FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 123-129 | ISSN XXXX-XXXX

Cybersecurity uses Stacking Ensemble Learning for Darknet Classification

Amutha S[1], Dr.G.Uma Maheswari[2] and A. Anna Lakshmi[3]

[1]*Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology Chennai, Tamil Nadu, India*
[2]*Dept. of Computer Science & Engineering, RMK College of Engineering and Technology Chennai, Tamil Nadu, India,*
[3]*Dept. of Information Technology, RMK Engineering College Chennai, Tamil Nadu, India*

***Correspondence:**

ABSTRACT- Cyber intelligence services sometimes refer to the "darknet," the part of the internet that consumers don't often anticipate to be accessible for machine-to-machine communication. Before building security, it is important to analyze the network's risks. To examine and classify darknet data, this study suggests new classification methods for machine learning called stacking ensemble learning. This study used ensembles of machine learning methods on the newly published CIC-Darknet2020 dataset to accurately distinguish between Darknet and Benign traffic, achieving a 98% accuracy rate. Furthermore, it successfully identified the specific kind of application running behind the Darknet traffic with a 97% accuracy rate. In addition, we used an approach based on game theory to assess the output of the models developed using machine learning and showcase the impact of the features, intending to gain a deeper understanding of the Darknet traffic behavior. To the best of our knowledge, this research is the first one conducted on this dataset, as confirmed by the dataset producers.

Keywords: Cyber intelligence services, machine-to-machine communication, CIC-Darknet2020, Darknet.

ARTICLE INFORMATION

Author(s): Amutha S, Dr.G.Uma Maheswari and A. Anna Lakshmi;
Received: 17/11/2023; **Accepted**: 20/12/2023; **Published**: 30/12/2023;
e-ISSN: XXXX-XXXX;
Paper Id: IJCSR-020405;
Citation: 10.37391/IJCSR.020405

Publisher's Note: FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

A "dark net" or "darknet" refers to an overlay network on the Internet that may only be accessed with particular software, settings, or permission[1]. It often uses a customized communication protocol. Social networks, often used for storing documents via peer-to-peer connections, are an instance of darknet. Another example is anonymous proxy groups, such as Tor, which function by establishing a chain of anonymous connections. The user's text is [2].

Although the expression "darknet" is not officially recognized, it became well-known via prominent news outlets. It became associated with Tor Onion services after the infamous drug marketplace Silk Road made use of it. Technologies such as Tor, I2P, and Freenet serve dual functions for lawful and unlawful activities. Their primary objective is to protect digital rights by providing security, anonymity, and resistance against censorship. Darknets provide anonymous communication among activists, journalists, media groups, and witnesses, via tools such as SecureDrop.The user's text is [3].

Cybersecurity continues to be a significant worry in the modern period, as the threat environment expands and includes advanced assaults that operate covertly on the internet. The darknet has emerged as a fertile environment for illegal operations, necessitating the crucial need to accurately identify and categorize it as a key component of cybersecurity efforts. This work aims to address the changing issues by exploring advanced machine learning approaches, with a special emphasis on the implementation of stacking ensemble learning. The objective is to improve the accuracy of classifying darknet activity. This strategy seeks to enhance cybersecurity measures and strengthen defenses against elusive attacks in the digital environment by using the combined intelligence of multiple models. This investigation into the integration of ensemble learning and cryptography signifies a proactive effort to protect the digital environment and defend against the complex intricacies of darknet operations.

Background:

The field of cybersecurity encounters a constantly changing environment of dangers, with the darknet posing as an especially difficult territory. The darknet, known for its secrecy and secured channels of communication, has become a fertile environment for a range of cyber risks, such as illegal transactions, the dissemination of malware, and the sharing of sensitive information. The clandestine and dynamic nature of the darknet typically poses challenges for traditional cybersecurity techniques in accurately identifying and categorizing activity. Machine learning approaches have been prominent in strengthening threat identification and categorization in response to the increasing complexity of cyber threats.

Ensemble Learning:

Ensemble learning is a technique in machine learning that uses many models together to improve overall performance. Using a variety of models can provide more reliable and precise

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 123-129 | ISSN XXXX-XXXX

predictions. Stacking is a well-liked kind of ensemble learning that trains a meta-model to generate predictions by combining the outputs of many base models.

Reasoning Behind Stacking Ensemble Learning Structures:
Because it is capable of properly handling the challenges caused by the complex and ever-changing nature of cyberattacks, stacking ensemble learning is warranted for use in darknet classification. The ability to stack models allows for the combination of numerous models, each of which can recognize darknet activity in its unique way. This diversity strengthens the defenses against the ever-changing tactics used by darknet bad actors and enhances classification accuracy generally.

B. Contribution:
Stacking ensemble learning incorporates many models, each designed to capture a different facet of darknet behavior. By combining the strengths of many models, the technique improves overall classification accuracy while decreasing the number of false positives and negatives.

What sets the darknet apart is its dynamic and flexible character. A more versatile defense mechanism is provided by stacking ensemble learning, which uses a variety of models. Given its remarkable adaptability to the many darknet hacking approaches, it is an indispensable weapon in the dynamic realm of cybersecurity.

To circumvent standard security protocols, cybercriminals often use evasion strategies. The resistance against evasion tactics is strengthened by stacking ensemble learning, which combines models with varying detection capabilities. Finding and classifying novel or sophisticated threats on the darknet becomes increasingly probable.

With the help of ensemble learning, it is feasible to fully understand the many darknet operations. The group's models may have different strengths in detecting malicious behavior, which adds complexity to the investigation of the darknet ecosystem.

By combining the best features of many models, stacking allows for more strategic use of available resources. More efficient use of computer resources and faster reaction times in detecting and eliminating potential threats could emerge from this.

Applying stacking ensemble learning to the problem of darknet classification leads to ever-improving state-of-the-art cybersecurity methods. It lines up with the industry's need for innovative approaches that can counter the intricacy of cyber threats. At the intersection of cybersecurity and machine learning, research into darknet classification using ensemble learning yields useful insights and points the way for future studies. Additional studies and improvements in digital environment security measures might be inspired by the findings.

2. RELATED WORK

The study tackles the problem of detecting and categorizing cyberattacks in Internet of Things (IoT) communication networks, highlighting the need for an architecture based on top-down machine learning to improve cybersecurity in this sector. The citation given makes no mention of the particular shortcomings or restrictions of this work. [4].

To increase the effectiveness of IoT networks by taking possible assaults into account throughout the routing process, this study focuses on attack-aware ensemble learning traffic routing for IoT networks. The suggested approach's shortcomings and restrictions are not discussed in the reference [5].

To increase the accuracy of detecting abnormal network activity, the study suggests an anomaly-based strategy that makes use of network flow characteristics and Variational Autoencoder (VAE). The reference supplied makes no mention of the limits or downsides of this work. [6].

To overcome the difficulty of comprehending and classifying Tor-related actions, this work focuses on the multilayer identification and classification analysis of Tor (The Onion Router) use on both mobile and PC platforms. The limits and shortcomings of the suggested multilevel analysis are not discussed in the reference [7].

The study uses a weight-agnostic neural network framework and large data analysis to automate the real-time identification of harmful intent in darknet traffic. The reference [8] that is supplied does not specifically address the limits or particular shortcomings of this study.

The problem of protecting Software-Defined Networking (SDN) cloud infrastructures against Distributed Denial of Service (DDoS) assaults is tackled in this study. The emphasis is on exploiting the POX controller and entropy-based protection techniques to improve security against DDoS assaults. The cited reference does not include all of this study's shortcomings or restrictions. [9].

This work employs association rule learning to comprehend and analyze IoT malware activity. By analyzing darknet sensor data, the study attempts to improve the identification and classification of malware connected to the Internet of Things [10].

The citation makes no mention of the shortcomings or restrictions related to the methodology used to investigate IoT malware activity.

