SENTINEL-AI: A Multi-Perspective AI Framework for Holistic Cyber Threat Demystification and Real-Time Interception

A Submission to The Fursah AI Competition - Tunisia Edition

Fares Mallouli
Ayoub Walha

May 30, 2025

Contents

List of Figures

Executive Summary

The escalating sophistication of cyber threats necessitates advanced, multi-faceted defense mechanisms. This report details **SENTINEL-AI**, an innovative framework developed for The Fursah AI Competition, designed to dissect complex network traffic, attribute its origin, discern operator intent, and preemptively identify malicious activities, including elusive AI-enhanced bots and automated attack tools.

SENTINEL-AI employs a **tri-partite AI architecture**, with each component specializing in a distinct analytical dimension:

1. **Main Threat Detector:** A high-fidelity classifier for identifying diverse attack types (e.g., DDoS, Botnets, Port Scans, Web Attacks) and benign traffic. This engine leverages the rich, NTLFlowLyzer-derived features from the BCCC-CIC-IDS2017 dataset, with web attack detection uniquely enhanced by CTGAN-based data augmentation.
2. **Operator Identification Model:** A behavioral profiler designed to distinguish between human-generated traffic and automated bot activity, utilizing nuanced flow characteristics indicative of scripted versus organic behavior.
3. **Source Classification Model:** An intelligent IP enrichment and categorization engine that determines traffic origin (e.g., ISP/Residential, Cloud Provider, Hosting/DataCenter, VPN/Proxy) by integrating ASN data, cloud provider ranges, and known malicious infrastructure lists.

This multi-perspective approach allows SENTINEL-AI to achieve robust performance across the competition's evaluation criteria. Key achievements include a Main Threat Detector accuracy of **0.9602** (F1-Macro: 0.6251), an Operator ID model Test Accuracy of **0.9891** (F1-Macro: 0.9779), and a Source Classifier Test Accuracy of **0.9761** (F1-Macro: 0.9762).

Crucially, SENTINEL-AI culminates in a **real-time inference pipeline built upon NFStream**, demonstrating the practical applicability of our framework for live network traffic analysis. The system dynamically extracts features, invokes the specialized models, and provides consolidated, actionable insights for each network flow. A significant innovation is the integration of **SHAP (SHapley Additive exPlanations)** into this real-time pipeline, offering instance-level explainability for the Main Threat and Operator ID models. Our architecture prioritizes transparency and explainability, moving beyond "black box" predictions. By strategically leveraging the enhanced BCCC-CIC-IDS2017 dataset and an innovative, modular design, SENTINEL-AI offers a comprehensive and effective solution to modern network security challenges.

Chapter 1

Introduction

The digital landscape is characterized by an unceasing evolution of cyber threats, ranging from widespread automated attacks to highly targeted, sophisticated campaigns. Traditional signature-based detection methods struggle against polymorphic malware, zero-day exploits, and the increasing use of legitimate infrastructure for malicious purposes. Addressing this dynamic threat environment requires intelligent systems capable of deep behavioral analysis, accurate source attribution, and timely operator identification.

The Fursah AI Competition - Tunisia Edition presents a comprehensive challenge: to develop AI-powered solutions that can:

- Classify the origin of incoming traffic.
- Differentiate traffic types, distinguishing legitimate services from malicious sources and automated bots.
- Determine if traffic is human-generated, bot-automated, or AI-enhanced.
- Detect behavioral anomalies, such as uniform vs. random request timing.
- Identify specific cyber threats like automated scanning tools or compromised machines.

A critical technical requirement is the capability for real-time analysis to ensure timely threat detection and response.

This report introduces **SENTINEL-AI**, our solution to this multifaceted challenge. SENTINEL-AI is not a monolithic system but rather a synergistic framework of specialized AI models, each designed to provide a unique perspective on network traffic. Our guiding philosophy is that robust cyber defense stems from a profound understanding of network activity – the *what* (threat type), the *who* (operator type), and the *where* (source origin). By integrating these insights, SENTINEL-AI aims to provide actionable intelligence for proactive network security.

The foundation of our work rests on the BCCC-CIC-IDS2017 dataset [1], an enhanced collection of network flows generated by the NTLFlowLyzer tool, providing a rich feature set for behavioral analysis. Our approach emphasizes modularity, explainability, and, crucially, the translation of batch-trained models into a practical real-time inference capability offering instance-level interpretations.

Chapter 2

System Architecture and Methodology

SENTINEL-AI is architected as a modular, multi-stage system designed for comprehensive network traffic analysis. Figure 2.1 illustrates the high-level flow from data ingestion and preprocessing through specialized AI engine analysis to consolidated, explainable insights.

