TECHNIUM

Technium Vol. 15, pp.26-32 (2023)
ISSN: 2668-778X
www.techniumscience.com

A CatBoost-Based Approach for High-Accuracy Botnet Detection

Abdulkader Hajjouz[1], Elena Avksentieva[2]

[1] PhD student, Faculty of Software Engineering and Computer Technology, ITMO University, Russia.
abdulkaderhajjouz@gmail.com
[2]Associate professor, Faculty of Software Engineering and Computer Technology, ITMO University, Russia.
eavksenteva@itmo.ru

The rising prevalence of network botnet attacks poses a significant threat to online security. Compromised networks controlled by malicious entities can perpetrate harm, including distributed denial of service attacks and data theft. In this study, we introduce a method to detect these botnets using the CatBoostClassifier. By analyzing network traffic for suspicious patterns, our system efficiently identifies potential botnet activities. Utilizing the CTU-13 dataset, we achieved an impressive 99.8699% accuracy, underscoring the efficacy of our approach. This research offers valuable insights into botnet attack detection and presents a robust solution for enhancing network security.
Keyword: *Botnet Detection, CatBoost Classifier, Cybersecurity, Machine Learning, Network Security, CTU-13 Dataset, Feature Selection, Intrusion Detection, Model Optimization, Classification Metrics*

Introduction:

The growing prevalence of network botnet attacks is becoming an urgent issue within the realm of cybersecurity. Botnets, essentially networks of compromised devices orchestrated by unauthorized users, could execute a variety of detrimental actions. These range from Distributed Denial of Service (DoS) attacks to unauthorized data extraction. The devices often exploited in these schemes tend to have inadequate security safeguards, underscoring the urgent need to bolster network security infrastructure. This is especially critical as digital engagement—from social media to financial transactions—becomes an ever more intrinsic part of our daily lives.

The purpose of this research is to investigate the following questions:

- Are machine learning approaches, effective in detecting botnet activities?
- How does the performance of the CatBoostClassifier compare to other machine learning models in botnet detection?

To address these questions, the research employs machine learning techniques, with a focus on the CatBoostClassifier, to enhance the detection of botnet activities. Machine learning methods are particularly suited for this application due to their ability to scrutinize expansive datasets and recognize intricate patterns that are symptomatic of botnet invasions.

In our methodology, we incorporate critical stages such as feature extraction and model training to construct a robust detection system. Utilizing the widely acknowledged CTU-13 dataset, we isolate key features indicative of anomalous network traffic for training our algorithm. The findings highlight the exceptional capabilities of the CatBoostClassifier in improving botnet detection rates. Consequently, these results make a compelling argument for the integration of the CatBoostClassifier into current cybersecurity strategies as an enhanced protective measure.

Technium Vol. 15, pp.26-32 (2023)
ISSN: 2668-778X
www.techniumscience.com

Related Work:

In this section, we review the existing literature on botnet detection, specifically focusing on machine learning algorithms, feature selection, data types, and ensemble methods. Our aim is to contextualize our own approach within the broader scope of existing research.

- **Feature Selection Techniques in Botnet Detection**

Safitri et al. [1] proposed a method focusing on feature selection analysis techniques for botnet detection. This paper aimed to understand how the number and types of features affect detection accuracy in various classification algorithms like Decision Tree, k-NN, SVM, Random Forest, and Naïve Bayes. Notably, the study found that implementing more than four features led to decreased detection accuracy. This research indicates that optimal feature selection is crucial for improving botnet detection algorithms.

- **Botnet Types and Detection Algorithms**

Moorthy's [2] work discussed different types of botnets and the methodologies for their detection [Moorthy, R. Sri Skandha]. The paper built a detection system based on network packet monitoring and used machine learning algorithms for classification. Using the CTU-13 dataset, the decision tree algorithm showed the highest accuracy at 92%. However, the study also pointed out weaknesses in the model such as slow packet classification time and vulnerability to attacks.

