Darknet Traffic Using Deep Learning

Muhammad Abusaqer and Quinn Sullivan

Department of Math and Computer Science

Minot State University

Minot, ND, USA

muhammad.abusaqer@minotstateu.edu

quinn.sullivan@minotstateu.edu

Abstract

This research investigates the potential of deep learning for the analysis of darknet traffic, a network operating outside the traditional internet and frequently associated with illegal activities such as drug dealing, trafficking, and exploitative content. The darknet also provides a secure communication and information exchange platform for privacy-conscious individuals. The study aims to classify darknet traffic into 8 categories - P2P, Audio-Streaming, Browsing, Video-Streaming, Chat, Email, File-Transfer, and VOIP - for accurate categorization of real-time applications and to support law enforcement agencies in detecting and preventing malicious activities. A custom-designed Artificial Neural Network (ANN) model was trained using the CIC-Darknet2020 dataset to perform multi-class classification of darknet traffic. The ANN model's performance was compared to two established machine learning algorithms, XGBoost and RandomForest. The results demonstrated that although the ANN model showed promise, it was outperformed by both XGBoost and RandomForest models. This paper presents a contribution by applying deep learning to the CIC-Darknet2020 dataset and comparing its performance with traditional machine learning models. The findings highlight the potential capabilities of deep learning models in analyzing darknet traffic and suggest avenues for future improvements.

1 Introduction

The Darknet, a hidden and encrypted part of the internet, has become a significant challenge for law enforcement and cybersecurity professionals due to its association with criminal activities and illicit services [1]. Operating on overlay networks such as Tor and I2P, the Darknet provides a high level of anonymity and privacy for its users, making it an attractive platform for illegal activities such as drug trafficking, hacking, and financial fraud [2]. The rapid growth and increasing complexity of Darknet ecosystems necessitate advanced monitoring and analysis tools to identify and combat these malicious activities [3]. Recent research has focused on developing novel methods for Darknet traffic classification, utilizing machine learning and deep learning techniques to detect and analyze suspicious activities and enhance cybersecurity efforts [4] [5]. As the Darknet continues to evolve, researchers and practitioners must adapt their approaches and develop innovative strategies to stay ahead of emerging threats and protect the integrity of online systems and networks.

This research paper focuses on the classification of a darknet dataset using both traditional machine learning and deep learning models. The results from the Artificial Neural Network (ANN) model will be compared with those from traditional machine learning models, namely RandomForest and XGBoost. The motivation behind this research is to explore the potential of machine learning in darknet classification and to gain an understanding of how various algorithms perform in this context. The study is based on the CIC-Darknet2020 dataset [6] [7].

2 Related Work

One of the relevant studies in the field of darknet traffic classification is the work by Iliadis and Kaifas in [7]. The authors explored the application of various machine learning models to classify darknet traffic effectively. The primary motivation behind their research was the growing importance of identifying and classifying darknet traffic for cybersecurity purposes, as understanding the nature of such traffic can provide valuable insights into potential threats and vulnerabilities. In their study, Iliadis and Kaifas [7] experimented with a range of machine learning algorithms, including Decision Trees, Random Forests, k-Nearest Neighbors (k-NN), and others. They aimed to find the most accurate and efficient approach for classifying darknet traffic. Their research highlighted the need for advanced methods that can accurately distinguish between different types of darknet traffic, which can, in turn, contribute to improved cybersecurity measures and threat detection. The findings of [7] serve as a valuable reference for the current research, as they provide insights into the effectiveness of different machine learning techniques in the context of darknet traffic classification. The comparison of their results with the outcomes of the present study, which employs an Artificial Neural Network (ANN) model along with RandomForest and XGBoost classifiers, can shed light on the relative performance of deep learning approaches versus traditional machine learning methods in this domain.

