Application of Machine Learning and XAI in Cybersecurity

M Sivamohan
210020025

November 11, 2023

Contents

1 Introduction

1.1 Background

Cybersecurity has become a critical concern in our interconnected digital world. The increasing sophistication of cyber threats poses a significant challenge to safeguarding sensitive information and critical infrastructure. Traditional rule-based security systems struggle to keep pace with the evolving nature of these threats. Machine Learning (ML) emerges as a promising solution, providing the capability to detect and mitigate cyber-attacks in real-time.

In the context of ML in cybersecurity, the CICIDS 2019 Dataset stands out as a valuable resource for researchers and practitioners. This dataset encompasses diverse cyber-attack scenarios, allowing for the development and evaluation of robust models. Among these, the *DrDoS_LDAP_data_2_0_per* subset specifically addresses LDAP (Lightweight Directory Access Protocol) attacks, a prevalent vector for cyber threats.

LDAP attacks exploit vulnerabilities in the LDAP protocol, a widely used protocol for accessing and maintaining distributed directory information services. LDAP is commonly employed for user authentication and authorization within networks. Attackers leverage various techniques, such as reflection and amplification, to overwhelm LDAP servers with malicious traffic, causing service disruptions and compromising the confidentiality and integrity of directory information.

Organizations relying on LDAP for centralized authentication, like enterprises and educational institutions, are particularly vulnerable to LDAP attacks. The potential impact extends beyond service disruption to unauthorized access and data breaches. As such, the detection and prevention of LDAP attacks are paramount to ensuring the security and integrity of networked systems.

In recent years, there has been a remarkable rise in machine learning (ML) models, allowing for data-driven decision-making in many fields. These models, on the other hand, frequently act as complex "black boxes" making it difficult to understand how and why they make specific predictions. In fields where trust, accountability, and regulatory compliance are critical, interpretability and explainability are important. We have people working on this called Explainable Artificial Intelligence (XAI). In this, we will be looking at how to use these XAI in ML models to make sense of the predictions more specifically, we will be using SHAP (SHapley Additive exPlanations).

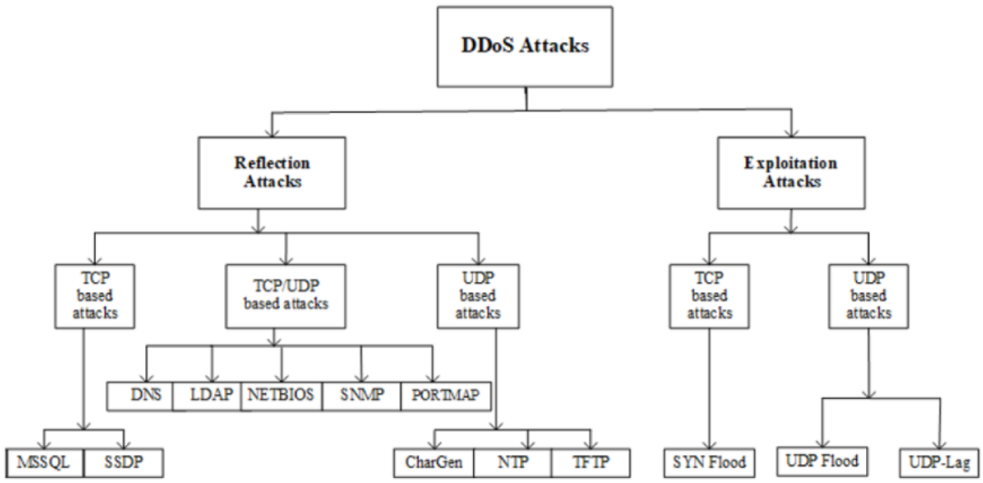


Figure 1: DDOS Attacks classification.

1.2 Problem Statement

Despite the advancements in ML-based cybersecurity solutions, accurately identifying and classifying specific types of attacks, such as LDAP attacks, remains a challenge. The intricacies of network traffic patterns and the dynamic nature of cyber threats necessitate the development of specialized models for effective detection. Addressing these challenges requires a nuanced approach, considering both the intricacies of the dataset and the complexity of the attacks.

