependency on labels
- **Disadvantages:**
  - Fails with non-Gaussian data
  - Cannot explain anomalies

MERIT

DIGITAL EUROPE PROGRAMME

Co-funded by the European Union

Statistical Detection: Pros and Cons

- **Advantages:**
  - Works with small datasets
  - No dependency on labels
- **Disadvantages:**
  - Fails with non-Gaussian data
  - Cannot explain anomalies

MERIT

Co-funded by the European Union

Statistical Detection: Pros and Cons

- **Advantages:**
  - Works with small datasets
  - No dependency on labels
- **Disadvantages:**
  - Fails with non-Gaussian data
  - Cannot explain anomalies

MERIT

Co-funded by the European Union

Statistical Detection: Pros and Cons

- **Advantages:**
  - Works with small datasets
  - No dependency on labels
- **Disadvantages:**
  - Fails with non-Gaussian data
  - Cannot explain anomalies

MERIT

Co-funded by the European Union

Statistical Detection: Pros and Cons

- **Advantages:**
  - Works with small datasets
  - No dependency on labels
- **Disadvantages:**
  - Fails with non-Gaussian data
  - Cannot explain anomalies

MERIT

Co-funded by the European Union

Statistical Detection: Pros and Cons

- **Advantages:**
  - Works with small datasets
  - No dependency on labels
- **Disadvantages:**
  - Fails with non-Gaussian data
  - Cannot explain anomalies

MERIT

Co-funded by the European Union

Machine Learning-Based Anomaly Detection

Core Idea

Algorithms learn patterns to classify behavior as **normal** or **anomalous**.

Unsupervised Learning

No labels needed

Examples: Clustering (k-means), Autoencoders

Supervised Learning

Requires labeled data

Examples: Random Forest, SVM

MERIT

Co-funded by the European Union

Unsupervised Learning Methods

Key Advantage

No labeled data required — ideal for detecting unknown threats.

Popular Techniques

- **Isolation Forest**
  Builds random trees to isolate outliers efficiently.
- **Autoencoders**
  Neural nets flag data with high reconstruction error.
- **Clustering (DBSCAN/k-means)**
  Anomalies fall outside dense clusters.

MERIT

Co-funded by the European Union

Supervised Learning Methods

Core Requirement

Labeled dataset (normal vs. attack samples)
Historical attack data enables high-accuracy detection.

Key Strengths

- **Precise** for known attack patterns
- **Explainable** with feature importance (e.g., Random Forest)
- **Adaptable** via transfer learning (e.g., Neural Nets)

Common Techniques

- **Random Forest**
  Handles mixed data types, resistant to overfitting
- **Support Vector Machines (SVM)**
  Effective for small-to-medium datasets
- **Neural Networks**
  Ideal for complex patterns (e.g., LSTM for sequences)

MERIT

Co-funded by the European Union

ML-Based Detection: Pros and Cons

Advantages

- **Detects complex patterns**
  Identifies non-linear relationships in data
- **Adaptive learning**
  Improves performance with more data
- **Scalable**
  Handles high-dimensional feature spaces

Disadvantages

- **Data preprocessing**
  Requires feature engineering and cleaning
- **Data dependency**
  Performance tied to training data quality
- **Adversarial vulnerability**
  Susceptible to specially crafted attacks

Best Use Case: Dynamic environments where traditional signature-based methods fail

MERIT

Co-funded by the European Union

Hybrid Approaches

Core Concept

Best of both worlds:
Combines statistical rigor with ML adaptability to reduce false positives while maintaining detection accuracy.

Implementation Strategies

- **Statistical pre-filtering + ML classification**
  Use stats to reduce noise before ML analysis
- **Rule-based triage with ML scoring**
  Apply business rules first, then ML for subtle cases
- **Human-in-the-loop validation**
  Analyst reviews ML-flagged anomalies

Key Benefit

30-50% fewer false positives than pure ML approaches
(Source: 2023 SANS Institute Report)

MERIT

Co-funded by the European Union

Hybrid Methods: Use Cases & Evaluation

Key Applications

- **Enterprise Security Monitoring**
  Combines SIEM rules with ML anomaly scoring
- **Threat Hunting Platforms**
  Augments human intuition with statistical baselines
- **User/Entity Behavior Analytics (UEBA)**
  Merges policy-based alerts with behavioral models