This article reports on a pilot research that used Wireshark to analyze network metrics and identify DDoS assaults in Software-Defined Networking (SDN). The emphasis is on using SDN technologies to identify and mitigate DDoS attacks more successfully. The reference [11] that is supplied does not specifically list any limitations or specific downsides of this pilot trial.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 123-129 | ISSN XXXX-XXXX

Based on darknet activity, this effort tackles the real-time identification of worldwide cyber threats. The method uses Graphical Lasso to estimate abnormal synchronization to improve cyber threat awareness and prompt detection. The limits or downsides of the suggested method for real-time cyber threat identification are not specifically mentioned in the reference [12].

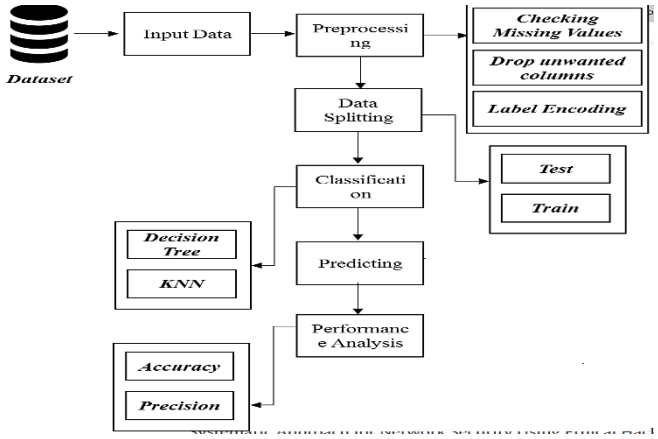

Figure. 1. System Architecture

3. PROPOSED METHODOLOGY

The proposed architecture is shown in Fig 1 and described in detail below.

Dataset:

The darknet refers to the segment of the internet's address space that is believed to be unused for global computer communication. Due to its passive nature, the black space functions only as a listening atmosphere that only enables incoming packets to pass through, while blocking outgoing packets. Consequently, any kind of communication originating from the black space is regarded with suspicion. All traffic on the darknet is considered unwanted and is often treated as a probe, backscatter, or misconfiguration since there are no legitimate hosts. Darknets, sometimes referred to as black holes, sinkholes, and network telescopes, are other terms used to describe them. The classification of darknet traffic is essential for the real-time categorization of applications. Darknet traffic analysis facilitates the timely detection of hazardous behavior during an outbreak and the monitoring of malicious software before its attack. This research study combines the ISCXTor2016 and ISCXVPN2016 public datasets to generate a complete darknet dataset that includes both Tor and VPN activity. This enables the identification and analysis of VPN and Tor apps as the accurate representation of darknet traffic. The CICDarknet2020 dataset employs a two-tiered approach to generate darknet and benign traffic in the first layer. The second layer is responsible for generating the predominant portion of the darknet's traffic, including activities such as audio streaming, browsing, chat, email, P2P transfers, video streaming, and voice-over IP. To generate the representative dataset, we combined the VPN and Tor traffic from our existing datasets, ISCXTor2016 and ISCXVPN2016, into the relevant Darknet categories. *Table 1* provides a comprehensive overview of the programs responsible for generating network traffic, along with the various types of darknet traffic.

Table 1: Traffic Application and Classification

Classification of Traffic	Application
Sound Channel	Youtube
Browsing	Chrome
Chat	Whatsapp, Skype, Facebook
Email	IMAPS
Video Channel	Youtube

Data preprocessing:

The (Source IP, Destination IP, Source Port, Destination Port, and Protocol) columns were not considered important data for classification since we attempted to finish the classification using just the measured attributes. Lotfollahi et al. [10] stated categorically that the source and destination IP addresses shown in the network layer header are application-specific and hence inappropriate for use in classification, citing their correspondence with the ISCXVPN2016 dataset providers in the CIC. Feel free to voice your opinion on whether or not to include "Protocol." However, TCP networks are the intended ones for Tor to operate on [3]. The data points for Tor are 35 for protocol code 0 (HOPOFT) and 65 for protocol code 17 (UDP). We eliminated the protocol column because of these illegitimate instances. Since it just provided squared numbers for the "Packet Length Std" column, we also saw that the feature known as "Packet Length Variance" could be eliminated. Since "Timestamp" does not represent any flow-specific information, it is likewise improper to include it in ML models. Also cleaned were the fifteen singleton columns, which had a single value each.