1.3 Objectives

The primary objective of this research is to leverage ML techniques for the detection of LDAP attacks using the CICIDS 2019 Dataset. Specific objectives include:

- Evaluate the model's performance in terms of accuracy, precision, recall, and F1 score.
- Investigate the interpretability of the model through eXplainable AI (XAI) techniques.

2 Motivation

2.1 Importance of Cybersecurity

In our digitally interconnected world, the importance of cybersecurity cannot be overstated. As businesses, governments, and individuals increasingly rely on digital platforms and networks, the potential for cyber threats to compromise sensitive information and critical infrastructure grows exponentially. Cybersecurity serves as the first line of defense against malicious activities such as data breaches, unauthorized access, and service

disruptions. A breach in security not only jeopardizes privacy and trust but also has severe financial and operational consequences. The imperative to secure digital assets underscores the critical role of cybersecurity in preserving the integrity and functionality of our modern society.

2.2 Rise of Cyber Threats

The landscape of cyber threats is continually evolving, presenting a formidable challenge to conventional security measures. Cybercriminals employ increasingly sophisticated tactics, ranging from malware and phishing attacks to advanced persistent threats (APTs). The proliferation of connected devices through the Internet of Things (IoT) further expands the attack surface, providing malicious actors with new avenues for exploitation. Nation-state-sponsored attacks and ransomware incidents have become more frequent and sophisticated, underscoring the need for adaptive and proactive cybersecurity measures to mitigate the growing risk of cyber threats.

2.3 Need for Machine Learning in Cybersecurity

Traditional cybersecurity approaches, often rule-based and reliant on known patterns, struggle to keep pace with the dynamic and adaptive nature of modern cyber threats. Machine Learning (ML) has emerged as a powerful ally in the fight against cyber threats due to its ability to analyze vast amounts of data, identify patterns, and adapt to new and evolving attack vectors. ML algorithms can detect anomalies, recognize patterns indicative of malicious activity, and enhance the efficiency and speed of threat detection. The integration of ML into cybersecurity practices represents a paradigm shift towards more robust and proactive defense mechanisms.

2.4 Role of eXplainable AI (XAI) in Enhancing Security

While ML enhances the capability to detect and prevent cyber threats, the "black box" nature of many complex models raises concerns about interpretability and trust. eXplainable AI (XAI) addresses this challenge by providing insights into the decision-making processes of ML models. In the context of cybersecurity, understanding why a model flags certain activities as malicious is crucial for effective decision support and response. XAI techniques, such as feature importance analysis and model-agnostic explanations, contribute to the transparency and trustworthiness of ML-based cybersecurity systems. By enhancing the interpretability of ML models, XAI facilitates collaboration between human analysts and automated systems, resulting in a more effective and accountable defense against cyber threats.

Together, these motivations highlight the pressing need for advanced technologies like ML and XAI in the realm of cybersecurity to fortify our defenses against the ever-evolving landscape of cyber threats.

3 Literature Review

3.1 Overview of ML in Cybersecurity

Machine Learning (ML) has emerged as a transformative force in cybersecurity, offering a dynamic and adaptive approach to threat detection and mitigation. ML algorithms excel at analyzing large volumes of diverse data to identify patterns and anomalies that may elude traditional rule-based systems. In cybersecurity, ML is applied to tasks such as intrusion detection, malware classification, and behavioral analysis. The ability of ML models to learn from data and adapt to evolving threats positions them as key components in building proactive and effective cybersecurity systems.

3.2 Significance of XAI in Model Interpretability

The significance of eXplainable AI (XAI) in the context of model interpretability is increasingly recognized as crucial for the adoption and trustworthiness of machine learning systems, especially in cybersecurity. XAI techniques aim to demystify the decision-making processes of complex ML models, making them more transparent and understandable to human analysts. In the cybersecurity domain, where the consequences of false positives or negatives can be severe, XAI plays a pivotal role in enhancing the interpretability of models. By providing insights into why a model classifies an activity as malicious, XAI fosters trust, aids in fine-tuning model parameters, and facilitates more effective collaboration between human analysts and automated systems.

4 Dataset

4.1 Introduction to CICIDS 2019 Dataset

The CICIDS 2019 (Canadian Institute for Cybersecurity Intrusion Detection Systems) dataset serves as a comprehensive and widely utilized resource in the field of cybersecurity research. Released for the purpose of fostering advancements in intrusion detection system evaluation, this dataset encompasses a diverse range of cyber attack scenarios, network traffic patterns, and anomalies. It provides a realistic simulation of real-world cyber threats, making it a valuable asset for developing and testing machine learning models. Comprising both benign and malicious network traffic, the CICIDS 2019 dataset enables researchers to address the complexities associated with false positives and negatives, essential for the robust evaluation of intrusion detection systems.

4.2 DrDoS_LDAP_data_2_0_per Subset

Within the expansive CICIDS 2019 dataset, the DrDoS_LDAP_data_2_0_per subset specifically focuses on LDAP-based Distributed Denial of Service (DrDoS) attacks. LDAP, a protocol widely used for accessing and maintaining directory services, becomes a target for malicious activities in this subset. DrDoS attacks leverage LDAP vulnerabilities to overwhelm servers with malicious traffic, causing service disruptions and compromising network integrity. The subset includes instances of both benign and attack scenarios,

creating a controlled environment for training and evaluating machine learning models specifically tailored for the detection of LDAP-based threats. The focused nature of this subset allows for a deep dive into the intricacies of DrDoS attacks, enabling the development of targeted and effective detection mechanisms.

4.3 Data Preprocessing

Data preprocessing is a crucial step in preparing the CICIDS 2019 dataset, and specifically the DrDoS_LDAP_data_2_0_per subset, for machine learning analysis. The dataset contains critical features such as Source IP address and Destination IP address, which were initially represented as non-integer values for our machine learning model. To address this, the IP addresses were segmented into four distinct integers—Source IP1, Source IP2, Source IP3, and Source IP4—by employing a period ('.') as the delimiter. For instance, if the Source IP column contains the value 104.110.151.222, it is split into 104, 110, 151, and 222, respectively.
Certain columns deemed less significant were removed from the dataset, specifically those whose names concluded with "max" or "min." As a result, attributes such as 'Fwd Packet Length Max' and 'Fwd Packet Length Min' were excluded.
Subsequently, the data underwent standardization using the formula $\frac{X-\mu}{\sigma}$ for each column. Prior to standardization, rows containing exceptionally high values that impeded the standardization process were eliminated.
Notably, the dataset exhibited substantial bias, with 1,612 instances labeled as 'attacked' and 2,179,930 labeled as 'BENIGN'. To mitigate this imbalance, both undersampling for 'BENIGN' and oversampling for 'attacked' were implemented. This resulted in a balanced dataset containing 50,000 samples for each class. Both undersampling and oversampling techniques were implemented utilizing the 'resample' function from the 'sklearn' library.

5 Methodology

5.1 PCA and Eigenvalues

5.1.1 Principal Component Analysis (PCA)

In this section, Principal Component Analysis (PCA) was employed to analyze the data and extract principal components. The eigenvalues associated with these components were computed to assess their significance.

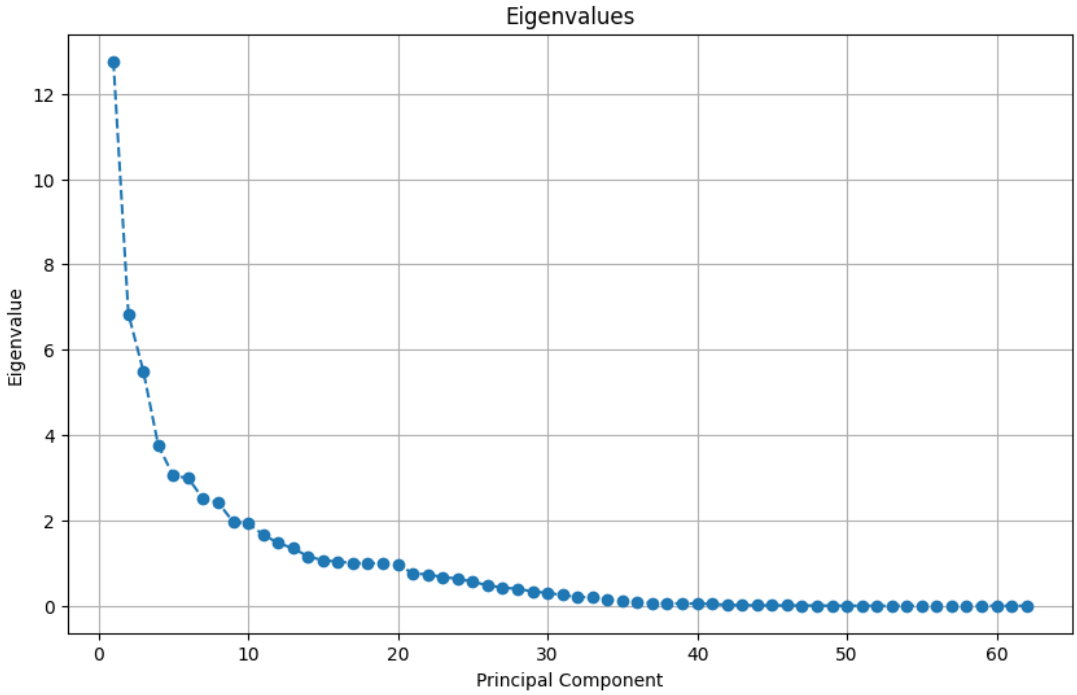
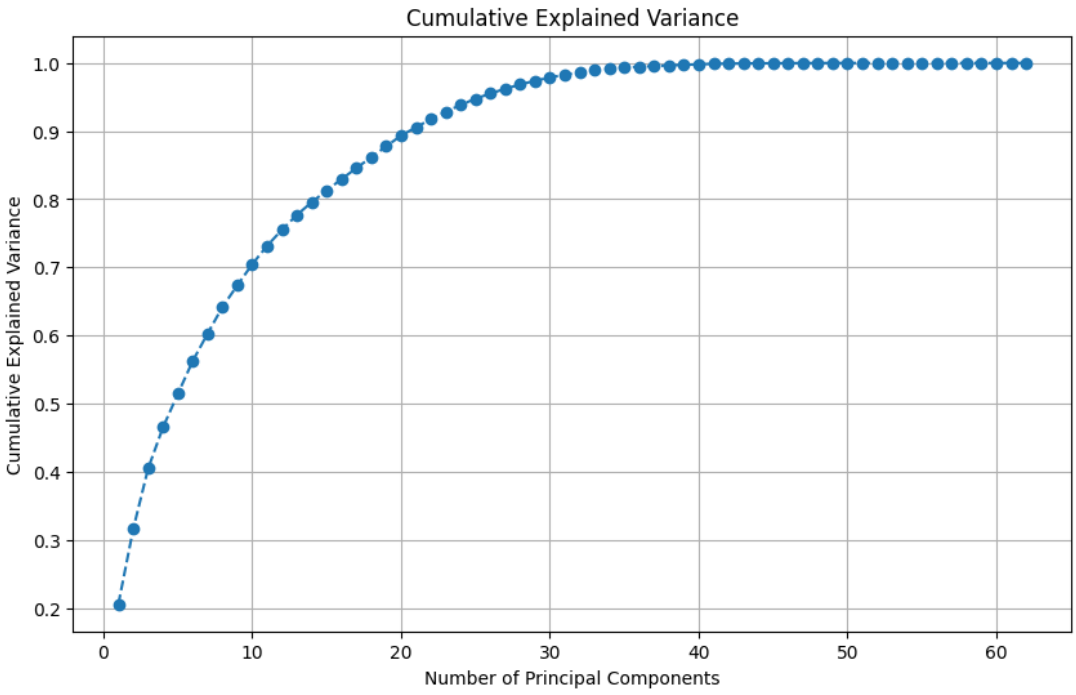


Figure 2: Eigen Values.