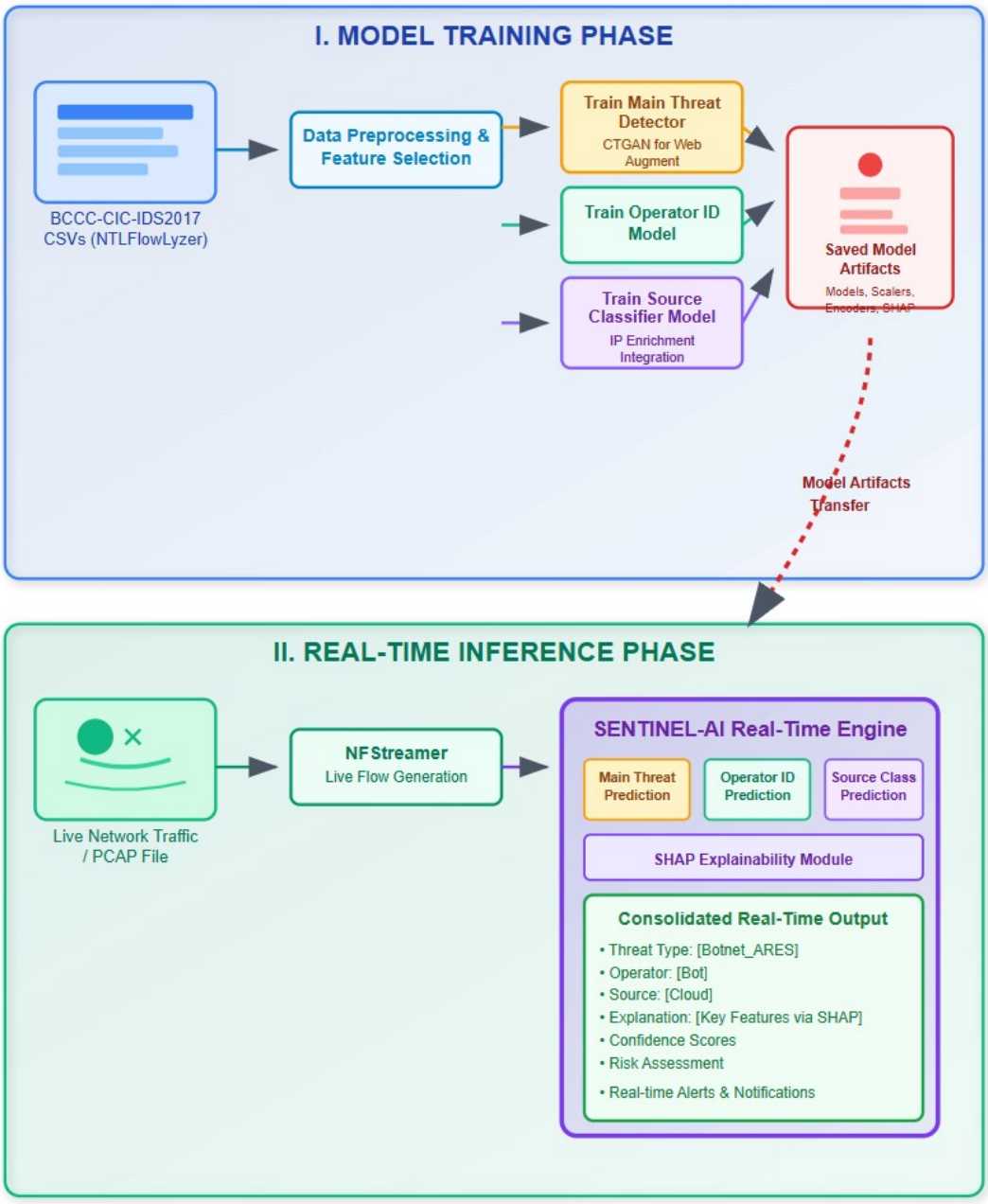


Figure 2.1: SENTINEL-AI High-Level System Architecture.

The core of SENTINEL-AI comprises three distinct AI engines, each addressing specific aspects of the competition's challenges.

2.1 Data Foundation: BCCC-CIC-IDS2017

Our primary data source is the BCCC-CIC-IDS-2017 dataset [1]. This choice was strategic, as the dataset is generated using NTLFlowLyzer, providing an enhanced and cleaned set of over 2000 flow features designed for nuanced behavioral profiling. This allowed us to focus on advanced model development rather than extensive preliminary data cleaning typically associated with older datasets. The specific CSV files utilized correspond to various benign and attack scenarios, as detailed by the dataset creators.

2.2 Engine 1: Main Threat Detector

2.2.1 Objective

To classify network flows into fine-grained categories, including various attack types (e.g., `DDoS_LOIT`, `Botnet_ARES`, `Port_Scan`, `Web_Brute_Force`) and `Benign` traffic.

2.2.2 Data and Labeling

Utilizes the inherent labels within the BCCC-CIC-IDS2017 dataset.

Advanced Data Handling: Augmentation for Imbalanced Classes

Network intrusion datasets are often characterized by significant class imbalance, where malicious traffic, especially specific attack types, is far less prevalent than benign traffic. Traditional resampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) can be beneficial but may sometimes generate less diverse or overly simplistic synthetic samples.

For SENTINEL-AI, particularly for the **Main Threat Detector**, addressing the imbalance of specific critical attack classes was paramount. While SMOTE combined with Random Under-Sampling (RUS) was employed for general class balancing, a more sophisticated approach was adopted for the highly imbalanced and nuanced web attack classes, specifically `Web_Brute_Force`, `Web_SQL_Injection`, and `Web_XSS`.