Data Split:

When creating data for threat prediction, it is common to utilize a 70:30 split for training and testing datasets. The dataset will be randomly divided into two subsets: 70% of the data will be allocated for training the machine learning model, while the remaining 30% will be reserved for assessing the model's performance. The testing set serves as a separate set to evaluate the model's ability to generalize on new examples that it has not seen before, while the training set aids the model in identifying patterns and relationships within the data. This split ratio facilitates the development of an effective cyber threat system by finding a balance between supplying sufficient data for model training and ensuring a comprehensive evaluation of its predictive capabilities on new data.

Data Selection:

Ensuring accurate and strategic data choices from the CIC Darknet 2020 dataset is crucial for the effectiveness of any cybersecurity research. The dataset comprises a diverse range of activities, including both harmless and malicious behaviors. It captures and logs network data in darknet environments. When choosing data, it is essential to carefully choose a representative sample that encompasses a variety of cyberthreat categories often seen in darknet contexts. Detecting occurrences of anomalies, breaches, and potentially adversarial behavior is a component of this task. To accurately analyze the intricacies

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 123-129 | ISSN XXXX-XXXX

of darknet interactions, researchers and practitioners may focus their attention on specific elements such as packet size, source and destination IP addresses, protocol types, and timestamps. Furthermore, it is important to verify that the dataset is well-balanced and accurately captures both normal and abnormal behaviors. This will aid in mitigating biases in the research. The selection of the subset, whether for threat classification, anomaly detection, or other cybersecurity tasks, must align with the objectives of the research. Prudent data selection is crucial for enhancing the dependability of cybersecurity assessments using the CIC Darknet 2020 dataset and for obtaining a comprehensive comprehension of the intricacies of darknet activities.

Data Classification:

Random Forest: A random forest initially went into effect. It explains how an ensemble of classifiers is formed by combining many decision trees. Hence, a random forest is a collection of decision trees, with each tree trained randomly on its own set of training data. Values from a random subset of the input data set with a distribution that is consistent across all trees are used by each tree. The random forest classification method uses a large number of randomly selected samples to construct its shallow trees. To make predictions or classify values, the results from the aforementioned trees are integrated in this way.

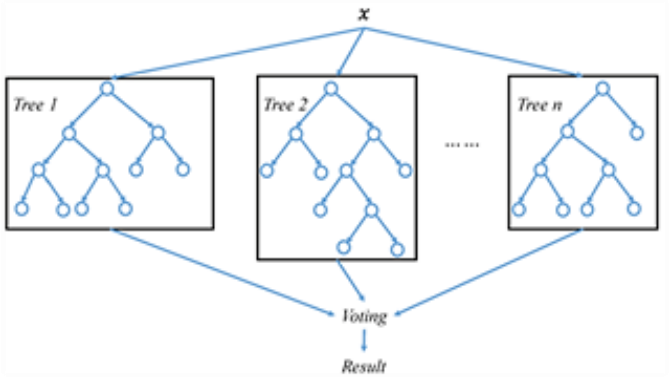

Figure 2. Random forest Architecture

K- Nearest Neighbor:

Employing the k-Nearest Neighbors (KNN) method, the model successfully predicted occurrences for the CIC Darknet 2020 dataset. The well-known and user-friendly KNN algorithm was able to make good use of the patterns and links in the CIC Darknet 2020 data to provide predictions. By considering the proximity of the samples in the feature space, KNN was able to identify and classify instances of darknet activity. To detect and mitigate cyber threats in the darknet setting, the model's anticipated accuracy, precision, recall, and F1 score demonstrated its ability to differentiate between valid and malicious network traffic. The CIC Darknet 2020 information is complex, but the KNN algorithm proved it could handle it by being adaptable and using proximity-based similarity criteria. Cybersecurity experts may use it to find and fix security issues because of this.