Comparative Approaches to Botnet Detection Algorithms

Gong et al. provided a comparison of different machine learning algorithms like KNN, Decision Forest, and LightGBM [3]. Their work found that LightGBM outperformed other algorithms in terms of accuracy and false alarm rate. This supports the idea that newer or less-common algorithms could offer advantages over established methods for botnet detection.

- **DDoS Specific Detection Mechanisms**

Ahmed focused on the detection of DDoS attacks at the application level using a Multi-Layer Perceptron (MLP) classification model [4]. The model achieved a high accuracy rate of 98.99%. The study suggests that specialized algorithms might be required for different types of botnet attacks, including DDoS at the application level.

Ensemble Methods for High Accuracy

Arshad et al. proposed a novel ensemble method named KDR, combining machine learning and deep learning approaches [5]. The model achieved an impressive accuracy of 99.7% on the CTU-13 dataset. The study emphasizes that ensemble methods can often outperform individual machine learning algorithms.

Previous research in network attack detection has often reported suboptimal accuracy rates. However, the analysis presented here achieved a commendable accuracy of 99.8699%. In summary, previous research has explored various avenues for botnet detection ranging from feature selection to specific algorithms and ensemble methods. Each approach has its advantages and limitations, and there's a growing consensus on the need for improved feature selection and algorithmic methods. Our current study aims to contribute to this evolving field by proposing a CatBoost-Based Approach for botnet detection, benchmarked against these existing methods.

Dataset:

For our experimental research, we employed the well-known CTU-13 dataset, initially collected by the Czech Technical University. This dataset is designed to provide authentic samples of network traffic, including both benign and malicious traffic attributable to botnet attacks. The dataset was generated by executing thirteen distinct malware scenarios, and it covers various types of network

attacks such as IRC, Click Fraud, DoS, Port Scan, Spam traffic, and Fast Flux. The dataset includes 58 features that are relevant to both benign and botnet traffic types. In terms of labeling, the data is categorized as either normal traffic (0) or botnet attack traffic (1). Upon analysis, the dataset shows a relatively balanced distribution with 53,314 samples classified as benign traffic and 38,898 samples classified as botnet traffic, making it well-suited for a robust and unbiased machine learning model.

Table. 1. Number of samples and network traffic in CTU dataset

Dataset	TrafficType	Number of Samples Remaining
CTU	Benign	53314
	Botnet	38898

Features and Selection Techniques:

In the realm of cyber intrusion detection, especially concerning safeguarding against botnet attacks, the significance of feature selection is unparalleled. These features, which are pivotal in discerning data patterns, must be pertinent, devoid of redundancy, and crucial for bolstering the model's precision. Proper feature selection not only augments the efficacy of machine learning models but also curtails computational demands and heightens interpretability.

Assessing Correlations and Hierarchies: Leveraging the Spearman rank-order correlation, depicted as

$\rho = 1 - \frac{6\Sigma\, d_i^2}{n(n^2-1)}$, where d_i symbolizes the rank differential, we crafted a heatmap to visually represent feature correlations. The dendrogram, employing hierarchical clustering and the Ward linkage method $\Delta v = \frac{|c_k||c_l|}{|c_k|+|c_l|} d(c_k, c_l)^2$, portrays the feature clusters' hierarchical arrangement.

Aggregating Features: Post visualization, a clustering methodology was employed, rooted in correlation linkage. The intent was to amalgamate features into clusters, marked by a specific distance metric. Within these clusters, the aim was to retain only the most emblematic feature, either recognized as the centroid in L^2 space or a medoid when dealing with non-Euclidean distances. This strategy is aligned with dimensionality reduction, with a focus on conserving critical information. The features earmarked for further analysis exemplified the essence of each cluster.

Concluding this stringent selection process, the dataset was refined to comprise its 25 most salient features. These carefully chosen features not only honed the dataset but also paved the way for a streamlined, effective model training. The culminating feature set ensured a harmonious blend of computational efficacy and precision in modeling.