We utilized **CTGAN (Conditional Tabular GAN)** [2] to generate synthetic data specifically for these web attack categories. CTGAN, a type of Generative Adversarial Network, learns the underlying data distribution of the minority class and generates new, high-fidelity synthetic samples that better capture the complex relationships between features compared to simpler interpolation-based methods. This targeted augmentation, with CTGAN epochs set to **20** and aiming for approximately **300** total samples per minority web attack class, was instrumental in enhancing the performance of our web attack detection sub-module, leading to improved precision and recall for these challenging threats.

2.2.3 Feature Engineering and Selection

We curated a set of **88 NFStream-compatible features** from the BCCC-CIC-IDS2017 dataset. To optimize performance and reduce dimensionality, a feature selection process involving 'SelectKBest' (using ANOVA F-value) and 'SelectFromModel' (with a RandomForestClassifier) was applied, resulting in a final set of **60 features** for the main model.

Rationale for Excluding Certain Identifiers as Direct Features

To prevent data leakage, ensure model generalizability, and focus on behavioral patterns, several common flow identifiers were deliberately excluded as direct input features for the Main Threat Detector:

- **Flow ID, Source/Destination IP Addresses, Timestamps:** These features exhibit extremely high cardinality. Using them directly could lead the model to memorize specific instances from the training set rather than learning generalizable

attack patterns. The objective is to identify malicious *behavior*, irrespective of the exact IP or timestamp. Temporal aspects relevant to behavior are captured by features like ‘duration‘ and Inter-Arrival Time (IAT) statistics.

- **Source Port:** Source ports are typically ephemeral and randomly assigned, offering little predictive value for general threat detection compared to destination ports, which indicate the targeted service and are retained.

While IP addresses are not used as direct features here, the Source IP is crucially utilized by Engine 3 (Source Classification Model) for origin attribution after enrichment.

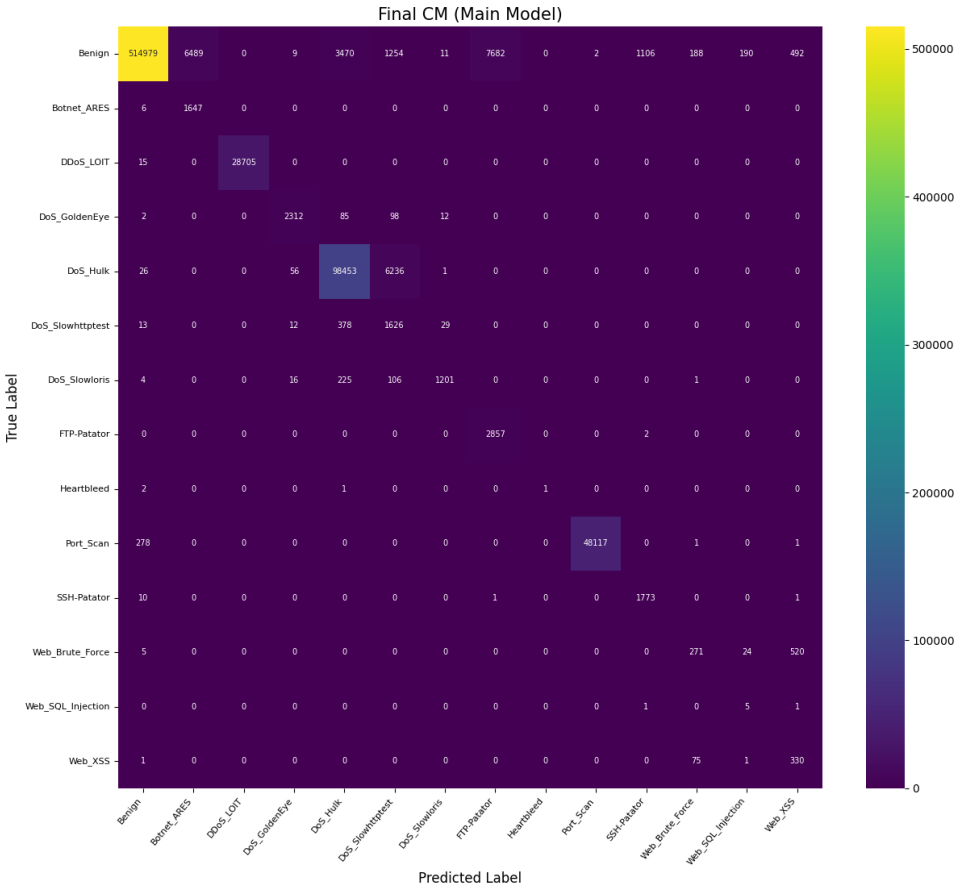
2.2.4 Modeling

A dual-model approach was implemented:

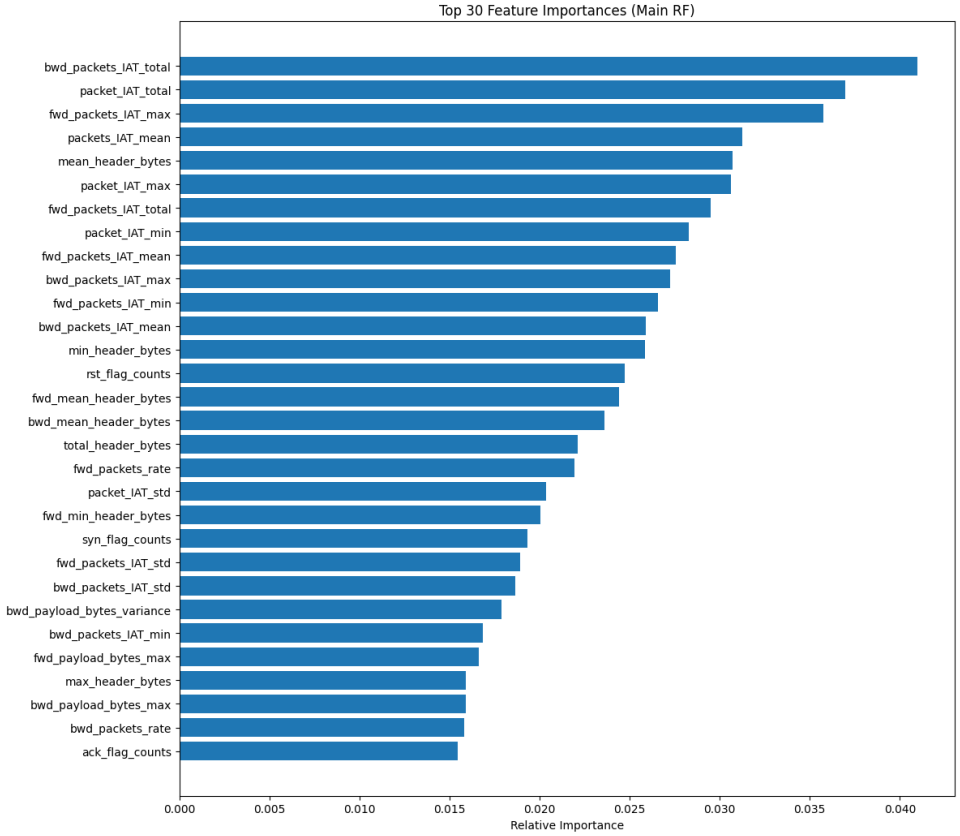
- **Primary Classifier:** A ‘RandomForestClassifier‘ (n_estimators=150, max_depth=30, class_weight='balanced_subsample') trained on the selected features.
- **Specialized Web Attack Detector:** A ‘LGBMClassifier‘ specifically trained to identify web attacks, benefiting from the CTGAN-augmented data. Its predictions refine the main classifier's output for web-related threats using a confidence threshold of **0.7**.

2.2.5 Key Results

The Main Threat Detector achieved an overall accuracy of **0.9602**, a macro F1-score of **0.6251**, and a weighted F1-score of **0.9699** on the test set. The confusion matrix (Figure 2.2a) illustrates per-class performance. Feature importance analysis (Figure 2.2b) highlights key discriminators such as `bwd_packets_IAT_total` and `fwd_init_win_bytes`.



(a) Confusion Matrix for Main Threat Detector.



(b) Feature Importances for Main Threat Detector.

Figure 2.2: Main Threat Detector Evaluation Metrics.

2.3 Engine 2: Operator Identification Model

2.3.1 Objective

To determine if network traffic is generated by a human operator or an automated bot.

2.3.2 Data and Labeling

Leveraging the BCCC-CIC-IDS2017 dataset, we established a binary labeling heuristic: flows from benign traffic capture files (e.g., `monday_benign.csv`) were labeled as 'Human' (0), while flows from attack scenario files (e.g., `ftp_patator.csv`, `dos_hulk.csv`) were labeled as 'Bot' (1).

2.3.3 Feature Engineering and Selection

A curated set of **78 base behavioral features** (e.g., 'duration', IAT statistics, payload characteristics, TCP flag counts, active/idle times) was initially considered. 'SelectKBest' (ANOVA F-value) was used to reduce this to an optimal set of **40 features**.

Exclusion of Direct Identifiers

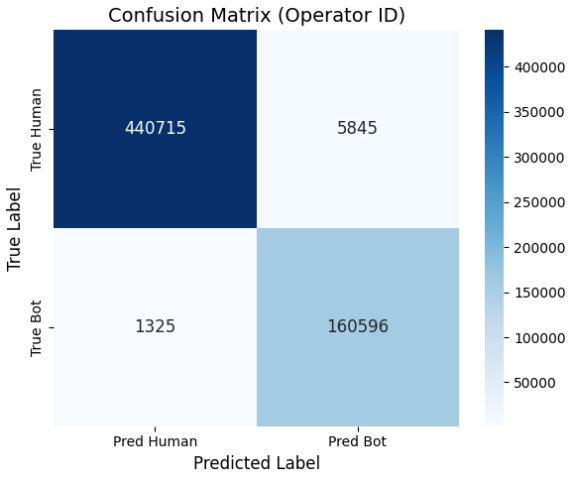
Similar to the Main Threat Detector, direct identifiers like Flow ID, specific IP addresses, and raw timestamps were excluded. This decision is critical for preventing overfitting and focusing on behavioral patterns (e.g., request timing, session duration) rather than specific instance memorization.