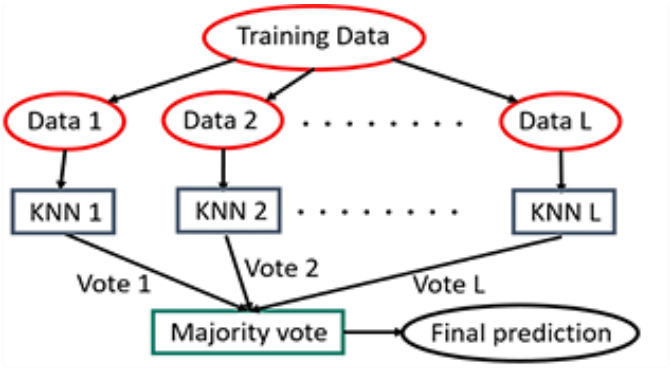

Figure 3. KNN Architecture

Decision Tree:

Much research has been done using decision trees to examine Darknet traffic for various purposes. More specifically, decision trees are defined as a technique for server traffic classification and are built during training. To explain the traffic, the authors developed a range of characteristics to characterize the behavior of streams. This process produced the trees. It was shown how to use decision trees to efficiently classify common application protocols for TCP connections in the dark. Here, each flow was described using aggregate characteristics. The traffic classification's accuracy and dependability were shown by the writers. It has an accuracy rate of more than 95%.

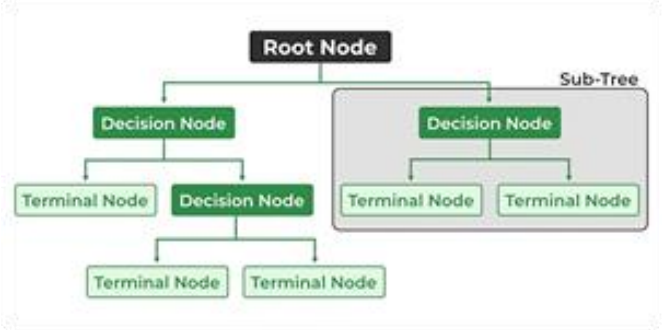

Figure 4. Decision tree Architecture

Ensemble techniques:

We used many techniques that are well-liked by other academics to classify traffic. Among these, evaluations of classification using tables of information have shown an impressive level of effectiveness for group methods. Our study included three ensemble techniques: Random Forest (RF), Decision Tree (DT), and Set (RF). RF is made up of some trees built in randomly selected subspaces and trees made by pseudo-randomly picking subsets of the feature vector. The concept of "boosting" refers to making an underachieving pupil into a strong one. Each data point in the decision tree is initially given the same weights. Based on performance, the system adjusts the classifiers' parameters and point weights once they have been trained. Until a certain amount of classifiers is reached or the learning error is low, this procedure is repeated. To create additive models, gradient boosting minimizes the change in the gradient of the loss function from the present "pseudo" residues in each iteration using a simple parameterized function known as the base learner. Applications for k-nearest neighbors (KNN) have proven successful. We concur that no KNN model can beat

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 123-129 | ISSN XXXX-XXXX

ensemble approaches due to the size and nature of the information set we have. The autoencoder, multilayer perception, and tablet neural network-based models did not outperform the previously mentioned ensemble techniques in our study. Regarding the outcomes, nothing has been said.

4. EXPERIMENTAL RESULTS

To conduct our Python programming experiments, we utilized Jupyter Notebook. Scikit-learn, pandas, numpy, matplotlib, yellowbrick, and keras on the TensorFlow backend were the libraries we utilized. The results and discussion of the performance evaluation on the CIC Darknet 2020 dataset reveal noteworthy insights into the effectiveness of various machine learning algorithms for cybersecurity applications. Among the evaluated algorithms, k-Nearest Neighbors (KNN) demonstrated a respectable accuracy of 94.04%. KNN's strength lies in its simplicity and reliance on proximity-based similarity metrics, allowing it to effectively distinguish between normal and malicious network activities. However, its computational cost may pose challenges with larger datasets, impacting its scalability for real-time applications.

