Interpretability of Intrusion Detection Models: An Information Visualization Approach

Tiago Martins Ferreira
São Paulo State University (UNESP)
Bauru, Brazil
tiago.ferreira@unesp.br

Carlos Eduardo Silva Bertazzoli
São Paulo State University (UNESP)
Bauru, Brazil
carlos.bertazzoli@unesp.br

Thiago José Lucas
São Paulo State Univ. of Tech. (FATEC)
Ourinhos, Brazil
thiago@fatecourinhos.edu.br

Eduardo Alves Moraes
São Paulo State University (UNESP)
Bauru, Brazil
eduardo.moraes@unesp.br

Alessandra de Souza Lopes
São Paulo State University (UNESP)
Bauru, Brazil
alessandra.lopes@unesp.br

Luis Augusto de Campos Alves
São Paulo State University (UNESP)
Bauru, Brazil
luis.alves@unesp.br

***Abstract*—This article explores information visualization to enhance the interpretability of intrusion detection models, focusing on Machine Learning and Explainable Artificial Intelligence (XAI). Given the complexity of cyberattacks and the ”black-box” nature of many models, this work proposes the use of techniques such as SHAP and visualizations to make model decisions, such as those from Random Forest, more understandable. Using the CICIDS2017 dataset, the study aims to apply preprocessing, train the model, interpret its decisions with SHAP, and generate explanatory visualizations. The objective is to increase confidence and adoption of intrusion detection systems, making them more transparent and auditable for security analysts. Results showed that the Random Forest model achieved an accuracy of 99.9%, indicating its high capability to distinguish between benign and malicious network traffic. More importantly, SHAP visualizations, including importance, summary, and dependence plots, provided valuable insights into model behavior.**

***Keywords*—Intrusion Detection, Machine Learning, Explainable Artificial Intelligence (XAI), Information Visualization, Cybersecurity.**

I. Introduction

In recent decades, the advancement of digital technology has driven a rapid expansion of connectivity among devices, exposing networks and systems to a variety of cyber threats. Complex and hard-to-detect attacks, such as zero-day exploits, have challenged traditional signature-based protection methods. At the same time, the increasing volume of traffic in corporate and critical networks demands solutions that operate efficiently and in real time. Conventional techniques, such as firewalls and antivirus software, are no longer sufficient given the current sophistication of attacks [1]. In vulnerable contexts, such as hospital networks or critical infrastructures, a detection failure can have severe consequences, impacting data integrity and human safety [2]. In this scenario, intrusion detection emerges as an essential component for mitigating risks and protecting digital assets. However, many Intrusion Detection Systems (IDS) still face challenges related to accuracy, scalability, and the ability to operate effectively in dynamic and highly complex environments [3].

Given these limitations, machine learning has gained prominence as a promising alternative to enhance the effectiveness of IDS systems. Algorithms such as Random Forest, SVM (Support Vector Machine), and deep neural networks have been widely employed to classify network traffic as normal or malicious based on patterns extracted from large volumes of data [1]. Despite achieving high accuracy, many of these models require significant computational power and face issues such as real-time latency and class imbalance [2]. Furthermore, complex models such as recurrent or transformer-based networks encounter additional interpretability challenges [1]. Consequently, there is a growing research need for lighter, more robust models suited for operational environments, such as Random-Forest, which aims to balance performance and response time [1]. In this context, the use of machine learning in intrusion detection represents progress; however, its applicability depends on overcoming obstacles related to transparency, understanding, and explanation of results by analysts.

Although machine learning models have revolutionized intrusion detection, their use in real-world environments remains limited due to their inherent lack of interpretability. In many cases, these algorithms act as “black boxes”, where only the inputs and outputs are visible, while the internal decision-making process remains inaccessible to the user [4]. This lack of interpretability undermines analyst trust and hinders adoption in environments that demand traceability and clear justifications, such as governmental or critical infrastructure networks [5], [6]. In this scenario, Explainable Artificial Intelligence (XAI) emerges as a complementary approach to traditional machine learning, promoting greater transparency in systems. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) have been applied to explain, globally or locally, the influence of variables on model decisions [2], [7]. Despite the growing adoption of XAI, its integration into IDS remains incipient [5]. Thus, the demand for explainable systems, also known as X-IDS, highlights the need for solutions that combine accuracy and auditability [6].

In this context, information visualization emerges as a strategic tool to support the interpretability of intrusion detection models. By leveraging human capabilities such as visual perception and pattern recognition, visualizations enable the extraction of meaning from complex datasets that are difficult to interpret in tabular form [8]. Interactive visual interfaces allow analysts to explore network traffic behaviors, identify anomalies, and understand model decisions more intuitively [9]. Despite advances in this direction, the combination of visual analytics and predictive algorithms still lacks in-depth exploration [9]. Initiatives such as the IMPAVID system demonstrate that visual analysis can enhance situational awareness and support decision-making in regulatory and technical contexts [10]. Thus, when combined with XAI techniques, information visualization enables security professionals to translate model outcomes intuitively, fostering greater acceptance and operational efficiency.