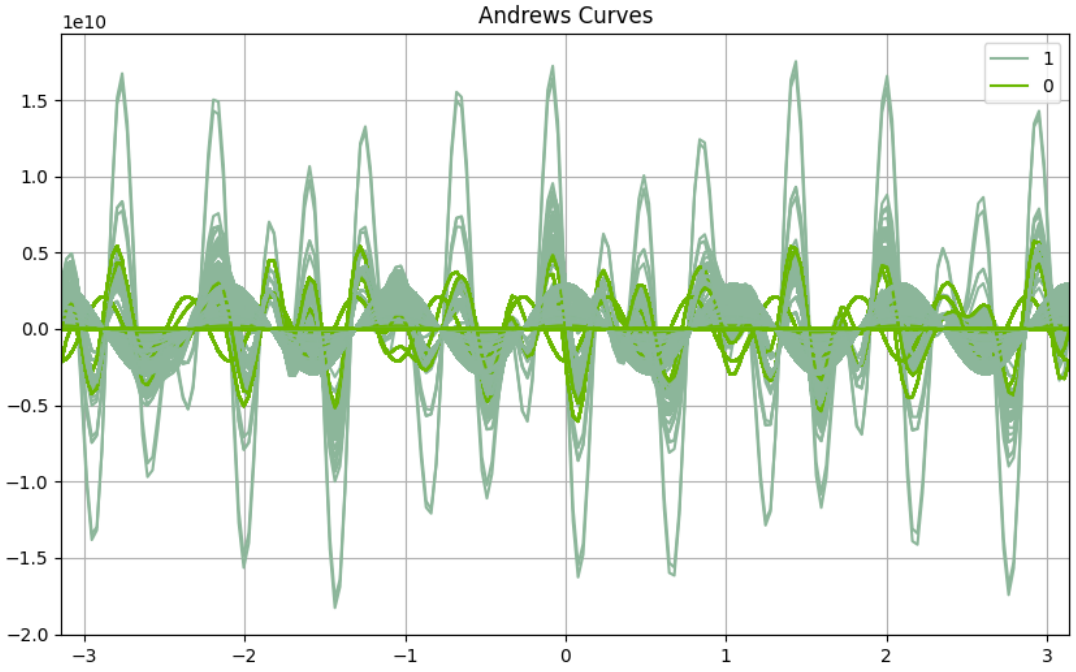


Analysis of the above graphs suggests that a subset of approximately 30 to 35 features holds significance in this dataset.

5.2 Andrew Curves

5.2.1 Introduction to Andrew Curves

Andrew Curves were utilized as a visualization technique to represent multivariate data in a 2-dimensional space. This method aids in identifying patterns and trends within the dataset.



Analysis of the above graphs tells that this dataset is linearly separable.

5.3 Model Selection and Training

Utilizing a Random Forest model, we identified crucial features and subsequently conducted training using the Logistic Regression algorithm.

5.4 SHAP

In our analysis, we employed SHAP (SHapley Additive exPlanations) to generate informative visualizations, including Waterfall Plot, Force Plot, Mean SHAP Plot, and Beeswarm Plot.

6 Results and Discussion

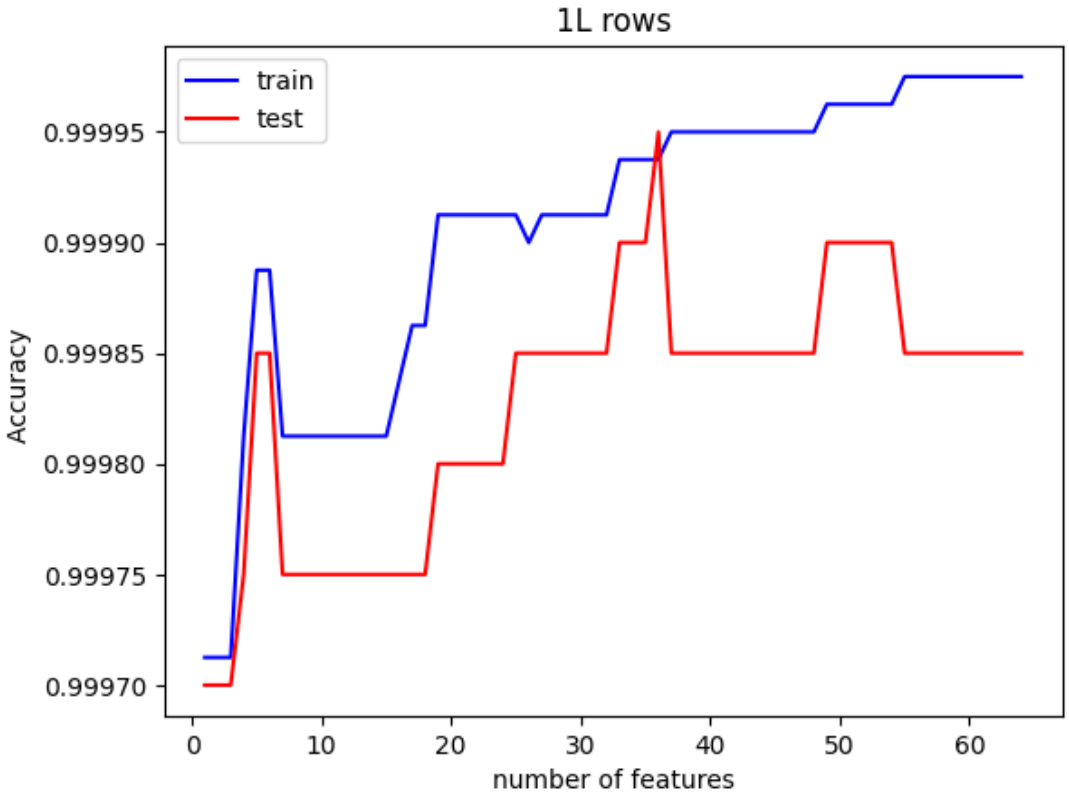
6.1 Performance of the Logistic Regression Model

