Intrusion Detection System Using ML Models

Dhrumil Shah, Shresta Kommera, Shashank Shashishekar Reddy, Venkata Karthik Patralapati
Department of Applied Data Science
San Jose State University
San Jose, CA

I. Motivation

The sophistication and frequency of network attacks are continually increasing. Traditional security measures are often unable to keep pace with these evolving threats, necessitating more advanced solutions like machine learning-based intrusion detection systems (IDS). In addressing the critical need for advanced security measures within cybersecurity, the utilization of the CICIDS2017 dataset for developing machine learning classifiers for intrusion detection is both timely and essential. The increasing complexity and frequency of cyber threats necessitate the adoption of sophisticated technologies that can dynamically adapt to new attack patterns (Sharafaldin, Lashkari, & Ghorbani, 2018).

The CICIDS2017 dataset, notable for its up-to-date and diverse attack simulations alongside realistic traffic patterns, provides an invaluable resource for training algorithms capable of identifying subtle anomalies indicative of cyber threats. By leveraging modeling using this dataset, the proposed research aims to enhance the accuracy, efficiency, and scalability of intrusion detection systems, contributing significantly to the resilience of network infrastructures against malicious activities. This endeavor not only aligns with the current needs of cybersecurity defenses but also contributes to the broader academic and practical discussions on effective strategies for threat detection and management in the digital age.

II. Background

Intrusion Detection Systems (IDS) are vital for protecting network integrity, confidentiality, and availability against rising cyber threats. Since their inception in the 1980s, IDS have evolved from simply monitoring network traffic for known threats to utilizing advanced strategies like anomaly detection and machine learning, enabling them to detect sophisticated, new attacks. These systems include Network-based IDS (NIDS), which analyze all network traffic, and Host-based IDS (HIDS), which focus on individual host activities. Modern IDS, integrated with Intrusion Prevention Systems (IPS), form a proactive defense mechanism that not only detects but also responds to threats. This integration enhances their capability to analyze vast data using artificial intelligence, improving anomaly detection. As cyber threats grow more complex, the continuous advancement of IDS technologies is crucial for effectively safeguarding digital environments, reflecting the ongoing need for robust security measures in our interconnected digital world.

III. Literature Review

In our research, we have meticulously incorporated a range of machine learning models to enhance the performance of intrusion detection systems (IDS). Building upon the groundwork laid by Rashid et al. (2020), who highlighted the necessity of data normalization and feature selection through techniques like Principal Component Analysis (PCA) and Genetic Algorithms (GA), our model integrates these preprocessing steps to optimize the detection accuracy.

Our approach advances these concepts by selectively employing 4 models from Decision Trees, Random Forest, Logistic Regression, Gaussian Naive Bayes, Stochastic Gradient Descent, K-Nearest Neighbors, XGBoost, and Multi-Layer Perceptrons. These models were chosen based on their demonstrated success in previous studies, such as their application in handling the complexities of the CICIDS-2017 dataset (Panwar et al., 2019). Specifically, models like Random Forest and XGBoost have shown superior performance across crucial metrics like accuracy, precision, recall, and F1-score.

In particular, the Multilayer Perceptron (MLP) classifier has shown remarkable precision in our tests, achieving an accuracy of 99.73% in complex detection scenarios, echoing the findings of Yin et al. (2019). Similarly, in the study by Abraham and Bindu (2021), the utilization of XGBoost and Random Forest delivered high precision, with accuracies of 81% and 78% respectively, in detecting network intrusions.By synthesizing these diverse methodologies and integrating robust sampling techniques such as Random Sampling for anomaly detection, our model represents a comprehensive and refined approach to IDS. This holistic application of advanced machine learning strategies and feature selection techniques positions our model to deliver highly effective results, making it a significant advancement over previous methods in the realm of network security.

IV. Methodology

A. Data Collection

The CICIDS2017 dataset was meticulously gathered at the Canadian Institute for Cybersecurity, simulating real-world network traffic to enhance IDS analysis. It captures a broad spectrum of network behaviors from benign activities to complex attacks like Brute Force, DoS, and Heartbleed, recorded over a week in July 2017 using the B-Profile system, which emulates the behavior of 25 users across standard internet protocols. This diverse traffic was labeled using CICFlowMeter

and is available in formats including full packet payloads in pcap and CSV files for machine learning purposes. The dataset ensures comprehensive data for developing and benchmarking intrusion detection systems, enhancing their accuracy and adaptability.