Table. 2. Selected features

0	Flow Duration	7	Bwd Pkt Len Min	14	Bwd PSH Flags	21	Down/Up Ratio
1	Tot Fwd Pkts	8	Bwd Pkt Len Mean	15	Bwd Header Len	22	Init Bwd Win Byts
2	Tot Bwd Pkts	9	Flow Byts/s	16	Fwd Pkts/s	23	Active Mean
3	TotLen Fwd Pkts	10	Flow Pkts/s	17	FIN Flag Cnt	24	Active Std
4	TotLen Bwd Pkts	11	Flow IAT Std	18	SYN Flag Cnt		
5	Fwd Pkt Len Min	12	Flow IAT Min	19	RST Flag Cnt		

Technium Vol. 15, pp.26-32 (2023)
ISSN: 2668-778X
www.techniumscience.com

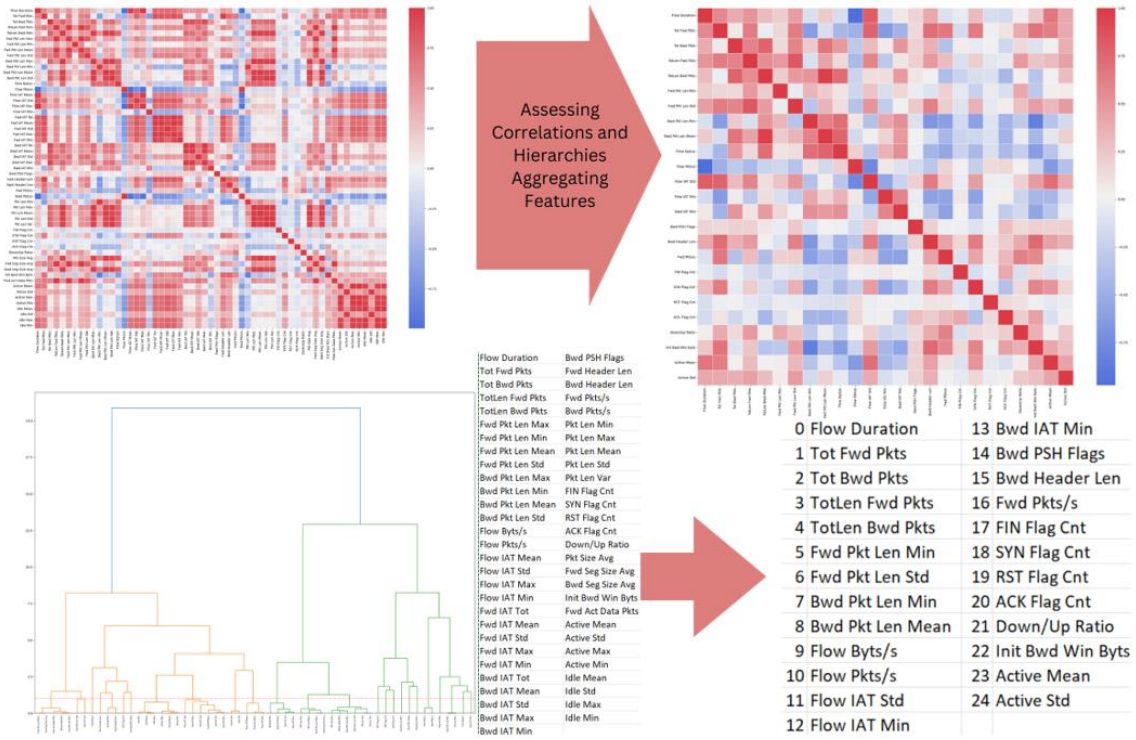

Fig. 1. Feature selection technology

Model Development and Optimization Using CatBoost for Botnet Detection:

To build an effective botnet detection model, we implemented a rigorous approach to data preprocessing and partitioning. The dataset was strategically divided into training and holdout subsets, utilizing stratified sampling based on class labels. This method preserved the class distribution across the splits, allocating 80% of the data for training and reserving the remaining 20% for holdout purposes. Subsequently, the holdout set was divided equally into evaluation and test subsets, each constituting 10% of the original dataset.

During preprocessing, categorical variables were transformed into a machine-learning-friendly format. We utilized the `getdum` function from the Pandas library to achieve one-hot encoding, applying this transformation to the training, evaluation, and test datasets.

We leveraged the CatBoost library's `Pool` function to establish data pools for training, evaluation, and testing. This structure is integral to CatBoost's efficient data management capabilities.

For model configuration, we employed the `CatBoostClassifier` with tailored hyperparameters. We set the loss function to 'Logloss' and focused on 'Recall' as our primary evaluation metric to ensure accurate identification of true positives. We also adjusted class weights to correct for any imbalances. Real-time training process monitoring was enabled by setting the verbosity parameter to true.

The training pool was used to fit the CatBoost model, with the evaluation pool serving for validation. The model was then fine-tuned to an optimal 671 iterations to prevent overfitting and improve generalization performance on unseen data.

Results and Discussion

The results obtained from the CatBoost Classifier in the context of botnet detection were nothing short of exemplary. A meticulous examination of the results underscores the prowess of the model in botnet intrusion detection tasks. The detailed metrics offer a comprehensive understanding of the

model's performance, especially when scrutinizing its capabilities to differentiate between normal traffic and botnet attack patterns.

The Area Under the Curve (AUC) values, both from the standalone metric and the ROC score, came tantalizingly close to perfection, registering a score of 0.998597. This indicates the model's near-impeccable ability to distinguish between the two classes, and the Average Precision Score of 0.997784 further accentuates this observation.

Examining the precision, recall, and F1-score metrics, it's evident that Class 0 (representing normal traffic) exhibited remarkable precision, recall, and F1-score values. Class 1 (representing botnet traffic) too showcased impressive numbers, with only a fractional divergence between precision and recall, indicating a very balanced classification. The overall accuracy was stellar, approximately 99.8699%, which is a testament to the model's efficacy.

Table. 3. Result analysis

	precision	**recall**	**f1-score**	**support**
Benign	0.998501	0.999250	0.998875	5332
Botnet	0.998971	0.997943	0.998457	3890
accuracy			0.998699	9222
macro avg	0.998736	0.998597	0.998666	9222
weighted avg	0.998699	0.998699	0.998699	9222

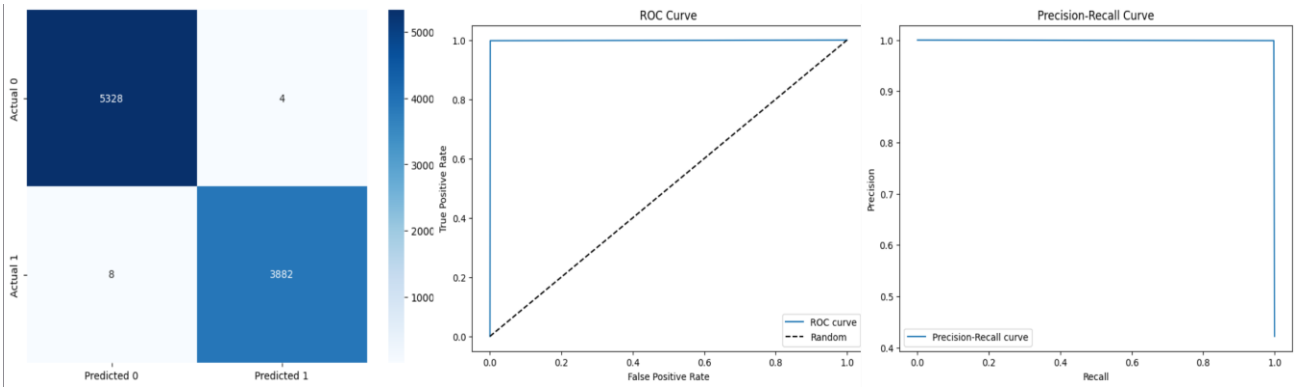

Fig. 2. Confusion Matrix, ROC Curve, Precision-Recall Curve

However, it's pertinent to point out that while the results are commendable, the inherent intricacies of real-world network traffic cannot be overlooked. The occasional nuances and sophisticated botnet patterns might sometimes resemble normal traffic, which can pose classification challenges. Notwithstanding, the exemplary performance of the CatBoost Classifier offers a solid foundation. With ongoing refinement and exposure to evolving botnet patterns, it holds immense promise to serve as a bulwark against increasingly sophisticated botnet attacks.

Table. 4. Result analysis

Ref	Year	Technique	accuracy
[1]	2022	DT	0.98
[2]	2023	DT	0.92
[3]	2023	KNN	0.96
[4]	2023	Multilayer perceptron	0.98
[5]	2023	KDR	0.997
Our study	2023	CatBoost	0.998699

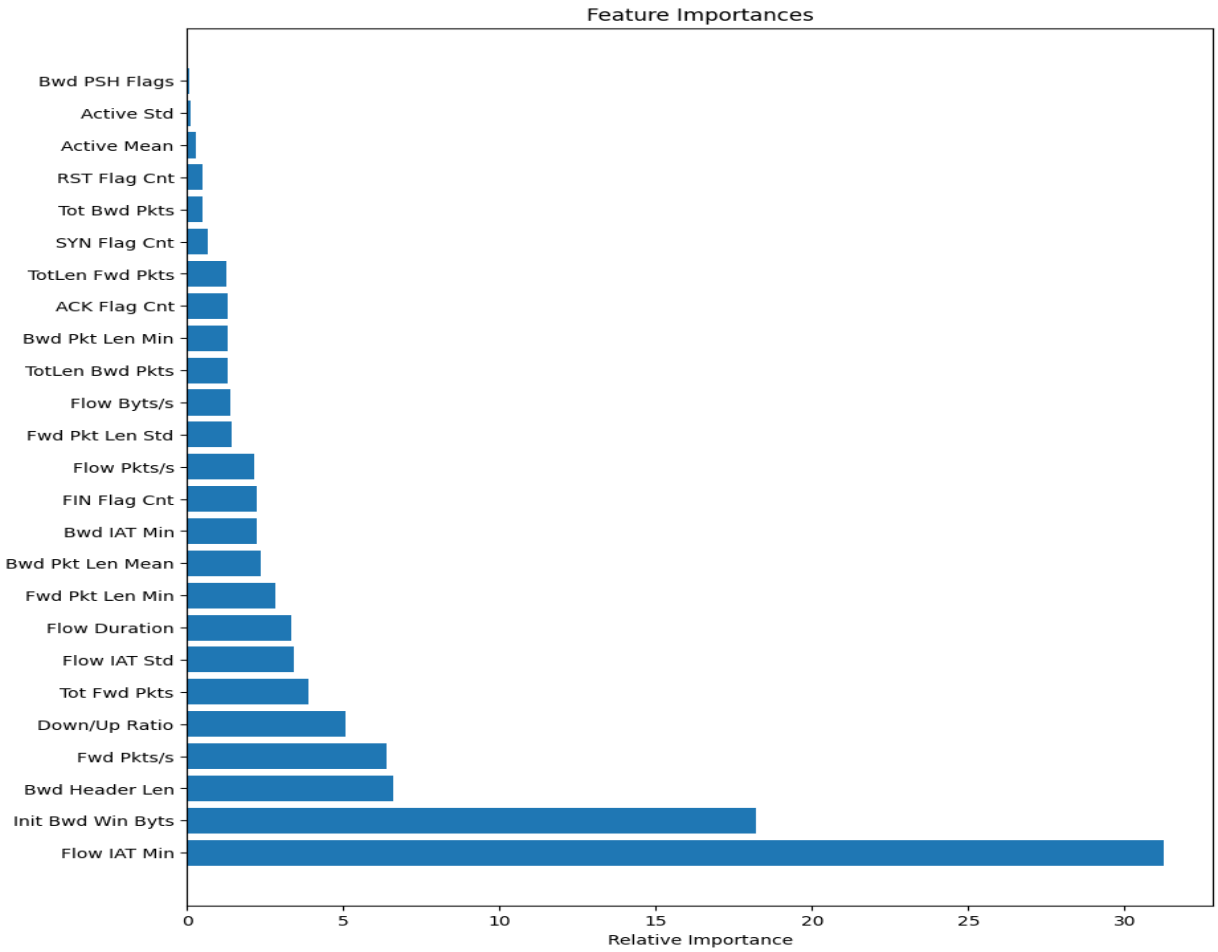

Fig. 3. Feature Importance

After training our CatBoost Classifier, we sought to determine the relative importance of each feature used in the model to provide insights into which network attributes are most indicative of botnet activities. To achieve this, we leveraged the feature importance attribute of the CatBoost model.

The feature importance score $FI(i)$for each feature i was computed using the CatBoost model's internal algorithm, which provides a numeric value representing the contribution of each feature towards the model's predictive accuracy.

The features were then ranked in descending order based on their importance scores, allowing us to understand the hierarchy of feature contribution in botnet detection. Formally, if n is the number of features, the sorted index $S = \{S_1, S_2, \dots, s_n\}$ is obtained such that $FI(S_1) \geq FI(s_2) \geq \cdots \geq FI(S_n)$.

We then plotted these scores on a horizontal bar chart with the y-axis representing the features and the x-axis representing the relative importance. This visualization aids in intuitively understanding which features are crucial for the model and therefore should be focused on in future iterations and practical applications.

Conclusions

In this study, we introduced a CatBoost-based approach, designed to provide a robust line of defense against botnet attacks. Our methodology was rigorously evaluated against the CTU-13 dataset, a widely acknowledged benchmark in the realm of network security. By leveraging the power of CatBoost, our approach demonstrated exemplary performance, achieving an astonishing accuracy of 99.8699%, setting it on a solid footing when compared against state-of-the-art methods in botnet detection.

As we look towards the future, the landscape of data analysis and machine learning continues to evolve, opening avenues for further enhancements in model accuracy and computational performance. One significant area of future focus will be the development of more refined data balancing and feature selection algorithms. These algorithms will aim to scrutinize large datasets to identify and rectify imbalances and inconsistencies, thereby improving the quality of the input data.

Future Work

For future research, computational complexity and resource planning for machine learning models warrant in-depth investigation. The newly proposed methods for computation capacity planning can guide cost-effective allocation of computing resources, particularly RAM [6]. Exploring the scope of these methods in botnet detection models like the CatBoost Classifier can lead to more efficient and cost-effective systems. This direction can significantly contribute to resource planning for deploying machine learning models at scale.

References:

1. Safitri, W. A., Ahmad, T., & Hostiadi, D. P. (2022, July). Analyzing Machine Learning-based Feature Selection for Botnet Detection. In 2022 1st International Conference on Information System & Information Technology (ICISIT) (pp. 386-391). IEEE.
2. Moorthy, R. S. S., & Nathiya, N. (2023). Botnet detection using artificial intelligence. Procedia Computer Science, 218, 1405-1413.
3. Gong, D., & Liu, Y. (2022, May). A Mechine Learning Approach for Botnet Detection Using LightGBM. In 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA) (pp. 829-833). IEEE.
4. Ahmed, S., Khan, Z. A., Mohsin, S. M., Latif, S., Aslam, S., Mujlid, H., ... & Najam, Z. (2023). Effective and Efficient DDoS Attack Detection Using Deep Learning Algorithm, Multi-Layer Perceptron. Future Internet, 15(2), 76.
5. Arshad, A., Jabeen, M., Ubaid, S., Raza, A., Abualigah, L., Aldiabat, K., & Jia, H. (2023). A novel ensemble method for enhancing Internet of Things device security against botnet attacks. Decision Analytics Journal, 100307.
6. Petrov, V., Gennadinik, A., & Avksentieva, E. (2022, May). Metrics for machine learning evaluation methods in cloud monitoring systems. In Proceedings of the 2022 8th International Conference on Computer Technology Applications (pp. 168-175).