In light of this scenario, the following research question arises: despite the effectiveness of machine learning models in intrusion detection, their “black-box” nature makes it difficult to understand the decision-making process, limiting trust and adoption by security analysts. How can these models be made more interpretable and transparent without compromising accuracy? The rationale for this study lies in the high complexity of cyberattacks and the significant increase in network traffic, factors that intensify the challenge of intrusion detection. Although algorithms such as Random Forest achieve high performance in identifying malicious behaviors, their applicability in critical environments depends on mechanisms that make their decisions comprehensible and visual. In this sense, XAI techniques, such as SHAP, combined with information visualization, offer valuable resources to enhance analysis and facilitate model interpretation by security professionals. Based on these considerations, this work aims to explore the use of information visualization and interpretability techniques to improve understanding and trust in intrusion detection systems based on Random Forest and the CICIDS2017 dataset. To this end, the objectives include: applying preprocessing techniques to the CICIDS2017 dataset; training an intrusion detection model using the Random Forest algorithm; applying the SHAP technique to interpret model decisions globally and locally; generating explanatory visualizations of the most relevant features for model decisions; and finally, assessing how information visualization contributes to understanding model behavior and network traffic patterns.

This article is organized as follows: Section 1 presents the introduction, including the context, research problem, rationale, and objectives. Section 2 discusses the theoretical foundation. Section 3 describes the materials and methods employed. Section 4 presents the results and their analysis. Finally, Section 5 provides conclusions and suggestions for future work.

II. Theoretical Background

The theoretical background of this study is grounded in four main pillars: the concepts and principles of information visualization, the fundamentals of Explainable Artificial Intelligence (XAI) and its relationship with visualization, visual techniques applied to the interpretability of machine learning models, and finally, the application of these approaches in the context of cybersecurity. These pillars support the proposal of using visual and interpretive resources to enhance understanding and trust in intrusion detection models, particularly in critical environments where system transparency and effectiveness are essential.

A. Information Visualization: Concepts and Principles

Information visualization is a field dedicated to the graphical representation of complex data with the goal of facilitating analysis and comprehension. By leveraging human visual perception as a tool, visualization enables the efficient interpretation of large volumes of data, allowing the identification of patterns, correlations, and anomalies that would hardly be detected in textual or tabular formats [8]. This approach is particularly relevant in contexts such as network security, where analysts must make rapid decisions based on large amounts of simultaneous information. The effectiveness of a visualization, however, depends on its clarity and well-defined purpose. Overloaded visualizations, with multiple objectives or excessive information, tend to confuse rather than clarify. Therefore, it is recommended that each visual representation have a specific motivation and utilize resources such as interactivity or parallel visualizations to convey complementary information [8]. Information visualization, therefore, not only represents data but serves as a strategic means of analytical reasoning, supporting users in generating insights from multivariate and highly complex datasets.

B. Explainable Artificial Intelligence (XAI) and Visualization

Explainable Artificial Intelligence (XAI) emerges as a response to the growing adoption of machine learning models in critical domains, where understanding the decision-making process is as important as achieving accuracy. XAI seeks to enhance the transparency of complex models, enabling users to understand the factors that drive specific decisions [11]. Techniques such as SHAP and LIME have been widely employed in this context. SHAP leverages game theory concepts to calculate the contribution value of each feature to a prediction, both on a global and local scale, clearly identifying which attributes influenced a decision [12]. LIME, on the other hand, constructs a simple local model to explain a specific prediction, approximating the behavior of the original model within a defined neighborhood [2]. When these techniques are combined with interactive visualizations, systems become more comprehensible, allowing analysts to graphically explore decision factors. This association between XAI and visualization reinforces trust, facilitates auditing, and improves decision-making processes in environments that rely on accurate interpretation of models [9], [13].

C. Visualization Techniques for Model Interpretability

The interpretability of machine learning models can be enhanced through visualization techniques that transform abstract information into user-accessible representations. Methods such as scatter plots, feature importance diagrams, and heatmaps allow for a more intuitive observation of model behavior. When combined with techniques such as SHAP and LIME, these visualizations emphasize the variables that most influence decisions, as well as the direction and magnitude of their contributions [2]. For example, dependence plots generated by SHAP illustrate how variations in feature values impact predictions, while summary plots provide a consolidated view of feature importance. These graphical tools are particularly useful when models involve high-dimensional or heterogeneous data, enabling a clearer analysis of the internal interactions within the algorithm [12]. Therefore, the use of visual techniques does not replace statistical methods but rather complements them by providing an interpretable interface between the mathematical logic of models and expert-driven analysis.