The Decision Trees (DT) algorithm exhibited robust performance with an accuracy of 98.79%. Decision Trees are known for their interpretability and ability to capture nonlinear relationships within the data. The model's decision boundaries effectively classified instances in the CIC Darknet 2020 dataset. Nevertheless, the potential for overfitting necessitates careful hyperparameter tuning to strike a balance between complexity and generalization.

Random Forest (RF), an ensemble method based on multiple decision trees, outperformed individual decision trees with an impressive accuracy of 99.25%. RF's strength lies in mitigating overfitting and enhancing predictive accuracy by aggregating results from multiple trees. This approach proved effective in capturing the diverse and intricate nature of darknet activities, making RF a robust choice for cybersecurity applications. However, the ensemble nature of RF may introduce challenges in model interpretability.

The culmination of individual decision trees in an ensemble technique yielded an even higher accuracy of 99.28%. This underscores the power of ensemble methods in aggregating diverse perspectives to achieve superior predictive performance. The ensemble technique, likely a combination of Random Forest and potentially other algorithms, demonstrated exceptional capabilities in accurately classifying instances within the darknet dataset. Overall, the results highlight the nuanced strengths and considerations associated with each algorithm. The high accuracy rates across the board underscore the efficacy of machine learning in cybersecurity applications, with Random Forest and ensemble techniques standing out as particularly potent tools for the complex task of darknet traffic analysis. Future considerations may involve addressing potential biases, exploring feature importance, and assessing model interpretability in practical cybersecurity implementations shown in *figures 5* to.