In [8], DarknetSec, a novel self-attentive deep learning framework, has been proposed to improve darknet traffic classification and application identification. It employs a cascaded

model combining a 1D Convolutional Neural Network (CNN) and a bidirectional Long Short-Term Memory (Bi-LSTM) network to capture local spatial-temporal features from packet payloads. The self-attention mechanism, integrated into the feature extraction network, uncovers hidden relationships among the extracted content features. DarknetSec also extracts side-channel features from payload statistics to enhance performance. Evaluated on the CICDarknet2020 dataset, DarknetSec outperforms state-of-the-art methods, achieving a multiclass accuracy of 92.22% and a macro-F1-score of 92.10%. It also maintains high accuracy in other encrypted traffic classification tasks.

Almomani proposed a novel darknet traffic analysis and classification system based on modified stacking ensemble learning algorithms in [9]. The study focused on utilizing stacking ensemble learning, a machine learning technique that combines multiple learning mechanisms to generate more accurate predictions. The system was evaluated on a dataset containing over 141,000 records from CIC-Darknet 2020, the same dataset used in this study. The experimental results showcased the classifiers' ability to distinguish between benign and malignant traffic, with accuracy rates exceeding 99% during the training phase and 97% in the testing phase. The study utilized a two-tiered learning stacking scheme that incorporated both individual and group learning, with three base learning methods, including neural networks, random forests, and support vector machines. The ensemble approach demonstrated better performance compared to single techniques, particularly when handling small historical windows, suggesting that the system becomes more robust and accurate as data grows. Despite limitations related to performance and privacy concerns, the proposed system offers a promising direction for future research in darknet traffic classification and analysis, exploring various ensemble schemes and methodologies to enhance its effectiveness against different types of attacks [9].

Sridhar and Sanagavarapu in [5] conducted a study on darknet traffic classification, aiming to enhance network security by detecting threats or risks. The authors used the standard CIC-Darknet2020 dataset, which contains instances of both benign and darknet traffic. They performed feature importance analysis using the Chi-Squared statistical score for feature selection and addressed the imbalance of classes by applying oversampling with Conditional Generative Adversarial Networks (Conditional GANs). The multi-class classification of traffic encryption types was carried out using the Random Forest classifier, achieving a 97.87 F1-Score for traffic encryption classification. In their conclusion, they suggested exploring feature extraction through Principal Component Analysis and employing Recurrent Neural Networks for detecting attacks over time as potential future work.

The paper [10] presents an approach to darknet traffic analysis using a weight agnostic neural network (WANN) framework for real-time detection of malicious intent. The authors propose a method that leverages big-data analysis techniques and network management practices to process and classify darknet traffic data. They aim to improve the efficiency and effectiveness of malicious intent detection in darknet traffic by using a WANN framework, which is capable of learning and generalizing from limited training data. This study contributes to the ongoing research on darknet traffic classification and detection of malicious activities. The proposed WANN framework offers a promising

approach to enhance cybersecurity efforts by automating the process of detecting threats in real-time.

Al-Qatf et al. in [11] proposed a deep learning approach for network intrusion detection that combines a sparse autoencoder with a Support Vector Machine (SVM). The authors recognized the importance of effective network intrusion detection systems to counter the growing number of cyber threats. They introduced a deep learning method that leverages a sparse autoencoder to extract relevant features from network traffic data and an SVM classifier to categorize the traffic as normal or malicious. The proposed system was trained and tested on a dataset consisting of various network traffic instances. The results indicated that the combined deep learning approach outperformed traditional machine learning techniques in terms of detection accuracy and generalization performance. This research highlights the potential of hybrid deep learning methods in enhancing network intrusion detection and providing more effective solutions for cybersecurity professionals.

3 Proposed Methodology

3.1 Dataset

The dataset used in this research is the CIC-Darknet2020 dataset, obtained from the Canadian Institute for Cybersecurity [6]. The Darknet-2020 dataset was chosen over other available datasets due to its recency and relevance to the research objectives. The dataset encompasses a mix of data types, including numerical, categorical, and text features. The 'Label' column in the dataset contains eight distinct classes, which are P2P, Audio-Streaming, Browsing, Video-Streaming, Chat, Email, File-Transfer, and VOIP. These classes represent different types of darknet traffic that the models aim to classify in this study.