D. Applications of Visualization in Cybersecurity

In the field of cybersecurity, information visualization plays an essential role in supporting threat detection and analysis. With the increasing complexity of networks and the large volume of data generated by sensors, logs, and detection systems, analysts face the challenge of synthesizing dispersed information for rapid and effective decision-making. The application of visualizations enables monitoring of packets and network flows, identification of suspicious patterns, mapping of vulnerabilities, and exploratory incident analysis [14]. Visual tools make it possible, for instance, to correlate events dispersed over time and space, facilitating the understanding of ongoing attacks. Studies have shown that visual approaches can reveal trends and complex relationships that might go unnoticed in traditional analyses [8]. Furthermore, initiatives such as the use of visual analytics applied to compliance processes in incident management demonstrate the potential of visualization to guide regulatory and technical decisions [10]. Thus, the integration of visualization and information security enhances the situational awareness of professionals in the field, making defense processes more effective and contextually informed.

E. Related Work

The study by [15], titled “Explainable AI for Comparative Analysis of Intrusion Detection Models”, aligns with the present work by exploring Explainable Artificial Intelligence to improve the understanding of intrusion detection models. Both studies aim to make machine learning model decisions more transparent, a crucial point for trust and adoption in cybersecurity environments. The main difference lies in the methodology: while the authors employ occlusion sensitivity and various machine learning models on the UNSW-NB15 dataset for comparative analysis, the present work focuses on applying SHAP and information visualization to a Random Forest model using the CICIDS2017 dataset. This methodological distinction allows the current study to deepen the interpretability of a specific model, Random Forest, which is central to intrusion detection.

The study “XAI-XGBoost: an innovative explainable intrusion detection approach for securing internet of medical things systems” by [2] shares with this work the focus on applying XAI to IDS systems, using techniques such as SHAP and LIME to provide insights into model predictions. Both articles recognize the fundamental need for interpretability in critical cybersecurity environments. However, the work of [2] differs by focusing on the Internet of Medical Things (IoMT) context and employing the XGBoost classifier, optimized with sampling and feature selection techniques on the WUSTL-EHMS-2020 dataset. In contrast, the present study investigates the interpretability of Random Forest models in traditional networks, using the CICIDS2017 dataset, which enables a more in-depth analysis of the characteristics and challenges specific to this type of network.

The article “IDRandom-Forest: Advanced Random Forest for Real-Time Intrusion Detection” by [1] is relevant to the present work because both focus on the Random Forest algorithm for intrusion detection. The similarity lies in the pursuit of improving the effectiveness of IDS systems based on Random Forest. However, the work by [1] distinguishes itself by proposing an advanced Random Forest, “IDRandom-Forest”, aimed at reducing testing time and increasing accuracy for real-time detection, incorporating techniques such as accuracy sliding windows and feature weighting. The present study, in turn, complements this approach by focusing on the interpretability of Random Forest, using SHAP and information visualization to explain model decisions, an aspect not directly addressed in the cited article but crucial for adoption in real-world environments.

The work of [7], "From explanations to feature selection: assessing SHAP values as a feature selection mechanism", is an important related study as it explores the SHAP technique, which is central to the present work. Both studies acknowledge SHAP's potential beyond explainability. The similarity lies in the use of SHAP to understand feature importance. The main difference, however, is that the authors investigate SHAP as a feature selection mechanism, assessing its effectiveness in various classification and regression tasks. In contrast, the present study employs SHAP to interpret the decisions of a Random Forest intrusion detection model, generating explanatory visualizations of the most relevant features to improve model understanding and trust, rather than primarily for feature selection.

The article by [5], "Explainable artificial intelligence models in intrusion detection systems" provides a comprehensive review of XAI models in IDS, making it an important related work. Both studies address the importance of XAI in overcoming the opacity of machine learning models in cybersecurity. The similarity lies in the exploration of techniques such as LIME and SHAP to increase transparency. The distinction, however, is that the cited work is a systematic review summarizing the state of the art, challenges, and opportunities of XAI in IDS. At the same time, the present study is an applied research focusing on the practical implementation of SHAP and information visualization to interpret a specific Random Forest model, offering a significant and visual contribution to the field.