Metric	Value
Model Accuracy	0.99972
F1 Score	0.9997

Table 1: Model Evaluation Metrics

	Predicted False	Predicted True
Actual False	10000	0
Actual True	5	9995

Table 2: Confusion Matrix



The preceding graph illustrates the fluctuation in accuracy between the training and test datasets as the number of features varies

6.2 Insights from XAI on Model Decisions

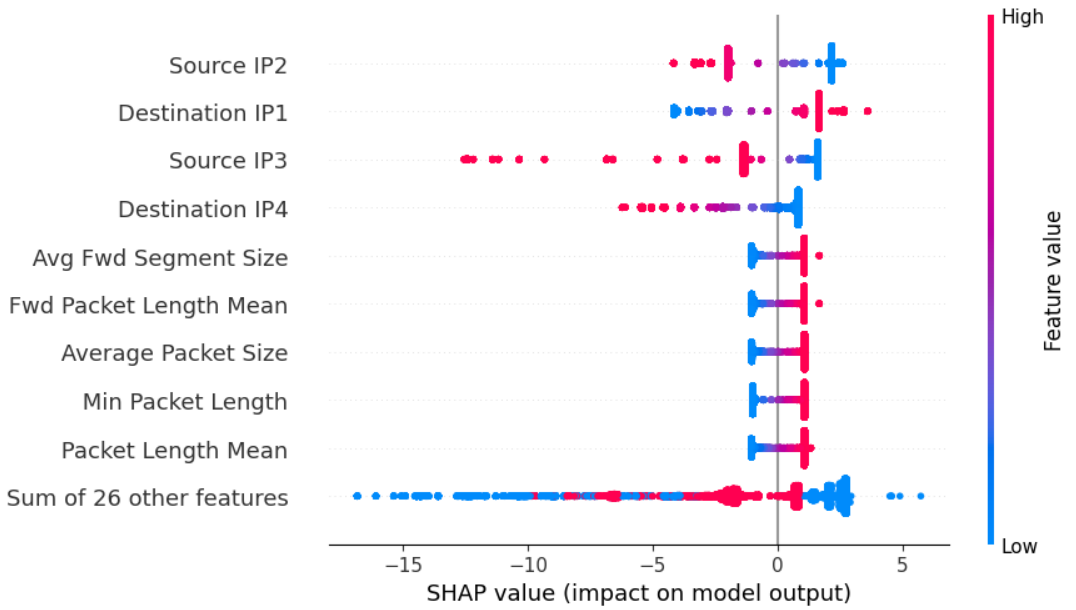


Figure 3: Beeswarm Plot.