2.3.4 Modeling

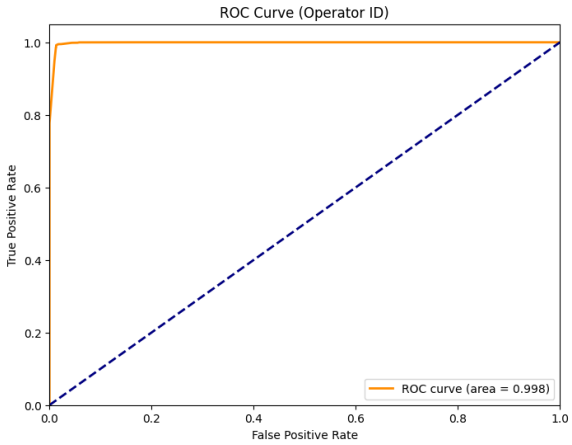
A 'RandomForestClassifier' (n_estimators=150, max_depth=20, class_weight='balanced') was trained. Data was scaled using a 'PowerTransformer'. Cross-validation (3-fold) on the training set yielded a mean F1-score of **0.9782**.

2.3.5 Key Results

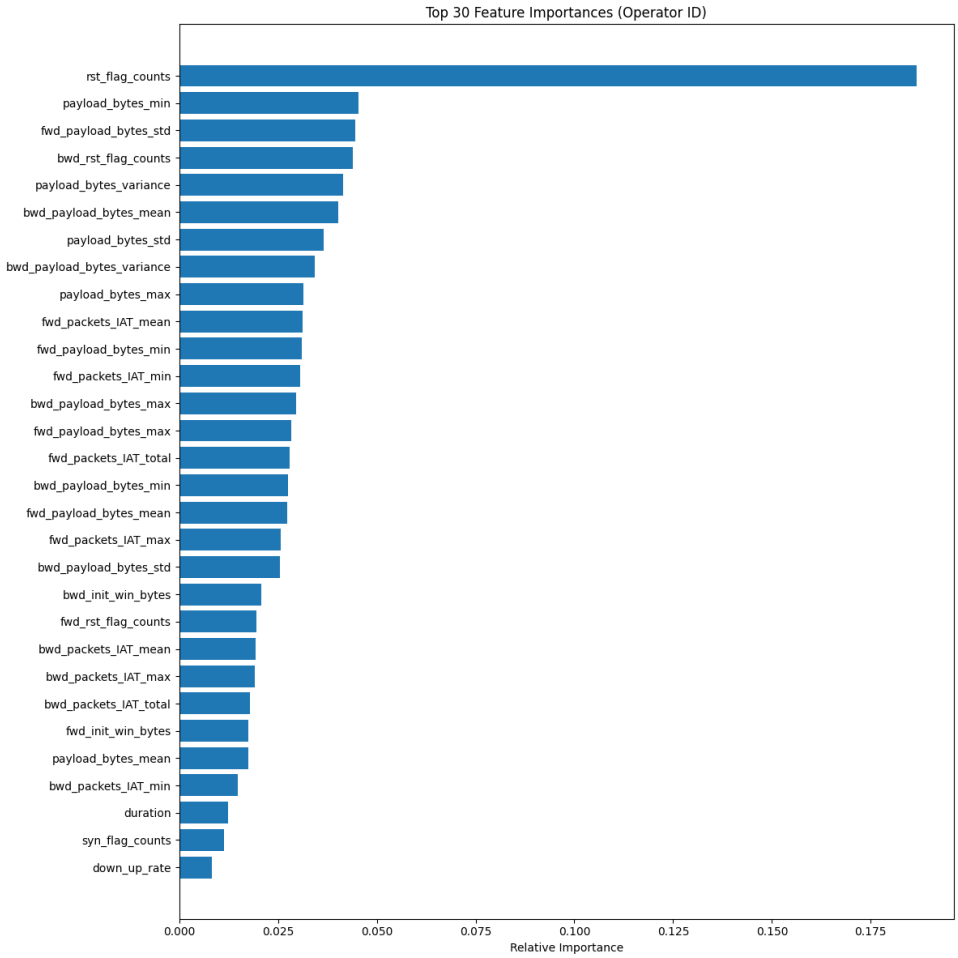
On the hold-out test set, the Operator ID model achieved an accuracy of **0.9891**, a macro F1-score of **0.9779**, and a ROC AUC of **0.9984**. Figures 2.3a and 2.3b show the confusion matrix and ROC curve. Feature importances (Figure 2.3c) indicate that features like `bwd_packets_IAT_mean` and `fwd_init_win_bytes` are highly discriminative.



(a) Confusion Matrix for Operator ID.



(b) ROC Curve for Operator ID.



(c) Feature Importances for Operator ID.

Figure 2.3: Operator Identification Model Evaluation Metrics.

2.4 Engine 3: Source Classification Model

2.4.1 Objective

To classify the origin of incoming traffic into categories such as ISP/Residential, Cloud, Hosting/DataCenter, Education, VPN/Proxy, or Unknown.