B. Data Preprocessing

The preprocessing of the Intrusion Detection System dataset encompasses a series of systematic steps designed to refine the data, ensuring it is suitable for effective analysis and model building. This section describes each step, elucidating why they are essential and how they contribute to enhancing the classification process.

The initial stage involves merging multiple CSV files into a single DataFrame. This integration is crucial as it simplifies data manipulation and analysis by providing a holistic view of the dataset. Consolidating data from various sources ensures consistency in data handling and eliminates the complexity of working with fragmented data sources, thereby enhancing the efficiency of subsequent preprocessing steps.

Missing data can introduce bias and inaccuracies in predictive modeling. By identifying and removing rows with missing values, we ensure the dataset's integrity,but in our project missing values are so important as we are doing anomaly detection. However, there were no missing values in the dataset.

Duplicates can skew the data distribution and influence the learning algorithms undesirably by over-representing certain observations. Eliminating duplicates from the dataset, which comprises 307,376 rows, is essential to prevent data skew and ensure unbiased model training. Removing these redundancies allows for accurate learning and effective generalization, enhancing model performance on new data. Columns with zero variance, such as 'Bwd PSH Flags', 'Bwd URG Flags', and various bulk rate metrics, were removed from the dataset because they provide no discriminatory value for predictive modeling. Eliminating these columns simplifies the model by reducing complexity and helps prevent overfitting during training.No infinite values were detected in the dataset. This check is crucial to prevent disruptions in processing, as infinite values can interfere with scaling and model training operations. Standardizing feature names by removing extraneous characters and unifying text cases ensures consistency, avoiding column misidentification and streamlining data processing workflows. Removing identical columns reduces data redundancy, lowers computational load during model training, and helps prevent overfitting by simplifying the dataset's dimensionality. Properly converting feature data types ensures that mathematical and statistical operations performed during model training are applicable and executed correctly, which is vital for the accuracy of the model.

The target label distribution is observed in the Fig. 1 focusing on the dataset's counts of various attack labels. DoS Hulk appears to have the largest count, followed by DDoS attacks and PortScan. DoS GoldenEye, FTP-Patator, SSH-Patator, and several online attacks like Brute Force, XSS, and

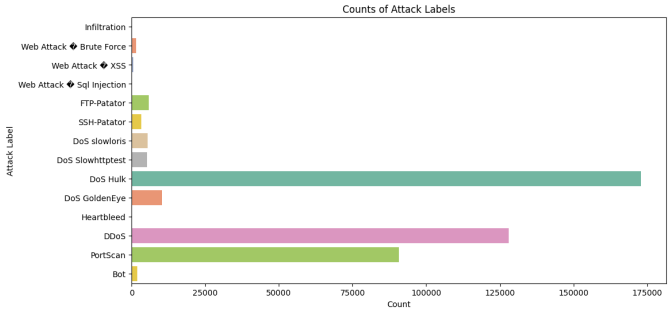

Fig. 1. Barplot of Various Types of Attacks.

SQL Injection are other prominent attack types. Attacks such as DoS Slowloris, DoS Slowhttptest, Heartbleed, Infiltration, and Bot are less common. By highlighting the frequency of specific attack types, this distribution sheds insights on the nature of the dataset and facilitates the creation of machine learning models for intrusion detection.

To create numerical consistency and allow for mathematical operations, the DataFrame df's columns "flow_packets/s" and "flow_bytes/s" were changed to numeric data types using pd.to_numeric. This was done to improve the accuracy, dependability, and clarity of our findings by accurately analyzing and reporting on packet and byte flow rates.

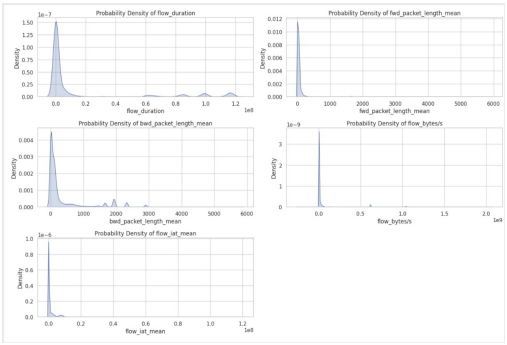

Fig. 2. Probability Density Plot of Network Traffic Features.

The Fig. 2 shows five probability density graphs for various network flow features. These features include flow duration, mean forward packet length, mean backward packet length, flow bytes per second, and mean flow inter-arrival time. Each plot shows the distribution of the corresponding feature values within the dataset, with the majority of values concentrated at the lower end of the range and exhibiting a right-skew distribution. This indicates that the majority of network flows have shorter durations, smaller packet lengths, lower bytes per second, and shorter inter-arrival intervals, as is typical of network traffic data.