The study by [8], "Effective Data Visualization in Cybersecurity", is a relevant related work as it addresses the importance of data visualization in cybersecurity, a fundamental pillar of the present project. Both studies converge on the premise that visual representation can transform large volumes of complex data into actionable insights for security analysts. The similarity lies in valuing human perception to identify patterns and anomalies. However, the authors provide a theoretical and technical overview of visualizations in network security, discussing challenges and techniques. The present study goes beyond the theoretical discussion by applying information visualization in conjunction with XAI techniques (specifically SHAP) to interpret the decisions of an intrusion detection model, providing a practical approach focused on the interpretability of machine learning models, distinguishing it from the work of [8].

III. Materials and Methods

This chapter details the methodology employed to explore information visualization applied to the interpretability of intrusion detection models. The approach encompassed data preprocessing, machine learning model training, and the application of interpretability and visualization techniques. The tools and libraries used were selected based on their effectiveness and relevance for each phase of the process[1], with particular focus on visualization capabilities to enhance result comprehension.

A. Dataset

The dataset used in this study was CICIDS2017 (Canadian Institute for Cybersecurity Intrusion Detection Dataset 2017). This dataset is widely recognized in the cybersecurity community for its comprehensiveness and realism, containing network traffic simulating common attacks such as DoS, DDoS, Brute Force, XSS, SQL Injection, among others, as well as benign traffic. CICIDS2017 consists of a large volume of network flow data, with 79 features and over 3 million records, making it a valuable resource for training and evaluating intrusion detection systems. The choice of this dataset is justified by its capacity to represent complex and dynamic network scenarios, essential for developing robust and interpretable models.

B. Data Preprocessing

The data preprocessing stage was crucial to ensure the quality and suitability of the dataset for machine learning model training. This phase was implemented using Python, leveraging the Pandas library for data manipulation and analysis. The main operations performed included:

- **Loading and Initial Cleaning:** The raw dataset was loaded, and columns were renamed to remove spaces and special characters, facilitating programmatic access. Missing (NaN) and infinite values were identified and removed, ensuring numerical integrity of the data.
- **Type Conversion:** All columns, except for the label column `Label`, were converted to numeric types. Values that could not be converted were treated as NaN and subsequently removed.
- **Removal of Low-Variance Columns:** Columns with very low variance (i.e., nearly constant values) were removed. These features contribute little to the model's predictive capability and may introduce unnecessary noise.
- **Min-Max Normalization:** After cleaning and feature selection, numerical data were normalized using Min-Max Scaling. This technique scales feature values to a fixed range (usually between 0 and 1), which is critical for

[1] Source codes are available at: https://git.tmferreira.tec.br/tiago.ferreira/cicids2017-visualization

distance-based algorithms and to prevent features with large scales from dominating the training process.

The outcome of this stage was a preprocessed dataset ready for training the intrusion detection model.

C. Intrusion Detection Model Training

For the intrusion detection task, the Random Forest algorithm was selected, implemented via the Scikit-learn library in Python. Random Forest is a tree-based machine learning algorithm that stands out for its robustness, ability to handle large volumes of data and high dimensionality, and lower susceptibility to overfitting compared to individual decision trees. The choice of Random Forest is justified by its proven performance in classification problems and its ability to provide a measure of feature importance, which is relevant for interpretability.

The training process involved:

- **Data Splitting:** The preprocessed dataset was divided into training and test sets, with a 70%–30% split. Stratification was applied to ensure that class distribution (benign vs. attack) was maintained across both sets, which is particularly important in imbalanced datasets such as CICIDS2017.
  - **Model Configuration:** The Random Forest model was configured with 100 estimators (decision trees), and the parameter `class_weight="balanced"` was used to mitigate class imbalance by assigning higher weights to minority classes during training.
  - **Model Evaluation:** After training, the model was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. The confusion matrix was also generated to provide a detailed view of the model's performance for each class.

The trained model was saved for subsequent use in the interpretability phase.

D. Interpretability and Visualization

The model interpretability was addressed using the SHAP technique, a game-theory-based approach that assigns each feature an importance value for a specific prediction. SHAP is model-agnostic, meaning it can be applied to any machine learning model, and provides both global explanations (overall feature importance) and local explanations (contribution of each feature to an individual prediction).

Visualizations were generated using the Matplotlib and Seaborn libraries in Python, which are widely used for creating high-quality statistical plots. The visualizations produced included:

- **Feature and Class Distribution (Raw and Preprocessed Data):** Histograms and count plots were generated to visualize the distribution of specific features (e.g., `Destination_Port`) and class distribution (benign vs. attack) in both the original and preprocessed datasets. These visualizations are fundamental to understanding the data composition and the impact of preprocessing steps.
- **SHAP Importance Plots (Bar and Summary):** SHAP generated bar plots showing the average importance of each feature for the model, and summary plots illustrating the distribution of SHAP values for each feature, revealing how each one influences the model output (positively or negatively).
- **SHAP Dependence Plot:** For the most important feature identified by SHAP (in this case, `Destination_Port`), a dependence plot was generated. This plot illustrates the relationship between the value of the `Destination_Port` feature and the corresponding SHAP value, revealing how different destination ports influence model predictions. The point dispersion and color indicate potential interactions with other features, providing insights into the model's behavior regarding this critical feature.

All visualizations were saved in image format (‘.png‘) for inclusion in the Results chapter, allowing a clear and concise visual analysis of model interpretability and data characteristics.

IV. Results

This chapter presents the results obtained from applying the methodology described in Section 3, focusing on the analysis of the CICIDS2017 dataset, the performance of the Random Forest model for intrusion detection, and, most importantly, model interpretability through visualization techniques. The results demonstrate the effectiveness of the proposed approach in providing insights into model behavior and feature importance, contributing to a deeper understanding of intrusion detection systems.

A. Analysis of the CICIDS2017 Dataset

Preprocessing the CICIDS2017 dataset was a crucial step to prepare the data for model training. Initially, the raw dataset, containing over 3 million rows and 79 columns, was loaded. The class distribution in the original dataset, observed prior to cleaning, revealed significant imbalance, with the `BENIGN` class (normal traffic) being predominant. After removing rows with NaN or infinite values and eliminating low-variance columns, the dataset was reduced to 3,053,587 rows and 71 columns. Min-Max normalization was applied to scale feature values to a consistent range.

To illustrate the dataset composition, visualizations of class distribution and a representative feature, `Destination_Port`, were generated for the raw data.

Figure 1 presents the class distribution in the original dataset, highlighting a strong imbalance: benign traffic accounts for the vast majority of records. At the same time, different attack types appear in significantly smaller proportions:

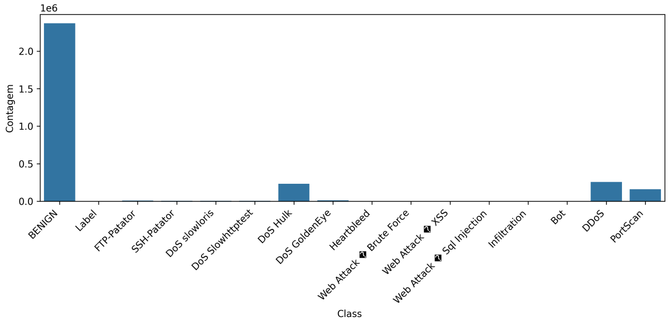

Fig. 1. Class Distribution in the Raw Dataset.

Figure 2 shows the distribution of the `Destination_Port` feature, emphasizing the predominance of connections to a restricted set of ports. This concentration suggests the existence of typical traffic patterns, useful for both exploratory analysis and predictive modeling:

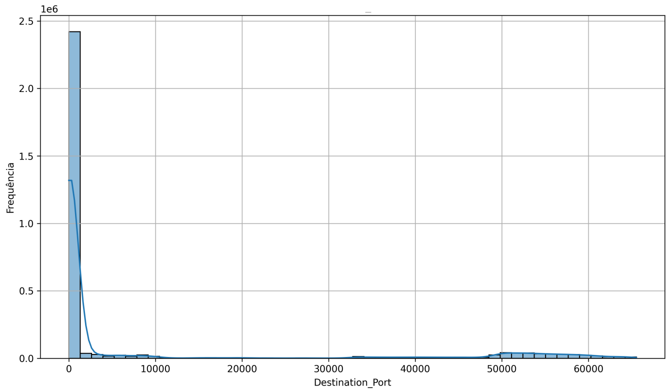

Fig. 2. Distribution of the `Destination_Port` Feature in the Raw Dataset.

After preprocessing, visualizations were generated to assess the integrity of the processed data. Figure 3 shows the class distribution in the training and test sets after the stratified split of the preprocessed dataset. Although benign traffic remains predominant, the proportion between classes was preserved in both subsets. This strategy ensures fair model evaluation while maintaining representativity of both classes during training and validation.

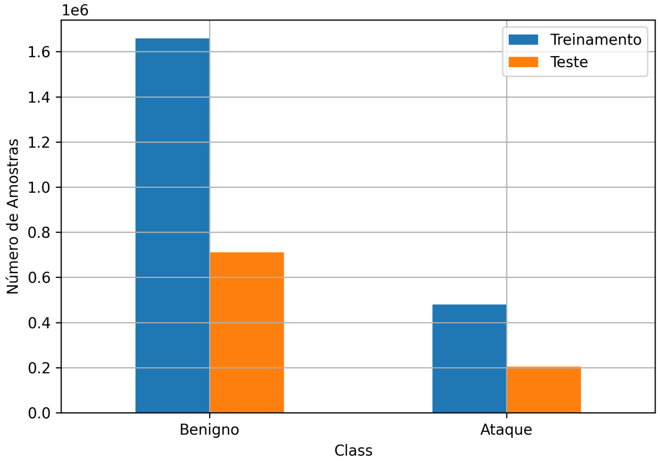

Fig. 3. Class Distribution in the Preprocessed Dataset.