2.4.2 Data and IP Enrichment

Source IP addresses (`src_ip`) were extracted from all BCCC-CIC-IDS2017 flows. A comprehensive IP enrichment pipeline was developed:

- **ASN Lookup:** GeoLite2 ASN data was used to map IPs to ASNs and organization names, optimized using an IntervalTree.
- **Cloud Provider Identification:** IP ranges for AWS, Azure, and GCP were loaded.
- **VPN/Proxy Detection:** Lists of known NordVPN servers and Tor nodes were integrated.
- **Categorization Logic:** A rule-based heuristic (using the function `categorize_source_inf`) assigned IPs to the defined `TARGET_SOURCE_CATEGORIES`.

2.4.3 Feature Engineering

Features for this model are derived characteristics: `ASN` (numerical), `ASN_Org` (categorical), `Is_Cloud` (boolean), and `Is_Proxy` (boolean). This approach avoids data leakage from using raw IP addresses as direct model inputs, focusing instead on their enriched attributes.

2.4.4 Modeling

A scikit-learn 'Pipeline' was used, incorporating a 'StandardScaler' for numerical features and a 'OneHotEncoder' for the categorical `ASN_Org` feature. This preprocessed data was then fed into a 'RandomForestClassifier' (n_estimators = 100, max_depth = 25, class_weight = 'balanced').

2.4.5 Key Results

The Source Classifier achieved a test accuracy of **0.9761** and a macro F1-score of **0.9762**. The confusion matrix (Figure 2.4) details per-category performance.

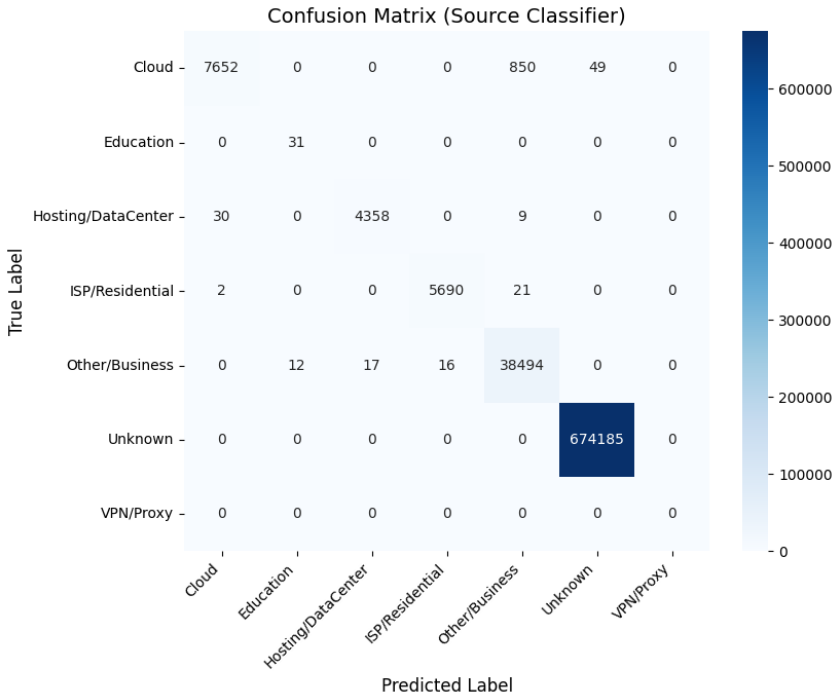


Figure 2.4: Source Classification Model Confusion Matrix.

Chapter 3

Real-Time Inference Pipeline with NFStream

A cornerstone of SENTINEL-AI is its capability for real-time analysis, achieved through an inference pipeline leveraging the **NFStream** library [3]. This pipeline processes live network traffic, extracts relevant features, and applies the trained SENTINEL-AI models for on-the-fly classification and explanation.

3.1 Pipeline Architecture

The real-time pipeline operates as depicted in Figure 3.1.

The specific stages involved are:

1. **Traffic Ingestion:** NFStreamer captures packets from a network interface or PCAP file.
2. **Flow Generation & Feature Extraction:** NFStream reconstructs flows. A custom function (`nfstream_flow_to_features_dict`) dynamically extracts attributes and maps them to the required feature sets for all three SENTINEL-AI engines.
3. **Model Invocation Enrichment:**
   - Features are scaled using pre-fitted scalers.
   - The Main Threat Detector, Operator ID Model, and Source Classifier are invoked.
   - For the Source Classifier, real-time IP enrichment (ASN, cloud, VPN/Tor) is performed using cached databases.
4. **Real-Time Explainability (SHAP):** For the Main Threat and Operator ID models, **SHAP (SHapley Additive exPlanations)** [4] values are calculated per-flow, providing instance-level explanations of feature contributions to the prediction.
5. **Aggregated Output:** Consolidated predictions (Threat Type, Operator, Source Category) and SHAP explanations are generated for each flow.

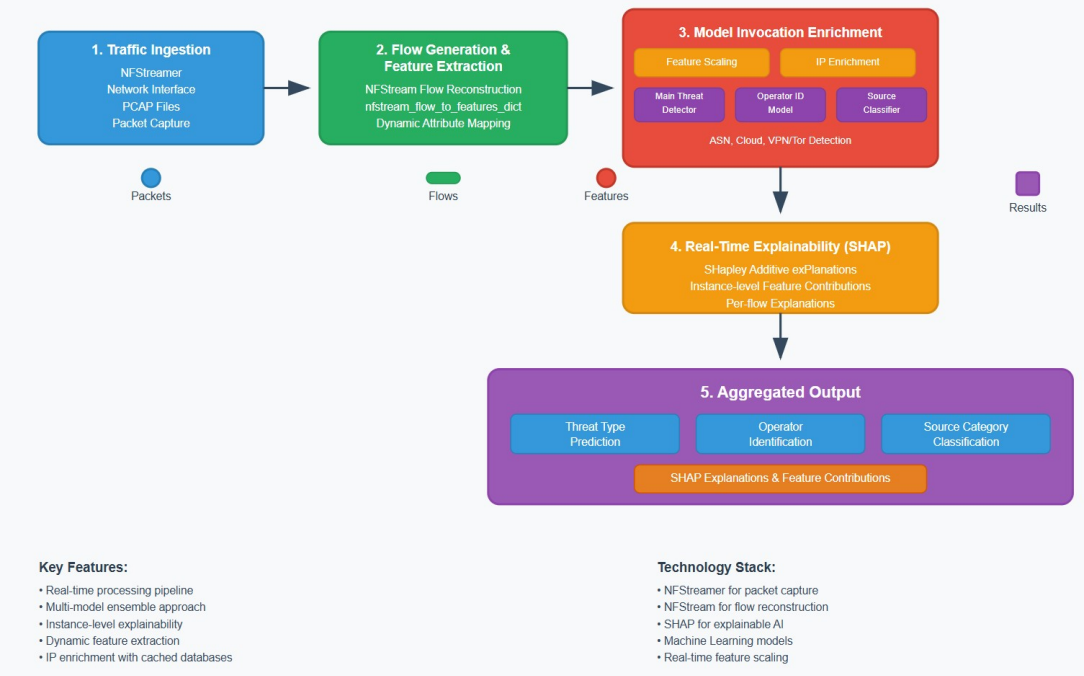


Figure 3.1: SENTINEL-AI Real-Time Inference Pipeline Architecture.

3.2 Key Implementation Features for Real-Time Performance

- **Efficient Initialization via Pre-loaded Artifacts:** To ensure rapid per-flow processing, all essential components—including trained models, data scalers, label encoders, selected feature lists, SHAP explainers, and IP enrichment databases—are loaded into memory once at the pipeline's startup.
- **Accelerated IP Enrichment with Optimized Lookups:** The `lru_cache` decorator is employed to significantly speed up repeated lookup queries for IP address enrichment (e.g., ASN and cloud/VPN/Tor status), minimizing latency.
- **Coordinated Analysis through Modular Prediction Logic:** A central function, `process_flow`, orchestrates the sequential invocation of each specialized AI engine (Threat, Operator, Source) and integrates their respective predictions and SHAP explanations for a comprehensive output.

3.3 Example Real-Time Output with Explainability

Figure 3.2 demonstrates the rich output from our real-time inference pipeline, including SHAP explanations that detail feature contributions to specific predictions. This example showcases the system processing a suspicious flow.

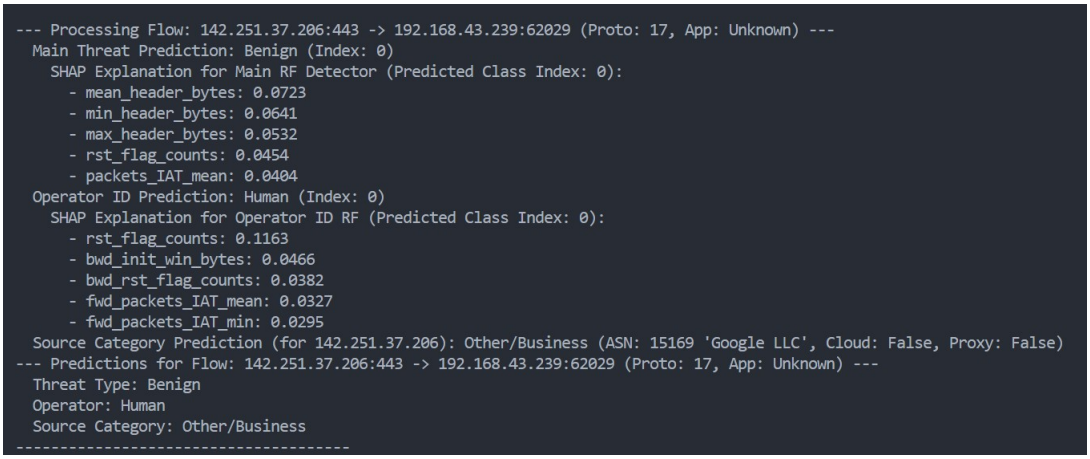


Figure 3.2: Example Console Output from Real-Time Inference Pipeline with SHAP Explanations.

This real-time, explainable output, as illustrated in Figure 3.2, significantly aids analysts in validating alerts and understanding the underlying characteristics and drivers of a detected threat. The instance-level SHAP explanations provide immediate, feature-level justification for the AI's classifications.

Chapter 4

Addressing Competition Objectives

SENTINEL-AI's multi-engine design directly addresses the diverse requirements of the Fursah AI Competition.

