# Automated Alert Classification Using Machine learning and NLP

Security Operations Centers (SOCs) are inundated with high volumes of alerts, which causes analysts to experience alert fatigue and struggle with timely incident response. The current reliance on manual triage processes amplifies false positive rates and risks critical security incidents being overlooked. Existing automated triage tools often fail to integrate unstructured alert descriptions with structured metadata, thereby missing opportunities to enhance classification accuracy through Natural Language Processing (NLP).

The objective of this project is to develop a hybrid system that combines traditional machine learning with NLP to automatically classify cybersecurity alerts as malicious or benign and assign interpretable priority scores. By leveraging both structured features—such as severity, timestamp, source and destination IP addresses, user and system entities, historical alert frequency, and temporal patterns—and NLP-derived features—such as alert title and description embeddings, TF-IDF vectors, named entities, and sentiment or urgency scores—this solution aims to reduce analyst workload while preserving high accuracy and low false negative rates.

This system will process a stratified sample of at least 1,000 alerts drawn from enterprise SOC environments, with a target of 1,500 to 2,000 records to optimize model performance. Stratification will ensure balanced representation across critical, high, medium, and low-severity categories. The model's output will consist of a severity prediction—categorized as Minor, Major, or Critical—accompanied by probability scores and explanations of the key features influencing each decision, thereby supporting analysts with both rapid and transparent insights.

Data will be sourced through a dual approach. Public datasets such as the AIT alert collection, CICIDS network intrusion logs, the NapierOne threat indicator repository, and the Microsoft Security Incident Prediction dataset will provide a foundational corpus. Where possible, partnerships with private NOC providers and managed SOC services will supply anonymized, expert-verified alerts. Throughout data collection, a minimum of twenty features per alert—split evenly between structured and NLP-derived attributes—will be ensured, with data spanning at least thirty separate days and reflecting a severity distribution of thirty percent critical or high, forty percent medium, and thirty percent low or informational. Expert analysts will validate ground truth labels for each alert.

Protecting sensitive information mandates a comprehensive anonymization strategy. User and entity identifiers will undergo pseudonymization via consistent mapping, while timestamps and location data will be generalized. Textual descriptions will be sanitized to remove any personally identifiable details. All transformations will be logged in an audit trail, reversible only by authorized personnel under strict key management, thereby supporting GDPR compliance and alignment with the NIST Cybersecurity Framework.

Success will be measured against ambitious performance targets inspired by the Automated Alert Classification and Triage (AACT) benchmark. The system is expected to achieve between fifty-five and sixty-five percent alert reduction, maintain a false negative rate below two percent, and exceed ninety percent classification accuracy. The end-to-end processing time per alert should remain under three seconds, with inference latency under two seconds and feature extraction within three seconds, enabling throughput above two hundred alerts per minute. NLP enhancements are projected to deliver at least a fifteen-percent accuracy gain over structured features alone, facilitate semantic clustering of related alerts, and provide interpretable text features to bolster analyst understanding.

Evaluation will employ temporal validation to prevent data leakage, comparing the hybrid model against baseline frequency-based and structured-only XGBoost approaches, an NLP-only TF-IDF SVM classifier, and leading commercial SIEM tools where feasible. Business impact metrics will include analyst workload reduction, mean time to detect, cost per alert processed, and analyst satisfaction scores. Explainability will be reinforced through SHAP and LIME analyses, along with attention visualizations for the NLP components.

The minimum viable product will process at least 1,000 alerts, achieve over eighty-five percent accuracy, maintain a false negative rate below two percent, and demonstrate a forty-percent reduction in analyst effort. Stretch objectives encompass real-time alert processing, seamless integration with major SIEM platforms, automated alert correlation, and an explainable AI dashboard to foster trust and transparency.

In summary, this project combines rigorous data collection, advanced ML and NLP techniques, and robust evaluation methodologies to deliver an automated alert classification and triage system that significantly enhances SOC efficiency without compromising accuracy or compliance.