Insert Your Title Here*

Insert Subtitle Here

Aisling McGillicuddy
Department Name
Institution/University Name
City State Country
email@email.com

ABSTRACT

Provide a short one paragraph abstract for your report..

KEYWORDS

Insert 3-5 keywords for your project.

1 Introduction

The project was undertaken to directly address the critical cybersecurity challenge of DDoS attacks, which are one of the major causes of service disruption and significant financial loss for organizations worldwide. Current methods used for detection are often affected by high latency or poor accuracy. The following is the central research question that guides the project: "How accurate and efficient are machine learning models at classifying real-time network traffic as Normal or DDoS using flow-based features, and which features are the most significant indicators of an attack?" The aim is to develop and assess high-accuracy, low-latency machine learning models appropriate for real-time deployment; the purpose is to advance existing cybersecurity defenses through data-driven anomaly detection techniques.

2 Data

The primary source for the data is the Real-Time DDoS Traffic Dataset for ML available on Kaggle: https://www.kaggle.com/datasets/kalireadhat/realtime-ddos-traffic-dataset.

2.1 Source of dataset

The dataset, titled Real-Time DDoS Traffic Dataset for ML, was accessed from Kaggle and is considered a reasonably credible source for a research project because it is explicitly designed for supervised machine learning and includes highly relevant flow-based features, though it is not a primary source like a major university lab's repository. The dataset's credibility is tempered by the creator's explicit note that it "may contain simulated attack patterns," a common and necessary practice in cybersecurity data generation to safely replicate threats. The dataset was publicly listed or significantly updated on Kaggle around a year ago, but the precise date the underlying network traffic was captured is not specified. The dataset was generated by the creator by compiling network traffic that either replicated real-time conditions or was simulated under carefully controlled network configurations to produce a mix of authentic DDoS and normal traffic instances, with flow-based metrics then extracted and labeled. This simulation process is standard for creating labeled flow-based security datasets, enabling researchers to build models with features like packet_count_per_second that are directly indicative of attack behavior.

2.2 Characters of the datasets

The Real-Time DDoS Traffic Dataset for ML is a structured and labeled collection of network flow instances, comprising both normal traffic and simulated/replicated DDoS attack instances, making it specifically suitable for supervised machine learning training and validation for real-time anomaly detection. The data contains flow-based metrics highly relevant to real-time analysis, with key features including the binary label traffic_type (Normal vs. DDoS), along with quantitative measurements such as packet_count, packet_count_per_second, byte_count, and byte_count_per_second, among other flow-based statistics. Before modeling, the initial data preparation will involve Exploratory Data Analysis (EDA) to check for null values and the distribution of the target variable. Subsequently, Feature Scaling/Normalization (such as standardization or Min-Max scaling) will be applied to the flow-based metrics to prevent features with wide value ranges from skewing the learning process, which will involve careful consideration of normalization techniques effective for real-world generalization. Finally, the binary target variable will be appropriately encoded such as 0 for Normal, 1 for DDoS, and while the Random Forest model provides inherent feature importance, preliminary Feature Selection/Engineering may also be explored to potentially reduce dimensionality and enhance real-time computational feasibility.

3 Methodology

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

WOODSTOCK'18, June, 2018, El Paso, Texas USA

The project's methodology is quantitative and data-driven, centering on supervised machine learning classification to address the research questions. The work will be implemented using the Python programming language, leveraging core libraries like Pandas and NumPy for data manipulation, Scikit-learn (sklearn) for model implementation and evaluation, and Matplotlib and Seaborn for visualization of EDA and results, all within an interactive environment like Jupyter Notebooks. The Random Forest Classifier will serve as the baseline model due to its excellence in classification, robustness to overfitting, and inherent ability to compute feature importance , which directly addresses the second research question. To ensure a comprehensive comparison, the Random Forest's performance will be benchmarked against at least two other algorithms: Logistic Regression valued for its simplicity and interpretability and a Gradient Boosting Machine like XGBoost or LightGBM known for often achieving higher accuracy. All models will undergo rigorous evaluation using a train/test split and cross-validation, with performance measured by key metrics including Accuracy, Precision critical for minimizing blocked normal traffic, Recall vital for detecting all attacks, F1-Score, and ROC AUC. Finally, a crucial step will be the Real-Time Feasibility Analysis, where the computational speed and prediction latency of the best-performing model will be measured to assess its viability for practical, real-time deployment.