Strategic Advantages

- **Enhanced Robustness**
  40% fewer false negatives than pure ML (MITRE 2023)
- **Balanced Performance**
  85% accuracy while maintaining explainability
- **Operational Efficiency**
  Reduces alert fatigue by 60% (Forrester 2024)

Industry Trend: 72% of enterprises now adopt hybrid approaches (Gartner)

MERIT

Co-funded by the European Union

Hybrid Methods: Challenges

Key Limitations

- **Implementation Complexity**
  Requires expertise in both statistical methods and ML
- **Resource Intensive**
  30-50% higher compute needs than single-method systems
- **Integration Overhead**
  Legacy system compatibility issues in 68% of deployments (Gartner 2023)

Mitigation Strategies

- **Phased Rollouts**
  Start with high-value use cases
- **Unified Platforms**
  Pre-integrated solutions reduce complexity
- **MLOps Pipelines**
  Automated model tuning saves 40% effort (McKinsey)

MERIT

Co-funded by the European Union

Choosing the Right Anomaly Detection Technique

Key Selection Criteria

- **Data Availability**
  Labeled vs. unlabeled data volume and quality
- **Threat Complexity**
  Simple rules vs. sophisticated attack patterns
- **Interpretability Needs**
  Regulatory requirements vs. black-box tolerance
- **System Constraints**
  Real-time processing vs. batch analysis capabilities

Decision Framework

- Statistical methods for stable, well-understood systems
- ML approaches for dynamic, complex environments
- Hybrid solutions for critical high-stakes scenarios

Source: NIST SP 800-172 Rev1 (Anomaly Detection Guidelines)

MERIT

Co-funded by the European Union

Comparison of Approaches

Approach	Requirement Level	Strengths	Limitations
Statistical	Low	Simple, interpretable, fast	Based on rigid assumptions, less flexible
ML-Based	Medium–High	Capable of detecting complex patterns and anomalies	Requires large datasets, often lacks transparency (black-box)
Hybrid	Medium	Combines strengths of statistical and ML approaches; more robust	More complex to implement and maintain

MERiT

Co-funded by the European Union

No One-Size-Fits-All

Key Insight

Anomaly detection must be tailored to its environment—there is no universal solution.

- **Context-Specific:** Detection strategies must align with the operational domain and threat landscape.
- **Multi-Method Approaches:** Effective systems often integrate statistical, ML-based, and rule-based methods.
- **Design Goal:** Focus on **reliability**, **scalability**, and **context-awareness**.

MERiT

Co-funded by the European Union

Clustering for Unsupervised Detection

What is Clustering?

Clustering = Unsupervised grouping of similar data points.

Why Clustering Matters in Cybersecurity:

- **Detects Unknown Threats:** Finds previously unseen or novel attack patterns.
- **Groups Similar Behaviors:** Clusters related malicious activities for better pattern recognition.
- **Works Without Labels:** Identifies anomalies in datasets where labeled data is unavailable.

MERiT

Co-funded by the European Union

What is Clustering?

Definition

Unsupervised grouping of similar data points

Key Characteristics

- Automatic pattern discovery
- No need for labeled data
- Measures similarity/distance

Common Uses

- Anomaly detection
- Data segmentation
- Pattern recognition

MERiT

Co-funded by the European Union

Applications in Cybersecurity

Clustering Reveals:

- Normal user behavior
- Types of malicious activities
- Traffic or access patterns
- Device or endpoint profiles

MERIT

Co-funded by the European Union

Why Use Clustering for Security?

Operational Benefits

- **No need for labels:** Ideal for unlabeled, real-world data
- **Data reduction:** Groups similar events for analysis

Security Benefits

- **Unknown threats:** Detects new or stealthy attack patterns
- **Behavioral profiling:** Reveals evolving user/network trends

MERIT

Co-funded by the European Union

Common Clustering Algorithms

Key Algorithms in Cybersecurity

- **K-Means** - Distance-based partitioning
- **DBSCAN** - Density-based spatial clustering
- **Hierarchical** - Nested cluster trees
- **Spectral** - Graph-based clustering

MERIT

Co-funded by the European Union

1. K-Means Clustering

Core Algorithm

- Partitions data into k clusters by minimizing distance to cluster centroids
- Requires the number of clusters (k) as input

Security Use Case

Grouping similar network sessions or user behaviors

Key Limitations

- Assumes spherical clusters
- Sensitive to outliers and initialization

MERIT

Co-funded by the European Union

2. DBSCAN

Density-Based Spatial Clustering of Applications with Noise

- Groups data based on density; no need to specify k
- Can identify outliers (noise)

Security Application

Detecting rare behaviors (e.g., port scans, abnormal logins)

Limitations

- Struggles with clusters of varying density

MERiT

Co-funded by the European Union

3. Hierarchical Clustering

Algorithm Characteristics

- Builds nested tree via merging (agglomerative) or splitting (divisive)
- Useful for exploratory analysis

Security Application

Visualizing relationships between attack behaviors (e.g., malware families)

Limitations

- Computationally expensive for large datasets

MERIT

Co-funded by the European Union

4. Spectral Clustering

Key Features

- Leverages graph theory and similarity matrices
- Effective with non-convex and complex-shaped clusters

Security Application

Analyzing communication patterns (e.g., lateral movement detection)

Limitations

- High computational cost
- Sensitive to parameter tuning

MERIT

Co-funded by the European Union

Applications of Clustering in Cybersecurity

Use Case	Role of Clustering
Network Traffic Analysis	Identify similar traffic patterns; detect anomalies or outliers in real-time flow data.
User Behavior Analytics (UBA/UEBA)	Profile typical user behavior and highlight deviations indicative of insider threats.
Malware Analysis	Group malware samples by behavior or code similarity for faster classification.
Threat Hunting	Uncover hidden relationships or suspicious patterns in massive log datasets.
Incident Response	Aggregate and correlate alerts to streamline investigation and reduce alert fatigue.

MERIT

Co-funded by the European Union

Challenges and Considerations

Clustering in Cybersecurity is Powerful—But Not Without Challenges

Effective use of clustering techniques requires careful consideration of data characteristics, algorithm behavior, and real-world constraints.

- **High-Dimensional Data:** Often requires dimensionality reduction (e.g., PCA, t-SNE).
- **Algorithm Selection:** Choice depends on data distribution, noise tolerance, and scalability.
- **Interpretability:** Clusters must yield actionable insights for security analysts.
- **Scalability:** Must efficiently handle large-scale or real-time streaming datasets.

MERIT

Co-funded by the European Union

Clustering for Security: Summary

Core Benefits

- Discovers unknown threats and structures complex data
- Effective in unsupervised settings without labeled data
- Supports proactive defense and complements ML pipelines

Key Insight

Careful feature selection and algorithm tuning are critical for actionable results.

MERIT

Co-funded by the European Union

Time-Series & Sequential Models in Cybersecurity

Core Concept

- Cybersecurity events occur as **temporal sequences**: login attempts, network flows, user actions
- **Time-aware models** detect patterns traditional models may miss

Key Models

- Long Short-Term Memory (LSTM) networks
- Hidden Markov Models (HMMs)

Why Time Matters

- Brute-force: repeated logins
- Insider threats: subtle access patterns
- Data exfiltration: multi-step behavior

MERIT

Co-funded by the European Union

Long Short-Term Memory (LSTM) Networks

Core Concept

A Recurrent Neural Network (RNN) variant that captures long-term dependencies in sequences

Key Features

- Handles long sequences of actions or events
- Learns both short-term and long-term context
- Trained in supervised or unsupervised modes

MERiT

Co-funded by the European Union

LSTM in Cybersecurity

Critical Applications

- Detecting anomalies in login or user activity logs
- Modeling normal system call sequences for intrusion detection
- Detecting botnet traffic in network flows
- Predictive models for threat escalation

Why LSTMs Excel

- Memory cells maintain state over long sequences
- Gate mechanisms filter irrelevant temporal noise
- Adapts to variable-length security event patterns

MERIT

Co-funded by the European Union

LSTM Networks: Strengths and Limitations

Advantages

- **Complex behavior modeling**
  Captures intricate temporal patterns
- **Real-time analysis**
  Enables live threat detection
- **Length flexibility**
  Handles varying sequence sizes