C. Data Transformation

Applying MinMaxScaler normalizes feature values, enhancing the performance of algorithms sensitive to the magnitude of data, such as gradient descent-based methods, by ensuring

that all features contribute equally to the model's learning process. Retaining features highly correlated with the target variable while discarding others helps focus the model on relevant attributes, improving the model's interpretability and reducing the risk of overfitting. From the figure 2 (Write about highest correlated features). Employed modeling for two sampling techniques (Random Undersampling and Smote Oversampling). Random undersampling addressed the class imbalance, which can lead models to bias towards the majority class. Balancing the classes ensures that the classifier does not overlook the minority class, thus improving the model's accuracy across all classes. Here all the attack records combined will have the number of records in benign. Also, SMOTE enhances training data balance by generating synthetic attack records, not just replicating existing ones. This method corrects class imbalances, facilitating more effective learning of rare but critical attack classifications in intrusion detection systems, thereby improving model accuracy and generalization.In the data preparation phase, we divided our balanced dataset into training, validation, and testing segments to facilitate thorough model evaluation and optimization. We allocated 70% of the data for training, while the remaining data was evenly split into validation and test sets, each comprising 15% of the total dataset.

V. Modeling

In our study on intrusion detection, we developed and evaluated binary and multi-class models to differentiate between benign activities and various network attacks. The binary models classified activities as either benign or malicious, while multi-class models discerned specific types of attacks, providing detailed insights. We employed algorithms such as Decision Trees, Random Forests, XGBoost, and Multi-Layer Perceptrons.

To ensure robust model performance, we standardized the data using Standard scaling, normalizing each feature to zero mean and unit variance. This normalization was crucial for consistent training and accurate evaluation. To address class imbalance—a common issue in intrusion detection—we implemented both random undersampling and SMOTE. Random undersampling reduced the majority class size, while SMOTE increased the minority class, creating balanced datasets that enhanced the effectiveness and reliability of our models.

The rigorous training and testing of these models involved splitting the data into training, validation, and testing sets. This structured approach, combined with our scaling and sampling strategies, enabled our models to handle diverse and complex intrusion scenarios effectively. Our comprehensive evaluation across precision, recall, and F1-score metrics confirmed the models' capability to detect and classify different network threats accurately, offering valuable insights for enhancing security measures in real-world applications.

VI. Experiments

In our research, we conducted an extensive evaluation of preprocessing and sampling techniques to determine their impact on intrusion detection models. This evaluation covered both binary and multi-class classifications, utilizing MinMax and Standard scaling methods combined with Random Undersampling and SMOTE (Synthetic Minority Over-sampling Technique). Our experiments systematically applied these methods across various models, including Decision Trees, Random Forests, XGBoost, and Multi-Layer Perceptrons. Additionally, we investigated the influence of Principal Component Analysis (PCA) on model performance, conducting each experiment with and without PCA integration.

Precision (Under Sampling) – MULTI CLASS

Model	Baseline			PCA		
	Train	Validation	Test	Train	Validation	Test
Decision Tree	96.79	76.56	72.06	90	78.25	79.99
Random Forest	94.56	79.29	85.28	96.37	73.65	91.27
XGBoost	92.33	72.74	75.28	94.22	80.40	88.51
MLP classifier	91.99	74.86	84.72	95.18	74.41	81.65

Precision (Oversampling) – MULTI CLASS

Model	Baseline			PCA		
	Train	Validation	Test	Train	Validation	Test
Decision Tree	92.95	76.62	76.12	100	77.35	72.04
Random Forest	89.78	69.20	83.25	99.91	79.12	85.84
XGBoost	96.78	77.87	78.20	92.09	78.27	85.49
MLP classifier	91.17	69.63	67.62	91.72	69.40	68.66

Precision (Under Sampling) – BINARY CLASS

Model	Baseline			PCA		
	Train	Validation	Test	Train	Validation	Test
Decision Tree	96.96	92.50	82.30	94.99	92.65	93.67
Random Forest	96.18	94.32	94.21	97.91	97.21	97.05
XGBoost	98.90	98.33	98.34	99.68	98.77	98.86
MLP classifier	93.39	93.24	90.14	96.90	96.90	96.92