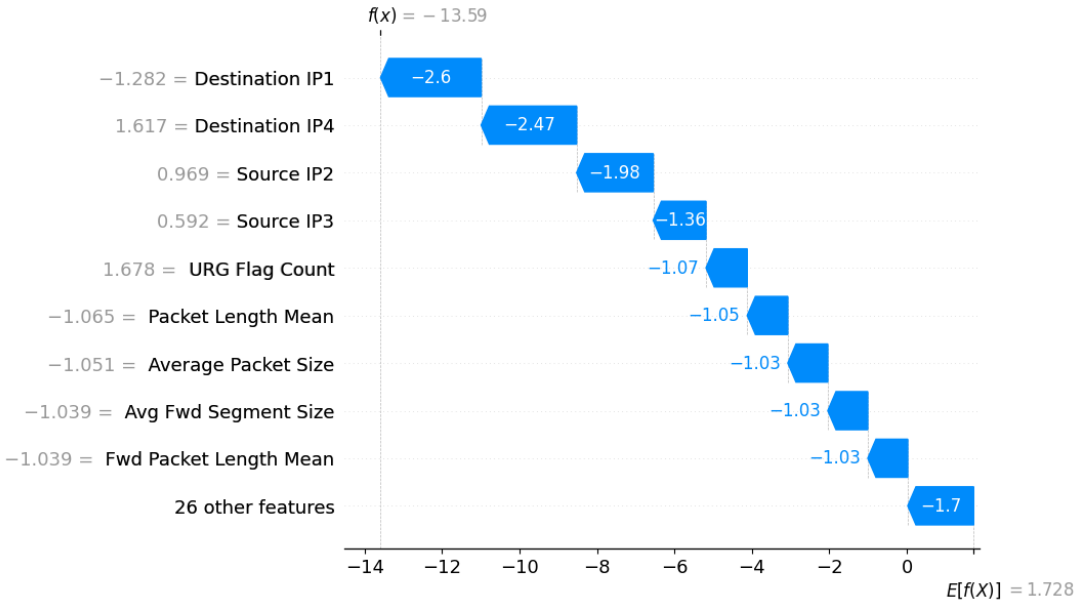


Figure 4: Waterfall Plot.

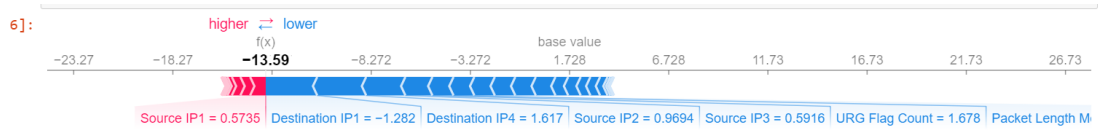


Figure 5: Force Plot.