Challenges

- **Data hungry**
  Requires extensive training data
- **Compute intensive**
  Needs GPU acceleration for efficiency
- **Black-box nature**
  Harder to explain than statistical models

Best for: High-value targets where detection accuracy outweighs resource costs

MERIT

Co-funded by the European Union

Hidden Markov Models (HMMs)

Core Concept

Probabilistic models that analyze:

- **Hidden states** (unobserved system status)
- **Observable events** (e.g., user actions, logs)

Key Advantages

- **Interpretable:** Provides probability scores
- **Categorical:** Ideal for discrete event sequences
- **Intent-aware:** Infers hidden system states

Security Applications

- Insider threat detection
- Malware phase analysis
- Authentication anomaly detection

MERIT

Co-funded by the European Union

Hidden Markov Models in Cybersecurity

Key Applications

- **User Behavior Profiling**
  Baseline normal access patterns
- **Keystroke Dynamics**
  Continuous authentication systems
- **System Call Analysis**
  Malware and intrusion detection
- **Lateral Movement Detection**
  Identify attacker pivoting between systems

MERIT

Co-funded by the European Union

HMM Evaluation

Strengths

- **Interpretable**
  Clear state transition diagrams
- **Data Efficient**
  Works with small discrete datasets
- **Mathematically Sound**
  Well-established theory

Limitations

- **Markov Assumption**
  Limited memory of states
- **Dimensionality**
  Struggles with complex features
- **Long Dependencies**
  Poor at extended temporal patterns

MERIT

Co-funded by the European Union

Feature Engineering for Security

Core Concept

Transforming raw security data (logs, packets) into structured ML-ready features

Critical Importance

- **Noise Reduction**
  Filters irrelevant data artifacts
- **Attack Signal Isolation**
  Amplifies subtle malicious patterns
- **Efficiency**
  Enables real-time processing
- **Performance Boost**
  2-3x accuracy improvement possible

MERIT

Co-funded by the European Union

Common Feature Categories in Network Traffic

Network Feature Taxonomy

- **Volume-based:** Total bytes sent/received, packet count
- **Time-based:** Flow duration, inter-packet arrival time
- **Behavioral:** Number of distinct IPs or ports contacted
- **Protocol-specific:** TCP flags, HTTP methods, DNS query types
- **Statistical:** Mean/variance of packet size or byte rate
- **Connection-level:** Initiator/responder roles, repeated connection attempts

MERiT

Co-funded by the European Union

Techniques and Tools for Feature Engineering

Key Techniques

- **Flow aggregators:** NetFlow, IPFIX, Zeek logs
- **Sessionization:** Group packets into sessions
- **Time-windowing:** Aggregate over time intervals

Practical Examples

- **DDoS detection:** Surges in flows or bytes
- **Port scanning:** Many destination ports with low traffic
- **Botnets:** Repetitive contacts to specific IPs/domains

MERiT

Co-funded by the European Union

Feature Engineering Benefits

Key Advantages

- Bridges the gap between raw traffic/logs and ML models
- Enables detection of nuanced and stealthy behaviors
- Reduces dimensionality and improves scalability
- Enhances explainability and trust in model outputs

Critical Insight

Well-designed features are key to effective ML-based cyber defense.

MERiT

Co-funded by the European Union

Feature Engineering for Log Data

Common Log Sources

- System and application logs
- Authentication logs
- Audit trails

Key Value Proposition

Logs provide contextual, event-driven information essential for detecting behavioral anomalies.