Precision (Oversampling) – BINARY CLASS

Model	Baseline			PCA		
	Train	Validation	Test	Train	Validation	Test
Decision Tree	97.97	98.42	97.44	98.99	97.99	97.72
Random Forest	96.82	96.64	96.45	95.67	89.24	88.45
XGBoost	97.82	94.96	91.37	98.94	97.79	97.97
MLP classifier	89.90	87.25	83.49	91.20	90.76	90.82

Fig. 3. Multi Class and Binary Classifications

For binary classification, our study explored the effectiveness of different combinations of scaling and sampling strategies. Initial experiments applied MinMax scaling followed by Random Undersampling, assessing model performances. Subsequent tests involved MinMax scaling with SMOTE, evaluating how these models adapt to oversampling. We also implemented Standard scaling with both undersampling and oversampling, aiming to identify the most robust preprocessing strategy for accurate intrusion detection.

The Fig. 3 presents precision scores for various models (Decision Tree, Random Forest, XGBoost, and MLP classifier) under binary classification with under-sampling, comparing baseline and PCA results. The XGBoost model stands out with consistently high precision scores across all datasets.

Specifically, under the PCA approach, XGBoost achieves 99.68% precision on the training set, 98.77% on the validation set, and 98.86% on the test set. These results highlight the superior performance of XGBoost, particularly when PCA is applied, demonstrating its effectiveness in handling high-dimensional data and improving model precision. Compared to the baseline, the PCA values for XGBoost show a slight increase, indicating that dimensionality reduction through PCA enhances the model's performance and robustness in detecting binary class intrusions. The XGBoost model performs exceptionally well on the intrusion detection project using the CICIDS2017 dataset due to its ability to handle complex, high-dimensional data efficiently and its robustness in distinguishing subtle patterns indicative of intrusions. Additionally, its advanced regularization techniques prevent overfitting, enhancing precision and reliability in identifying various attack types.

The Fig. 3 illustrates precision scores for various models (Decision Tree, Random Forest, XGBoost, and MLP classifier) under binary classification with oversampling, comparing baseline and PCA results. The XGBoost model consistently demonstrates high precision across all datasets. Specifically, under the PCA approach, XGBoost achieves 98.94% precision on the training set, 97.79% on the validation set, and 97.97% on the test set. These results highlight the superior performance of XGBoost, particularly when PCA is applied, showcasing its ability to handle high-dimensional data and improve model precision. Compared to the baseline, the PCA values for XGBoost show a noticeable increase, indicating that dimensionality reduction through PCA enhances the model's performance and robustness in detecting binary class intrusions.

In the context of developing an intrusion detection system using the CICIDS2017 dataset, XGBoost excels due to its advanced regularization techniques and ability to efficiently process complex, high-dimensional data, making it highly effective in accurately identifying various attack types.

Overall, XGBoost performs well in intrusion detection project when the benign features are undersampled using random under samplers than compared to using SMOTE and oversampling the attacks.

In multi-class classification, we replicated the binary classification experiments to discern specific types of network attacks with increased granularity. This approach allowed us to understand the scalability and effectiveness of our methods in a more complex classification context. Our findings aim to provide insights into optimizing intrusion detection systems, highlighting the best practices for deploying these models in real-world scenarios where accurate and efficient threat detection is critical.

The Fig. 3 presents precision scores for various models (Decision Tree, Random Forest, XGBoost, and MLP classifier) under multi-class classification with both under-sampling and oversampling, comparing baseline and PCA results.

In the under-sampling scenario, the XGBoost model shows strong performance, particularly when PCA is applied. With PCA, XGBoost achieves 94.22

In the oversampling scenario, XGBoost continues to perform, train well, with PCA-enhanced scores of 92.09

Overall, XGBoost performs well in intrusion detection project when the benign features are undersampled using random under samplers than compared to using SMOTE and oversampling the attacks.

In the context of developing an intrusion detection system using the CICIDS2017 dataset, XGBoost excels due to its advanced regularization techniques and ability to efficiently process complex, high-dimensional data, making it highly effective in accurately identifying various attack types. The consistent improvement with PCA underscores its robustness and suitability for real-world intrusion detection applications.

Our results indicated that Standard scaling consistently outperformed MinMax scaling across both classification types. The models processed with Standard scaling and Random over-samplers showed the best performance metrics, demonstrating superior precision, recall, and F1-scores. This suggests that Standard scaling, when combined with Random over-samplers, provides a more effective normalization technique in handling the inherent complexities and class imbalances found in network intrusion datasets. These findings underscore the importance of choosing the right preprocessing and sampling strategies to enhance the predictive accuracy and reliability of intrusion detection systems.