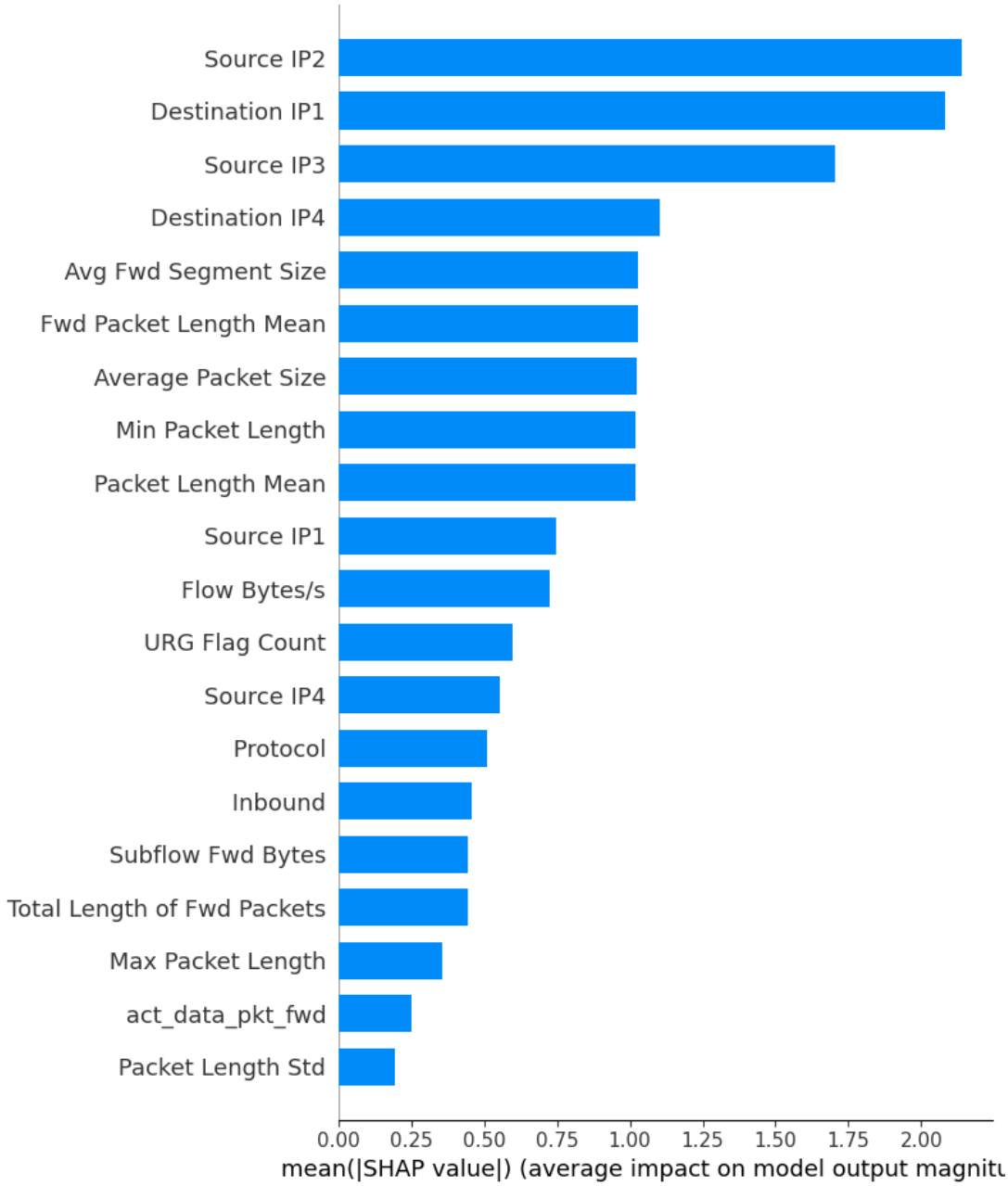


Figure 6: Mean SHAP Plot.

7 Conclusion

The alignment between the crucial features identified through the implementation of the Random Forest classifier and the top features highlighted by SHAP (SHapley Additive exPlanations) is of paramount significance. This congruence strengthens the reliability and interpretability of the model's decision-making process. The utilization of SHAP in conjunction with black-box models serves as a pivotal approach to shedding light on the intricacies of the model's inner workings.
In essence, the integration of XAI methodologies in black-box models serves as a powerful tool for enhancing the transparency and comprehensibility of machine learning systems. This approach not only aids in model validation but also empowers stakeholders to make informed decisions based on a clearer understanding of the model's behavior and the relevance of its identified features.

8 References

[1] Iman Sharafaldin, Arash Habibi Lashkari, Saqib Hakak, and Ali A. Ghorbani, "Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy", IEEE 53rd International Carnahan Conference on Security Technology, Chennai, India, 2019.

[2] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018.

[3] P. Kabiri and A. A. Ghorbani, "Dimension reduction and its effects on Clustering for Intrusion Detection," in Privacy, Intrusion Detection, and Response: Technologies for Protecting Networks, P. Kabiri, Ed. IGI Global, 2011.

[4] Journal of Defense Modeling and Simulation: Applications, Methodology, Technology 2022, Vol. 19(1) 57–106 The Author(s) 2020 DOI: 10.1177/1548512920951275

[5] Explainable Artificial Intelligence in CyberSecurity: A SurveyNICOLA CAPUANO1, GIUSEPPE FENZA2, (Member, IEEE), VINCENZO LOIA 2, (Senior Member, IEEE), AND CLAUDIO STANZIONE3, (Member, IEEE)

[6] Explaining Network Intrusion Detection System Using Explainable AI Framework Shraddha Mane1 and Dattaraj Rao2 Persistent Systems Limited, India

[7] Machine Learning and Deep Learning Methods for Cybersecurity YANG XIN1,2, LINGSHUANG KONG 3, ZHI LIU 2,3, (Member, IEEE), YULING CHEN2, YANMIAO LI1, HONGLIANG ZHU1, MINGCHENG GAO1, HAIXIA HOU1, AND CHUNHUA WANG4