MERIT

Co-funded by the European Union

Feature Categories in Log Data

Key Feature Types

- **Event frequency:** Login failures per time window
- **Temporal patterns:** Time of day, weekday vs. weekend
- **User behavior:** Unique commands, file access patterns
- **Device behavior:** Process tree depth, parent-child relationships
- **Text patterns:** Suspicious keywords in command logs
- **Aggregated statistics:** Avg. session duration, attempts per IP

MERIT

Co-funded by the European Union

Techniques and Use Cases

Key Techniques

- Regex, Logstash, custom parsing scripts
- Time-based aggregation (e.g., 5-min windows)
- Text embedding for command sequences

Security Applications

- **Insider threats:** Rare file access by non-admins
- **Privilege escalation:** Sudden shift in commands
- **Ransomware:** Mass file changes in short time

MERIT

Co-funded by the European Union

Best Practices in Feature Engineering

Critical Guidelines

- **Use domain knowledge:** Understand context behind logs
- **Normalize features:** Prevent bias in scale-sensitive attributes
- **Handle categorical data:** Encode ports, users, protocols
- **Dimensionality reduction:** PCA or feature selection
- **Avoid data leakage:** Only use info available at prediction time

Ultimate Objective

Build accurate, interpretable, and context-aware features.

MERIT

Co-funded by the European Union

Summary: Feature Engineering in Cybersecurity

Key Benefits

- Converts raw logs into structured, ML-friendly inputs
- Enables detection of both common and rare attack behaviors
- Enhances interpretability and operational utility

Core Insight

Careful feature design bridges raw data and actionable insights.

MERIT

Co-funded by the European Union

Real-World Applications of ML in Cybersecurity

ML Enhances Cybersecurity By:

- Adaptability to evolving threats
- Scalability to large data volumes
- Ability to detect unknown (zero-day) threats

Key Applications:

- Intrusion Detection Systems (IDS)
- Malware Detection
- Behavior Analysis

MERIT

Co-funded by the European Union

Intrusion Detection Systems (IDS) with ML

Purpose:

Monitor network/system activity for unauthorized or abnormal behavior

ML Techniques in IDS:

- *Anomaly Detection:* Learn normal network behavior, flag deviations
- *Supervised Classification:* Use labeled data to separate normal vs malicious

Example Features:

- Packet size, byte rate, TCP flags
- Number of connections per host
- Protocols used, destination ports

MERiT

Co-funded by the European Union

IDS Tools and Benefits

Popular Tools:

- Snort + ML plugins
- Zeek/Bro + Python ML scripts
- OpenIDS, Suricata with anomaly detection modules

Benefits:

- Detects zero-day and polymorphic attacks
- Learns continuously from emerging patterns
- Improves detection accuracy over rule-based systems

MERIT

Co-funded by the European Union

Malware Detection with ML

Traditional vs ML-based:

- Traditional antivirus relies on known signatures
- ML detects unseen or obfuscated malware by analyzing behavior and structure

ML Techniques:

- Static Analysis: Code features without execution (API calls, file headers)
- Dynamic Analysis: Behavior during execution (network requests, file changes)
- Hybrid Approaches: Combine static + dynamic features for better detection

MERIT

Co-funded by the European Union

Malware Detection with ML

Example Features:

- Frequency of system calls
- Entropy of binary files
- Registry modifications
- Network connections initiated

MERiT

Co-funded by the European Union

Malware Detection Tools and Benefits

Real-World Tools:

- VirusTotal + ML integrations
- Microsoft Defender for Endpoint
- Cuckoo Sandbox with ML classifiers

Benefits:

- Detects novel malware families
- Resistant to signature evasion
- Scalable to large volumes of files

MERiT

Co-funded by the European Union

Behavior Analysis with ML

Focus:

Detect insider threats, account takeovers, compromised devices

ML Approaches:

- Time-series models (LSTMs) for sequential user actions
- Clustering for unusual user groups
- Anomaly detection on login, file access, commands

MERiT

Co-funded by the European Union

Behavior Analysis with ML

Example Features:

- Login times and frequency
- Accessed resources/folders
- Commands executed on endpoints
- Email or communication patterns

MERiT

Co-funded by the European Union

Behavior Analysis Tools and Benefits

Real-World Tools:

- Splunk UEBA
- IBM QRadar
- Exabeam, Vectra AI

Benefits:

- Effective against insider threats
- Supports real-time monitoring
- Reduces false positives by modeling context

MERIT

Co-funded by the European Union

Summary: Machine Learning in Cybersecurity

Key Takeaways:

- ML empowers detection of evolving, unknown threats
- Enhances IDS, malware detection, and behavior analysis
- Selecting appropriate models and features is key
- Enables scalable, adaptive, and context-aware security

MERiT

Co-funded by the European Union