VII. Discussion and Future Improvements

Existing methods in intrusion detection systems (IDS) and intrusion prevention systems (IPS) have proven highly effective, utilizing a range of sophisticated techniques to secure networks from threats. However, as cyber threats continue to evolve in complexity and frequency, there remains significant room for enhancement in these systems through the integration of newer technologies and innovative approaches.

One of the primary areas for future improvement lies in the integration of advanced machine learning models. Traditional machine learning algorithms, while effective, often struggle to capture the intricate and dynamic nature of modern network traffic. To address this, future research should explore the incorporation of deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs can be particularly effective in identifying spatial patterns in network data, while RNNs are well-suited for temporal pattern recognition, making them ideal for detecting anomalies in network traffic over time. The use of these architectures could significantly enhance the accuracy of intrusion detection systems by capturing complex patterns that traditional methods might miss.

Additionally, ensemble methods that combine multiple models could further improve the accuracy and robustness of IDS. By leveraging the strengths of different algorithms, ensemble methods can provide a more comprehensive analysis of network traffic, reducing the likelihood of false positives and false negatives. Techniques such as bagging, boosting, and

stacking could be employed to create robust ensemble models that enhance overall system performance.

Advanced feature selection methods also hold promise for optimizing IDS. Methods like mutual information and recursive feature elimination can help in identifying the most relevant features for intrusion detection, improving both interpretability and efficiency. By focusing on the most critical features, these methods can enhance the performance of machine learning models, making them more efficient and effective in detecting threats.

Adaptive sampling techniques like Adaptive Synthetic Sampling (ADASYN) are another area of potential improvement. These techniques can dynamically adjust the training data to better reflect current network behaviors and emerging threats. By generating synthetic samples for minority classes, ADASYN can address class imbalance issues, ensuring that the IDS remains effective even as the nature of network traffic evolves.

Moreover, integrating cost-sensitive learning and stability metrics into IDS could further enhance their reliability. Cost-sensitive learning allows systems to consider the impact of different types of misclassifications, ensuring that critical threats are prioritized over less severe ones. Stability metrics can help in evaluating the robustness of the IDS against new, unseen data, ensuring that the system maintains high performance over time.

Finally, ongoing research should focus on the development of real-time intrusion detection capabilities. The ability to detect and respond to threats in real-time is crucial for minimizing the impact of cyber attacks. Leveraging streaming data analytics and real-time processing frameworks, future IDS could offer immediate threat detection and response, significantly improving network security.

In summary, while existing IDS and IPS are effective, there is considerable potential for enhancement through the integration of advanced machine learning models, deep learning architectures, ensemble methods, advanced feature selection, adaptive sampling techniques, cost-sensitive learning, and real-time capabilities. These improvements could collectively strengthen intrusion detection systems, making them more robust and reliable against the ever-evolving landscape of cybersecurity threats.

ACKNOWLEDGMENT

Our profound appreciation goes out to Professor Shayan Shams for his constant leadership and helpful support during the course of the coursework, which greatly aided in the project's successful completion.

REFERENCES

[1] S. S. Panwar, Y. P. Raiwani and L. S. Panwar, "An Intrusion Detection Model for CICIDS-2017 Dataset Using Machine Learning Algorithms," 2022 International Conference on Advances in Computing, Communication and Materials (ICACCM), Dehradun, India, 2022, pp. 1-10, doi: 10.1109/ICACCM56405.2022.10009400.

[2] Y. Rbah et al., "Machine Learning and Deep Learning Methods for Intrusion Detection Systems in IoMT: A survey," 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Meknes, Morocco, 2022, pp. 1-9, doi: 10.1109/IRASET52964.2022.9738218.

[3] Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity, 2*(1). https://doi.org/10.1186/s42400-019-0038-7

[4] Rathee, A., Malik, P., & Kumar Parida, M. (2023). Network Intrusion Detection System using Deep Learning Techniques. In "2023 International Conference on Communication, Circuits, and Systems (IC3S)" (pp. 1-6). Bhubaneswar, India. https://doi.org/10.1109/IC3S57698.2023.10169122

[5] Rashid, A., Siddique, M. J., & Ahmed, S. M. (2020). Machine and Deep Learning Based Comparative Analysis Using Hybrid Approaches for Intrusion Detection System. In *2020 3rd International Conference on Advancements in Computational Sciences (ICACS)* (pp. 1-9). Lahore, Pakistan. https://doi.org/10.1109/ICACS47775.2020.9055946