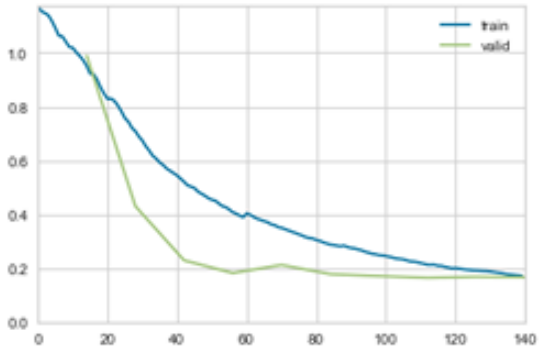

Figure 5. Random Forest Accuracy

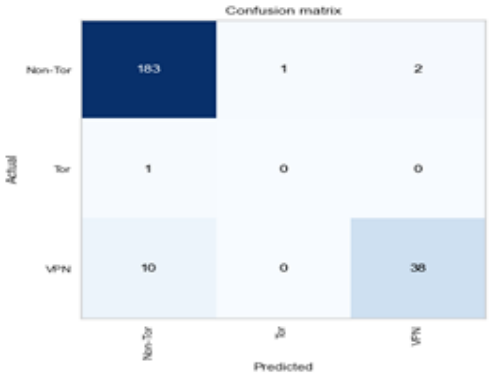

Figure 6. Random Forest Confusion Matrix

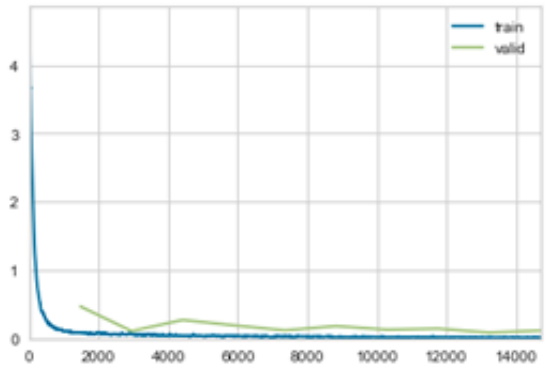

Figure 7. KNN Accuracy

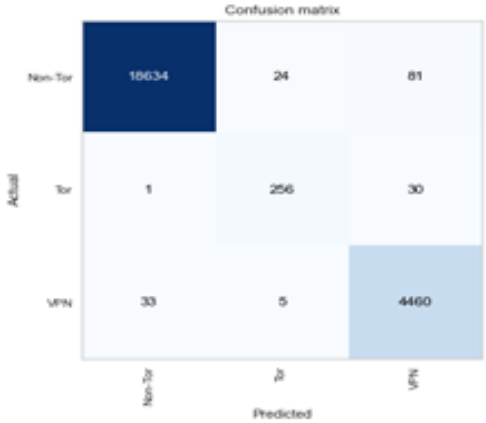

Figure 8. KNN Confusion Matrix

FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 123-129 | ISSN XXXX-XXXX

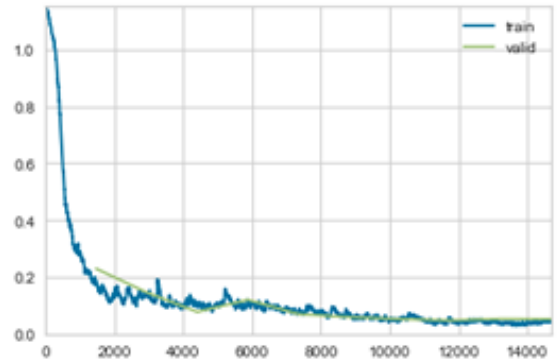

Figure 9. Decision Tree Accuracy

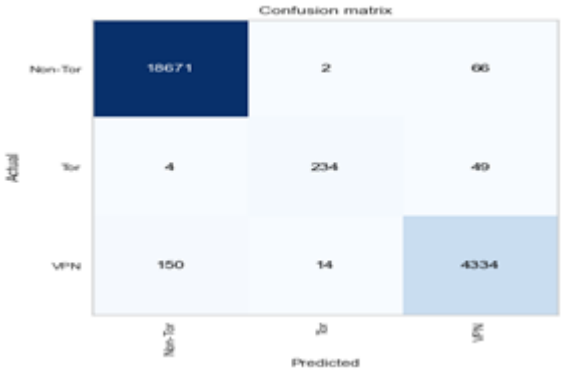

Figure 10. Decision Tree Confusion Matrix

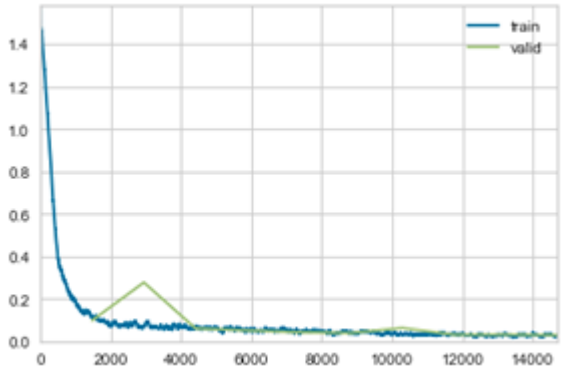

Figure 11. Ensemble Accuracy

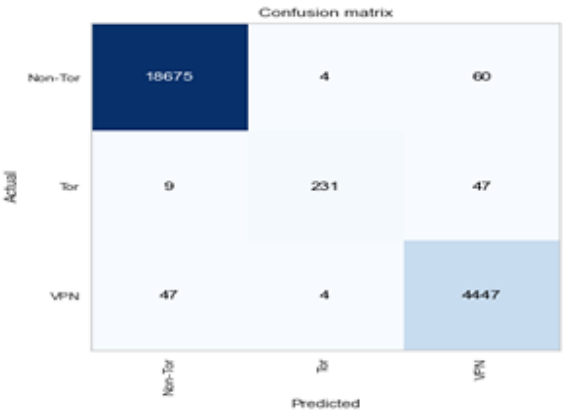

Figure 12. Ensemble Confusion Matrix

5. CONCLUSION AND FUTURE WORK

Reviewing the effectiveness of machine learning algorithms on the CIC Darknet 2020 dataset ultimately shows how critical it is to choose the correct model for cybersecurity applications. The ensemble method outperformed the other algorithms tested, with an accuracy of 99.28%. Despite the complex and diverse nature of darknet data, this comprehensive approach achieved superior results in accurately detecting occurrences, likely using Random Forest with other algorithms. Despite k-Nearest Neighbors (KNN)'s remarkable accuracy of 94.04%, larger datasets or real-time scaling may be beyond its capabilities due to its proximity-based technique and user-friendliness. Decision Trees (DT) displayed impressive performance with a 98.79% accuracy rate. Overfitting can be prevented with proper hyperparameter manipulation, but DT is interpretable. Random Forest (RF) outperformed individual decision trees in detecting intricate patterns in darknet activity, with an accuracy of 99.25%. The fact that ensemble techniques consistently provide the highest levels of accuracy demonstrates their potential for enhancing prediction performance via the integration of diverse perspectives. Cybersecurity relies heavily on this finding since darknet scenarios need complex and nuanced approaches for detecting and classifying suspicious behavior. Incorporating the best features of many models into an ensemble strategy yields comprehensive findings for darknet traffic analysis, according to the results. Security researchers and practitioners should take note of this discovery, which highlights the need to use intricate ensemble methodologies to improve the accuracy and reliability of threat detection systems. Research in the future may concentrate on enhancing ensemble techniques, removing potential biases, and studying interpretability difficulties to make them more applicable to real-world cybersecurity scenarios. In sum, the results show that machine learning is essential for cybersecurity efforts and that ensemble techniques are the gold standard for more accurate and resilient darknet traffic classification.