3.1 Random Forest Classifier

The Random Forest (RF) Classifier is an ensemble learning method that constructs a multitude of decision trees during training, ultimately outputting the class that represents the mode of the predictions made by the individual trees. The model relies on the assumptions of approximate independence among the individual tree errors—achieved through bootstrapping the data and using a random subset of features for splitting—and the belief that the majority vote of many weak classifiers yields a strong, accurate result. Its primary advantages include its high accuracy, robustness to overfitting, and its inherent ability to provide a valuable feature importance measure, Mean Decrease in Impurity, which directly addresses a core research question. However, RF can be computationally expensive and less interpretable than simpler models. The model is chosen for this project specifically because of its high accuracy for the critical DDoS classification task and its built-in feature importance capability. It will be implemented using the sklearn.ensemble.RandomForestClassifier function in Python, with extra work involving Hyperparameter Tuning such as optimizing n_estimators and max_depth, Feature Scaling to normalize flow metrics, and applying Class Weighting to address potential data imbalance, thereby improving the robustness and generalizability of the results.

3.2 Heading Level 2

…

Example format: The updated template, user manuals, samples, and required fonts, all are available at the URL https://www.acm.org/publications/proceedings-template. It contains said information for all three versions of MS Word (Windows and 2 versions of Mac). There are also separate links to the user guide, which can be referred to by the user. This URL also contains some useful video links, which describe how to add the template, structure the paper, and generate the layout, in different clips. **Display Formula with Number**

$$\sqrt{b^2-4ac}\,\frac{-b\pm\sqrt{b^2-4ac}}{2a} \tag{1}$$

Continuation part of Paragraph Text The user must style this paragraph in **ParaContinue** style, which follows immediately after the **DisplayFormula** (numbered equation). The **DisplayFormula** style is applied only in case of a numbered equation. A numbered equation always has a number to its right. Insert paragraph text here. **Display Formula without Number**

$$\sqrt{b^2-4ac}\,\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

The **DisplayFormulaUnnum** style is applied only in case of an unnumbered equation. An unnumbered display equation never contains an equation number to its right, and this unique property distinguishes it from a numbered equation.

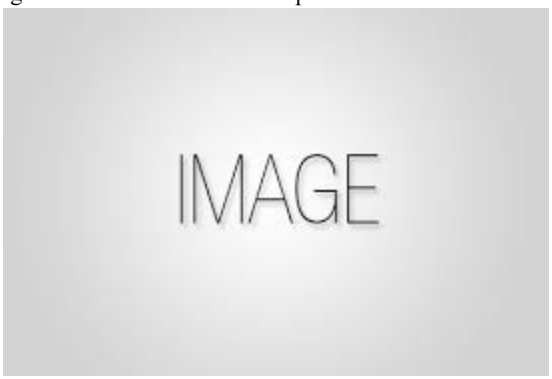

Figure 1: Figure Caption and Image above the caption [In draft mode, Image will not appear on the screen]

Theorem/Proof/Lemma. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement.

> ....Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract.

4 Results

In this part, you need to select a reasonable way to deliver the result of your topic. For example, equation or numerical results, or visualization of your result. You also need to provide a clear explanation of all results and how to understand the results. If there exist any unexpected results, please explain why or possible

cause of this special result. You can use subsection 4.1, 4.2, … to separate your results.

4.1 Heading Level 2

Example format: In the below paragraph, it is explained how alt-txt value is placed in **MS Word 2010**. To add alternative text to a picture in Word 2010, follow these steps:

1. In a Word 2010 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. Select the **Alt Txt** option from the left-side panel options.
4. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

Below are steps to place alt-txt value in **MS Word 2013/2016**. To add alternative text to a picture in Word 2013/2016, follow these steps:

1. In a Word 2013/2016 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. In the settings at the right side of the window, click on the "Layout & Properties" icon (3rd option).
4. Expand **Alt Txt** option.
5. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

1.1.1 Heading Level 3. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

1.1.1.1 Heading Level 4. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

5 Discussion

Every method/project has its shortage or weakness. Please discuss the unsatisfied results in your project. And discuss the feasible suggestions of future work to revise/improve your result.

6 Conclusion

In this part, you should summarize your project. What important results did you find for your topic and what's the effect of this result on the real-world?

ACKNOWLEDGMENTS

Insert paragraph text here. Insert paragraph text here.

REFERENCES

Use the following ACM Reference format for your citation

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

[1] Patricia S. Abril and Robert Plant, 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan, 2007), 36-44. DOI: https://doi.org/10.1145/1188913.1188915.

[2] Sten Andler. 1979. Predicate path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL '79).* ACM Press, New York, NY, 226-236. DOI:https://doi.org/10.1145/567752.567774

[3] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.

[4] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY..