Malicious Network Traffic Detection

Vincent Carter
University of Colorado Boulder
Boulder CO, USA
Cybersecurity Network Analysis
vincent.carter@colorado.edu

Leanica Tiongson
University of Colorado Boulder
Boulder CO, USA
Cybersecurity Network Analysis
leti3086@colorado.edu

***Abstract* —In this comprehensive study, we have endeavored to enhance the capabilities of Intrusion Detection Systems (IDS) by employing a multi-faceted machine learning approach to analyze network traffic and detect anomalies. Utilizing the well-curated CICIDS2017 dataset, which features labeled real-world attack scenarios, our research undertook a systematic application of advanced machine learning techniques to address the complexities of network intrusion detection.**

Our final analysis included a detailed evaluation of several algorithms, including K-Means Clustering, Decision Trees, and Neural Networks, each chosen for their potential to effectively identify subtle and overt cybersecurity threats. Throughout our exploration, we faced significant challenges that impacted the direction and outcomes of our research. Notably, we encountered suspiciously high accuracy rates in some of our models, suggesting potential data leakage, which prompted a thorough revision of our data handling and model validation processes. Additionally, the performance of unsupervised learning models was initially poor, highlighting the difficulties in tuning these models to effectively capture the nuanced behaviors characteristic of sophisticated network intrusions.

Despite these setbacks, our iterative approach allowed us to refine our strategies and ultimately improve our models. Neural Networks, with their deep learning capabilities, provided the most robust detection rates, balancing precision and recall effectively. Moreover, our study ventured beyond mere algorithmic application; it also refined feature selection and data preprocessing to enhance model accuracy and computational efficiency.

Conclusively, our research not only advances the field of cybersecurity by providing actionable insights into IDS enhancement but also sets the groundwork for future explorations into adaptive learning models capable of countering evolving network threats. The implications of this work extend to real-world applications, where maintaining the integrity and security of network systems is paramount.

Keywords — Network Traffic Analysis, Anomaly Detection, CICID Dataset, Cybersecurity, Machine Learning

I. Introduction

Intrusion Detection Systems (IDS) are vital in safeguarding network infrastructures against an ever-evolving landscape of cyber threats. Traditional security measures often fall short in addressing the sophisticated and dynamic nature of modern cyber attacks, highlighting the need for more adaptive and intelligent security solutions. This study leverages the CICIDS2017 dataset, which is well-regarded for its detailed representation of real-world network traffic and attack scenarios, to enhance IDS effectiveness through advanced machine learning techniques.

Our research explores a blend of machine learning strategies, including supervised methods like Decision Trees and Neural Networks, alongside unsupervised approaches such as K-Means Clustering. Each technique has been carefully selected and rigorously evaluated to ascertain its efficacy in detecting and classifying network anomalies. The primary aim is to identify robust methods that not only address current security challenges but also adapt to new and emerging threats, ensuring that IDS remain effective as attack methodologies evolve.

However, our investigative journey revealed some challenges, notably suspiciously high accuracy rates that raised concerns about potential data leakage, and the initial poor performance of unsupervised models. These issues prompted substantial methodological revisions, underscoring the complexities involved in developing scalable and adaptive IDS solutions.

II. Supervised Learning

In the field of cybersecurity, specifically in the development and refinement of Intrusion Detection Systems (IDS), the accuracy and reliability of threat detection are paramount. Supervised learning offers significant advantages in this context due to its ability to learn from labeled datasets that include both normal and malicious activities. This learning approach enables the model to understand complex patterns and behaviors associated with different types of network attacks, providing a foundation for accurately identifying potential threats.

Supervised learning algorithms are particularly suited for IDS because they can be trained on a diverse set of features derived from network traffic, learning to classify these features into predefined categories of behavior (e.g., benign or various types of malicious activities). The use of a labeled dataset, such as CICIDS2017, ensures that each training instance is annotated with a ground truth, allowing the model to learn the distinct characteristics of each class effectively.

We utilized Logistic Regression with a softmax function for multiclass classification. This model provides a solid baseline for multiclass problems due to its efficiency and effectiveness. It is particularly valuable in scenarios where probabilistic outcomes are beneficial for decision-making, offering probabilities for each class and thereby aiding in analyzing the model's confidence in its predictions.

Known for its robustness, the Random Forest algorithm builds numerous decision trees during training and outputs the class that is the mode of the classes predicted by the individual trees. This method is favored for its high accuracy, capability to manage large data sets with extensive dimensionality, and its facility to evaluate the importance of features, which is instrumental in determining which features crucially influence the model's decisions.

To address the nonlinear complexities and interactions in network traffic data, we also implemented a Neural Network model. This model comprised multiple dense layers, which effectively captured the intricate relationships in the data. Neural networks are particularly adept at pattern recognition and classification tasks in high-dimensional spaces, making them well-suited for the nuanced detection requirements of IDS. The deep learning architecture allows for learning high-level features in data, which is crucial for identifying subtle anomaly patterns that might elude simpler models.

All models were rigorously evaluated using standard metrics such as accuracy, precision, recall, and the F1-score. Despite the high accuracy observed, challenges such as class imbalance necessitated a deeper interpretation of the metrics, particularly focusing on the precision and recall for minority classes. This evaluation was not only aimed at validating the models' effectiveness but also at identifying opportunities for optimization, such as

modifying class weights or enhancing feature engineering techniques.

III. Unsupervised Learning

In addressing the multifaceted nature of cybersecurity threats, unsupervised learning provides a critical advantage. It allows Intrusion Detection Systems (IDS) to identify novel or unforeseen patterns of attacks, independent of labeled data. This is crucial for the detection of zero-day threats and subtle network intrusions that may not be represented in labeled datasets.

Unsupervised learning algorithms like Isolation Forest are invaluable for detecting anomalies in an environment where labels may not exist, particularly for novel or zero-day attacks. These algorithms operate by isolating outliers without prior knowledge of their nature, which is critical for proactive threat detection in Intrusion Detection Systems (IDS). Isolation Forest relies on a set of hyperparameters that influence its ability to discern between normal behavior and anomalies. Identifying the optimal combination of these parameters is pivotal for the algorithm's success.

In our research, we systematically experimented with a range of values for the following key parameters.

- N_estimators
  - The number of trees in the forest, which impacts the ensemble's ability to generalize and reduce variance.
- Max_samples
  - The subset size to train each base estimator, which affects the diversity and depth of the forest.
- Bootstrap
  - Whether or not to allow bootstrap sampling when building trees, which can introduce randomness and affect the variance of the model.
- Contamination
  - The proportion of outliers expected in the data, crucial for threshold setting in anomaly scoring.

The search for the best parameters was conducted through an exhaustive manual grid search, testing combinations of n_estimators, max_samples, bootstrap, and contamination to find the setup that maximizes the model's performance. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) metric was employed to evaluate the quality of the Isolation Forest's anomaly detection. This metric is particularly suitable for imbalanced datasets common in intrusion detection, as it provides a balanced measure of a model's ability to distinguish between the classes.

For each parameter configuration, the model was trained on the encoded dataset, and its performance was assessed based on the AUC-ROC score obtained from the model's decision function. This score reflects the model's ability to rank predictions rather than merely classify, providing insight into its certainty and precision.

To complement our unsupervised algorithms, we leveraged PCA for dimensionality reduction. PCA assists in distilling data to its most informative components, which can be crucial when working with extensive feature sets typical of network traffic data. This step not only streamlines the analysis but can also enhance the detection capabilities of algorithms like Isolation Forest and K-Means by reducing noise and focusing on the most significant features.

Unsupervised learning presents its own set of challenges, including sensitivity to the parameter selection and the interpretation of results without ground truth labels. In our study, we carefully adjusted the parameters of the Isolation Forest and evaluated different cluster numbers for K-Means to ensure optimal performance. Despite initial performance concerns, these algorithms ultimately provided valuable insights and a method for detecting anomalies that may have gone unnoticed using supervised methods alone.

IV. Findings and Insights

Our application of the Random Forest classifier unveiled a nuanced perspective on feature significance within the IDS context. The analysis revealed that features such as BwdPacketLengthStd, PacketLengthVariance, and PacketLengthStd hold

substantial importance. These features likely capture variability and irregularity in packet sizes, which are indicative of malicious traffic patterns like DDoS attacks, where packet sizes may be manipulated to disrupt service.

The AvgBwdSegmentSize and AveragePacketSize also emerged as significant, suggesting that average sizes within a session can be distinguishing characteristics of attack traffic, as many attack strategies either use large packets to flood the network or small packets to probe and exploit vulnerabilities quietly. [2]

Additionally, the prominence of BwdPacketLengthMean and BwdPacketLengthMax aligns with the expected behaviors of certain attacks that either use consistently sized packets or include bursts of large packets, thus revealing themselves as outliers compared to benign traffic.[1]

Interestingly, time-based features such as IdleMin, IdleMax, and IdleMean were also identified as important. These attributes likely relate to the timing and duration of traffic flow, which can be significantly different in attack scenarios where malicious entities might have irregular communication patterns.[1]

The classification report for the Logistic Regression model reveals a high precision and recall for the majority 'BENIGN' class, reinforcing the model's capability to accurately identify non-malicious traffic. However, the model struggled with minority attack classes such as 'Bot', 'Heartbleed', 'Infiltration', 'SSH-Patator', 'Web Attack – Brute Force', 'Web Attack – Sql Injection', and 'Web Attack – XSS', where both precision and recall fell to zero. This indicates a deficiency in the model's ability to detect less frequent but potentially more dangerous threats. [2]

The contrast in the model's performance across classes underscores the challenge of class imbalance in IDS datasets. While overall accuracy is high, the low recall for several attack categories is concerning as it suggests that the model fails to identify a substantial proportion of actual attacks, a critical shortfall in an IDS. The 'DDoS', 'DoS GoldenEye', 'DoS Hulk', and 'PortScan' attacks had better recall, implying the model's relative effectiveness in identifying these more frequent or more patterned types of attacks.

The insights gained from feature importance and the classification report paint a detailed picture of our IDS's capabilities and limitations. The importance of packet size and time-related features suggests a strong dependency on traffic patterns for anomaly detection. The high accuracy, yet poor recall for several attack types, emphasizes the necessity for continued refinement of our model, particularly in improving its sensitivity to the less prevalent but harmful attack types. Enhancing our feature set, perhaps by engineering new features or by integrating external threat intelligence, could prove beneficial. Moreover, adopting a multi-model or ensemble approach may help to mitigate the weaknesses observed in using a single-model approach.

V. Challenges and Limitations

In the course of advancing Intrusion Detection Systems (IDS) through unsupervised learning techniques, our project navigated a complex landscape fraught with inherent challenges. A significant issue was the discordance between the high accuracy achieved and the considerable rate of false positives and false negatives, particularly in the application of the Isolation Forest algorithm.

Despite the Isolation Forest algorithm yielding an overall accuracy of 92%, a deeper dive into the confusion matrix revealed a substantial count of false positives and false negatives. This disparity is indicative of the model’s struggle to maintain precision and recall balance, an aspect of paramount importance in the IDS domain where the cost of misclassification is high. The high number of false positives could result in alert fatigue, where security teams become desensitized to warnings, potentially overlooking genuine threats. Conversely, the false negatives represent actual attacks that slipped through undetected, which poses a direct risk to the security of the network.

The fundamental challenge lies in enhancing the model's sensitivity to accurately detect attacks (true positives) without an untenable increase in false alarms.

The model's specificity, or its ability to correctly identify benign instances (true negatives), must be calibrated against this sensitivity. Our findings illuminate the difficulty of striking this balance in an unsupervised learning framework, which inherently lacks the guidance of labeled data.

VI. Conclusion

Our investigation into enhancing Intrusion Detection Systems (IDS) through both supervised and unsupervised learning approaches has yielded significant findings and practical insights. The performance of supervised models, in particular, stands out as a cornerstone for robust cyber defense mechanisms.

A. Strengths of Supervised Learning Models

The supervised learning models employed in this study demonstrated remarkable efficacy in detecting known types of network attacks. Models like Random Forest and Logistic Regression performed exceptionally well in classifying diverse attack scenarios, thanks to their ability to learn from labeled datasets which detailed examples of both benign and malicious network activities. These models were adept at recognizing and responding to the patterns and signatures of known attacks, which are crucial since the majority of network security threats exploit well-documented vulnerabilities.

B. Limitations of Supervised Learning

Despite their strengths, supervised models are not without limitations. Their dependency on labeled data for training is a double-edged sword. While it enables them to learn detailed attack patterns effectively, it also restricts their ability to detect novel or zero-day attacks, for which no prior labeled data is available. This reliance on comprehensive and up-to-date labeled datasets can be a significant hurdle in dynamic threat environments where new attack vectors emerge continuously.

C. The Role of Unsupervised Learning

Unsupervised learning, through methods like Isolation Forest, plays a vital complementary role by attempting to identify unusual patterns that do not conform to expected behavior. This capability is indispensable for spotting anomalies that could signify new, previously unrecognized threats. However, the challenges of high false positive and negative rates, as discussed in the previous sections, highlight the difficulties in solely relying on unsupervised methods for threat detection.

D. Integrative Approach for Enhanced Detection

In light of these observations, our project advocates for an integrative approach that leverages the strengths of both supervised and unsupervised learning techniques. By combining the precision of supervised models in detecting known threats with the exploratory power of unsupervised models to flag novel anomalies, we can create a more comprehensive defense strategy. This synergy is particularly crucial as the majority of attacks exploit known vulnerabilities, making the robust performance of supervised models invaluable.

In conclusion, this project has not only demonstrated the potential of machine learning in revolutionizing IDS but also highlighted the dynamic interplay between different types of learning strategies. The combination of supervised and unsupervised learning presents a promising pathway towards developing resilient, adaptive, and highly effective intrusion detection systems.

VII. References

[1] C. Chio and D. Freeman, Machine Learning & Security: Protecting Systems with Data and Algorithms. Sebastopol, CA: O'Reilly Media, 2018.

[2]A. Parisi, Hands-On Artificial Intelligence for Cybersecurity. Birmingham, UK: Packt Publishing, 2019.