REFERENCES

[1] Gayard, Laurent (2018). Darknet: Geopolitics and Uses. Hoboken, NJ: John Wiley & Sons. p. 158. ISBN 9781786302021.

[2] Pradhan, Sayam (2020). "Anonymous". The Darkest Web: The Dark Side of the Internet. India. p. 9. ISBN 9798561755668.

[3] Press Foundation, Freedom of the. "SecureDrop". GitHub. Freedom of the Press Foundation. Retrieved 28 January 2019.

[4] Abu Al-Haija, Q. Top-Down Machine Learning-Based Architecture for Cyberattacks Identification and Classification in IoT Communication Networks. Front. Big Data 2022, 4, 782902.

[5] Abu Al-Haija, Q.; Al-Badawi, A. Attack-Aware IoT Network Traffic Routing Leveraging Ensemble Learning. Sensors 2021, 22, 241.

[6] Zavrak, S.; Iskefiyeli, M. Anomaly-Based Intrusion Detection from Network Flow Features Using Variational Autoencoder. IEEE Access 2020, 8, 108346–108358.

[7] Wang, L.; Mei, H.; Sheng, V.S. Multilevel Identification and Classification Analysis of Tor on Mobile and PC Platforms. IEEE Trans. Ind. Inform. 2021, 17, 1079–1088.

[8] Demertzis, K.; Tsiknas, K.; Takezis, D.; Skianis, C.; Iliadis, L. Darknet Traffic Big-Data Analysis and Network Management for Real-Time

International Journal of
Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 123-129 | ISSN XXXX-XXXX

Automating of the Malicious Intent Detection Process by a Weight Agnostic Neural Networks Framework. Electronics 2021, 10, 781.

[9] Mishra A, Gupta N, Gupta B. Defense mechanisms against DDoS attack based on entropy in SDN-cloud using POX controller. Telecommun Syst. 2021;77(1):47–62. doi: 10.1007/s11235-020-00747-w.

[10] Ozawa S, Ban T, Hashimoto N, Nakazato J, Shimamura J. A study of IoT malware activities using association rule learning for darknet sensor data. Int J Inf Secur. 2020;19(1):83–92. doi: 10.1007/s10207-019-00439-w

[11] Varghese, J.E.; Muniyal, B. A Pilot Study in Software-Defined Networking Using Wireshark for Analyzing Network Parameters to Detect DDoS Attacks. In Information and Communication Technology for Competitive Strategies (ICTCS 2020); Springer: Singapore, 2021; pp. 475–487.

[12] Han, C.; Shimamura, J.; Takahashi, T.; Inoue, D.; Takeuchi, J.I.; Nakao, K. Real-Time Detection of Global Cyberthreat Based on Darknet by Estimating Anomalous Synchronization Using Graphical Lasso. IEICE Trans. Inf. Syst. 2020, E103-D, 2113–2124.