Figure 4 displays the normalized `Destination_Port` feature distribution, preserving the concentration pattern observed in the raw data. This preservation indicates that normalization did not significantly distort the original characteristic of the variable:

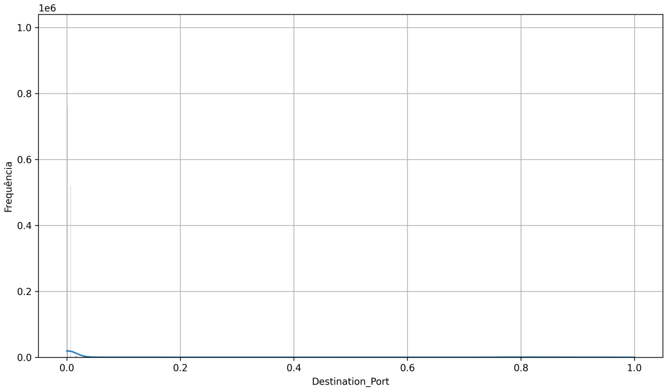

Fig. 4. Distribution of the `Destination_Port` Feature in the Preprocessed Dataset.

B. Random Forest Model Performance

The Random Forest model was trained using the processed dataset, and its performance was evaluated on a separate test set. The obtained results were significant, as shown in Table I. Evaluation metrics — accuracy, precision, recall, and F1-score — demonstrate the model's robust performance, with a high ability to distinguish between benign and malicious traffic.

Table I
Evaluation Metrics of the Random Forest Model.

Metric	Value
Accuracy	0,9990
Precision	0,9979
Recall	0,9978
F1-score	0,9978

Table II presents the confusion matrix generated from the model's predictions, providing a detailed view of correct and incorrect classifications. The values demonstrate that the model correctly classified the vast majority of instances, with very few false positives and false negatives.

Table II
Confusion Matrix of the Random Forest Model.

	Predicted Benign	**Predicted Attack**
Actual Benign	710.271	431
Actual Attack	455	204.920

C. *Model Interpretability with SHAP*

The application of the SHAP technique was essential to understand how the Random Forest model makes decisions and which features are most influential. Three types of SHAP plots were generated to visualize feature importance and impact:

1) SHAP Importance Plot (Bar): The bar plot of feature importance, shown in Figure 5, summarizes the average importance of each feature for the model. Features are ordered from most to least important, providing a global view of which network traffic attributes are most relevant for intrusion detection. As expected, the `Destination_Port` feature stood out as the most important, followed by other features related to network flow.

2) SHAP Summary Plot: The SHAP summary plot, displayed in Figure 6, provides a more detailed view of the distribution of SHAP values for each feature. Each point represents a dataset instance, and color indicates the feature value (red for high values, blue for low values). This plot allows observation not only of the overall feature importance but also how different feature values impact model predictions (positive SHAP values indicate contribution to predicting attack, while negative values contribute to predicting benign).

3) SHAP Dependence Plot: For the most important feature, `Destination_Port`, a SHAP dependence plot was generated, shown in Figure 7. This plot illustrates the relationship between the `Destination_Port` feature value and the corresponding SHAP value, revealing how different destination ports influence model predictions. Point dispersion and coloring indicate possible interactions with other features, providing insights into model behavior for this critical feature.

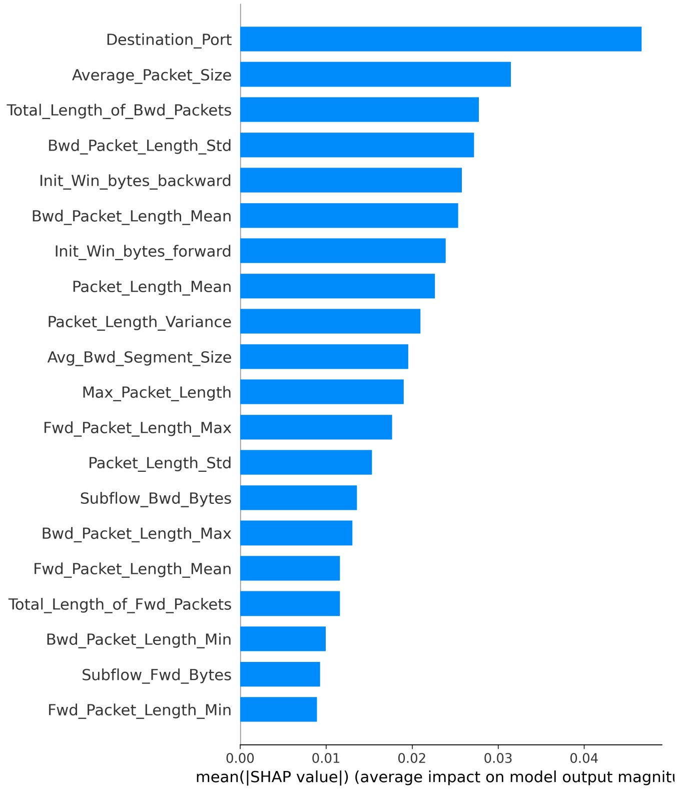

Fig. 5. SHAP Importance Plot (Bar).

In summary, the visualizations generated with SHAP provide a clear and intuitive understanding of the internal logic of the Random Forest model, enabling security analysts to identify which network traffic characteristics most contribute to intrusion detection. This interpretability is crucial for building trust in the system and supporting informed decision-making in cybersecurity environments.

Figure 6 shows a SHAP summary plot representing the impact of different network traffic features on the output of a machine learning model. Each point corresponds to an instance in the dataset, with the horizontal position indicating the SHAP value — i.e., the contribution of that feature to the model's prediction. Features are ranked by overall importance, with *Destination_Port* appearing as the most influential, followed by *Average_Packet_Size* and *Total_Length_of_Bwd_Packets*.

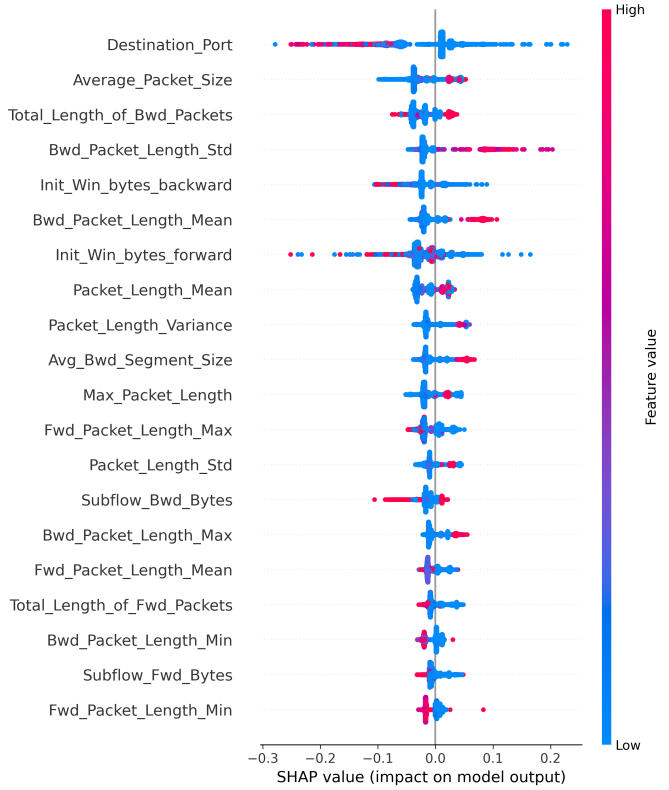

Fig. 6. SHAP Summary Plot.

In Figure 7, it is possible to observe a SHAP dependency graph for the *Destination_Port* feature:

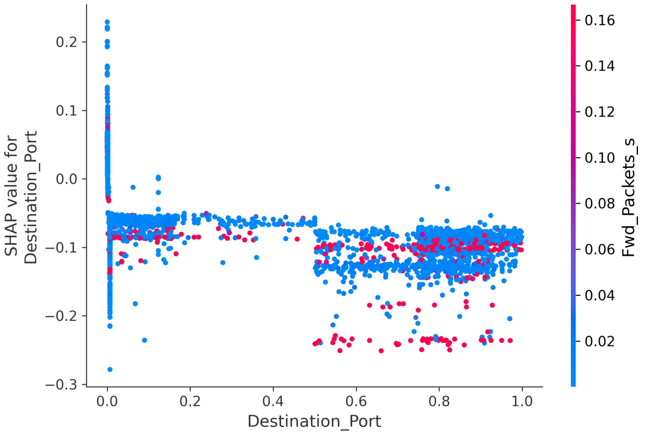

Fig. 7. SHAP Dependence Plot for `Destination_Port`.

V. Conclusion

This study explored the application of information visualization and interpretability techniques to enhance the understanding of machine learning-based intrusion detection models. Using the CICIDS2017 dataset and the Random Forest algorithm, it was demonstrated that high attack detection performance can be achieved while making the model's decision-making process transparent and understandable to security analysts.

The adopted methodology, including data preprocessing, training a robust model, and applying SHAP for interpretability, proved effective. Results showed that the Random Forest model achieved an accuracy of 99.9%, indicating its high capability to distinguish between benign and malicious network traffic. More importantly, SHAP visualizations, including importance, summary, and dependence plots, provided valuable insights into model behavior. The `Destination_Port` feature was identified as the most influential for intrusion detection, and the contribution of its values and other features to model predictions was clarified.

The main contribution of this work lies in the practical demonstration of how combining machine learning, interpretability (XAI), and information visualization can overcome the ”black-box” challenge in intrusion detection systems. By providing clear visual explanations, the proposed approach increases model trust, facilitates incident analysis, and supports decision-making. Understanding why the model classifies a specific traffic as malicious is fundamental in cybersecurity environments, where both accuracy and justification of actions are critical.

Future work may explore other XAI techniques, such as LIME, to compare generated explanations and obtain a more comprehensive view of model behavior. Additionally, applying the proposed approach to other intrusion detection datasets and machine learning algorithms, such as deep neural networks, could reveal new insights and generalize the results presented here. Developing interactive dashboards that integrate real-time interpretability visualizations would also represent a significant advancement, allowing security analysts to monitor and understand threats more dynamically.

Acknowledgments

The authors are grateful to State Center for Technological Education “Paula Souza” (CEETEPS) for research conditions at the Defensive Cybersecurity and Artificial Intelligence Laboratory (Detect.AI), and to Research Support Foundation of São Paulo State (FAPESP), Brazil grants #2023/12830-0.

References

[1] M. Azhar, S. Perveen, A. Iqbal, and B. Lee, “Idrandom-forest: Advanced random forest for real-time intrusion detection,” *IEEE Access*, vol. 12, pp. 113 842–113 854, 2024.

[2] Y. Hosain and M. Çakmak, "Xai-xgboost: an innovative explainable intrusion detection approach for securing internet of medical things systems," *Scientific Reports*, vol. 15, no. 1, p. 22278, 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-07790-0

[3] S. Reynaud and A. Roxin, "Review of explainable artificial intelligence for cybersecurity systems," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 78, 2025. [Online]. Available: https://doi.org/10.1007/s44163-025-00318-5

[4] C. Molnar, *Interpretable Machine Learning*, 3rd ed., 2025. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[5] S. AL and S. Sagiroglu, "Explainable artificial intelligence models in intrusion detection systems," *Engineering Applications of Artificial Intelligence*, vol. 144, p. 110145, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197625001459

[6] S. Quincozes, C. Albuquerque, D. Passos, and D. Mossé, "A survey on intrusion detection and prevention systems in digital substations," *Computer Networks*, vol. 184, p. 107679, 01 2021.

[7] W. E. Marcílio and D. M. Eler, "From explanations to feature selection: assessing shap values as feature selection mechanism," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2020, pp. 340–347.

[8] C. N. Adams and D. H. Snider, "Effective data visualization in cybersecurity," in *SoutheastCon 2018*, 2018, pp. 1–8.

[9] S. Miksch, C. Di Ciccio, P. Soffer, and B. Weber, "Visual analytics meets process mining: Challenges and opportunities," *IEEE Computer Graphics and Applications*, vol. 44, no. 6, pp. 132–141, 2024.

[10] A. Palma and M. Angelini, "Impavid: Enhancing incident management process compliance assessment with visual analytics," *Computers and Graphics*, vol. 130, p. 104243, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0097849325000846

[11] C. K. I, B. A, M. B, C. V, and D. R. N, "Explaining aha! moments in artificial agents through ike-xai: Implicit knowledge extraction for explainable ai," *NEURAL NETWORKS*, vol. 155, pp. 95–118, 2022.

[12] S. Jagatheesaperumal, V. Pham, R. Ruby, Z. Yang, C. Xu, and Z. Zhang, "Explainable ai over the internet of things (iot): Overview, state-of-the-art and future directions," *IEEE Open Journal of the Communications Society*, vol. PP, pp. 1–1, 01 2022.

[13] MixMode, "The imperative of explainability in ai-driven cybersecurity," https://mixmode.ai/blog/the-imperative-of-explainability-in-ai-driven-cybersecurity/, 2024, publicado em 5 de setembro de 2024. Acesso em: 13 jul. 2025.

[14] H. Shiravi, A. Shiravi, and A. A. Ghorbani, "A survey of visualization systems for network security," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 8, pp. 1313–1329, 2012.

[15] P. Corea, Y. Liu, J. Wang, S. Niu, and H. Song, "Explainable ai for comparative analysis of intrusion detection models," 06 2024.