- **Classify the origin of incoming traffic:** Explicitly handled by the **Source Classification Model** (Section 2.4), achieving **0.9761** accuracy.
- **Differentiate traffic types (legitimate, malicious, automated bots):**
  - Legitimate vs. Malicious: Core function of the **Main Threat Detector** (Section 2.2), accuracy **0.9602**.
  - Automated Bots: Identified by the **Operator ID Model** (Section 2.3), F1-score **0.9779** for 'Bot' class, and specific botnet classes in the Main Threat Detector.
- **Determine if traffic is human-generated, bot-automated, or AI-enhanced:**
  - Human vs. Bot: Directly classified by the **Operator ID Model**.
  - AI-enhanced: While not an explicit output class, SENTINEL-AI provides indicators. AI-enhanced bots might be flagged as 'Bot' but exhibit novel attack signatures, warranting investigation. SHAP explanations can highlight unusual feature contributions for such traffic.
- **Detect behavioral anomalies (uniform vs. random request timing, scripted input):** Features in the Operator ID Model and Main Threat Detector (IAT statistics, duration, active/idle times) are sensitive to these. Feature importance plots and real-time SHAP values (Figures 2.2b, 2.3c) confirm their relevance.
- **Identify potential cyber threats (automated scanning, adversarial AI, compromised machines):**
  - Automated Scanning: The `Port_Scan` class in the Main Detector effectively identifies such activities.
  - Adversarial AI Models: These may present as novel patterns, potentially flagged by our models or through SHAP explanations indicating unusual feature usage.
  - Compromised Machines: Often exhibit bot-like behavior, which would likely be identified by botnet classes in the Main Detector or as 'Bot' by the Operator ID Model.

4.1 Performance and Efficiency

- **Accuracy:** Strong metrics achieved across all models (Chapter 2).
- **Efficiency:** The choice of RandomForest/LGBM classifiers, coupled with strategic feature selection, ensures efficient inference. The real-time pipeline (Chapter 3) is designed for flow-by-flow processing, with pre-loaded models and cached lookups to optimize speed.
- **False Positive/Negative Rates:** Confusion matrices detail these. For instance, the Main Threat Detector shows high recall for critical attacks like `Botnet_ARES` (0.996), minimizing False Negative Rates for such threats.

4.2 Explainability

SENTINEL-AI prioritizes model transparency:

- **Global Feature Importance:** Plots for each model provide general insights into feature contributions (e.g., Figures 2.2b, 2.3c).
- **Instance-Level Real-Time SHAP Explanations:** For the Main Threat and Operator ID Models, SHAP values are calculated per-flow in real-time. This quantifies individual feature contributions to specific predictions, offering dynamic, on-the-fly reasoning (Section 3).
- **Modular Design & Transparent Data Handling:** The distinct engines and clear rationale for feature exclusion Rationale for Excluding Certain Identifiers as Direct Features enhance system comprehensibility.

This multi-faceted approach ensures SENTINEL-AI's outputs are interpretable and actionable.

Chapter 5

Innovation and Effectiveness

SENTINEL-AI's primary innovation lies in its **holistic, multi-perspective approach** to network threat analysis, culminating in an **explainable real-time inference system**.

Key innovative aspects include:

1. **Synergistic Multi-Model Architecture:** Integration of distinct Threat, Operator, and Source classifiers.
2. **Advanced IP Enrichment Pipeline:** Robust real-time IP origin attribution.
3. **Targeted Data Augmentation (CTGAN):** Enhancing web attack detection robustness.
4. **Practical Real-Time Deployment with NFStream & SHAP Explainability:** Demonstrating operational viability with instance-level interpretation, a significant step beyond static model evaluation.
5. **Methodical Feature Engineering and Exclusion Rationale:** Demonstrating a deep understanding of ML best practices and data leakage prevention.

The effectiveness of SENTINEL-AI is underscored by its strong performance metrics, its comprehensive addressing of competition objectives, and its ability to provide rich, interpretable insights in real-time.

Chapter 6

Conclusion and Future Work

SENTINEL-AI successfully addresses the Fursah AI Competition's challenges by delivering an innovative, multi-engine framework capable of comprehensive network traffic analysis, classification, and real-time, explainable inference. Its modular design, strategic data handling, and focus on practical deployment provide a robust solution for modern cybersecurity threats.

6.1 Future Work

- **Explicit AI-Enhanced Bot Detection:** Develop dedicated anomaly detection modules.
- **Unified Risk Scoring:** Consolidate outputs into a single risk score.
- **Graph-Based Relational Analysis:** Explore relationships between network entities.
- **Advanced Real-Time Feature Engineering:** Integrate more complex behavioral features via NFStream plugins.
- **Continuous Adaptation Scalability:** Investigate federated learning and further optimize the pipeline for high-throughput environments.

In conclusion, SENTINEL-AI offers a powerful and extensible platform for AI-driven network security, with significant potential to enhance threat detection and response capabilities.

Bibliography

[1] "NTLFlowLyzer: Towards generating an intrusion detection dataset and intruders behavior profiling through network and transport layers traffic analysis and pattern extraction," *Computers & Security*, vol. 140, p. 103749.

[2] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[3] Z. Saad, "NFStream: a Flexible Network Data Analysis Framework.," Mar. 2020.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, vol. 30, 2017.