3.2 Dataset Preprocessing

The original dataset comprised 141,530 rows with 85 columns. However, given the computational resource constraints faced by the authors, a random subset of 8,000 rows was selected for the experiment. During the preprocessing stage, the authors encountered issues with certain columns that negatively impacted the model's performance. Consequently, these problematic columns were dropped, along with a few others that did not contribute significantly to the output. Additionally, several preprocessing steps were applied to handle missing values and encode categorical features. Large entries in the dataset were replaced with NaN, missing values in numeric columns were imputed using mean imputation, and non-numeric columns with missing values were imputed using the most frequent (mode) imputation method. All non-numeric features, excluding the 'Label' column, were encoded as integers. The target variable ('Label') was encoded as integers using the LabelEncoder from the scikit-learn library.

3.3 Machine Learning Models

In this research, three different models were employed: an Artificial Neural Network (ANN) model, a RandomForest classifier, and an Extreme Gradient Boosting (XGBoost) classifier. The comparison aimed to evaluate the performance of the deep learning approach, as represented by the ANN model, against the well-established machine learning techniques of RandomForest and XGBoost in the context of darknet traffic classification. For all models, the dataset was split into training (80%) and testing (20%) sets, with the features scaled using the StandardScaler from the scikit-learn library.

3.3.1 Artificial Neural Network Model

Artificial Neural Networks (ANNs) are a class of machine learning algorithms that mimic the structure and function of the human brain, allowing them to learn patterns from data [12]. The neural network was designed with three layers: an input layer with 64 nodes and a ReLU activation function, a hidden layer with 32 nodes and a ReLU activation function, and an output layer with a softmax activation function [13] [14] [15] [16]. The model was trained for 10 epochs with a batch size of 32, and the optimizer used was the Adam optimizer [17].

3.3.2 RandomForest Model

RandomForest is an ensemble learning method that constructs multiple decision trees and combines their output to improve overall model performance and reduce overfitting [18] [19].

The RandomForest classifier in this research was instantiated with 100 estimators and a random state of 42 to ensure reproducibility. The model was then trained on the training set and used to make predictions on the testing set.

3.3.3 XGBoost Model

Extreme Gradient Boosting (XGBoost) is an advanced implementation of gradient boosting machines that uses a combination of decision trees and optimization techniques to improve model accuracy and speed [20] [21] ,

The XGBoost classifier [19] [20] in this research was trained on the training set, with the 'use_label_encoder' parameter set to 'False' and the 'eval_metric' parameter set to 'mlogloss'. After training, the classifier was used to make predictions on the testing set.

For all three models, evaluation metrics, including accuracy, precision, recall, and F1-score, were calculated to assess their performance in the context of darknet traffic classification.

3.4 Evaluation Metrics

To compare the performance of the Artificial Neural Network (ANN) model with RandomForest and XGBoost models in the context of darknet traffic classification, the following evaluation metrics were employed: accuracy, precision, recall, and F1-score.

3.4.1 Accuracy

Accuracy is the proportion of correct predictions (both true positives and true negatives) made by the model out of the total number of instances in the dataset. It is a commonly used metric to measure the overall performance of a classifier [22].

Accuracy = (True Positives + True Negatives) / (True Positives + False Positives + True Negatives + False Negatives)

However, accuracy alone may not be an appropriate measure when the data is imbalanced, as it can be misleading when the majority of the instances belong to one class [23].

3.4.2 Precision

Precision is the proportion of true positives out of the total number of instances predicted as positive by the model. In other words, it measures the ability of the classifier to correctly identify the positive instances among all the instances predicted as positive [24].

Precision = True Positives / (True Positives + False Positives)

Precision is a useful metric in the context of darknet traffic classification when the cost of false positives is high, such as in identifying malicious activities where incorrectly labeling benign traffic can lead to unnecessary investigations or countermeasures [4].

3.4.3 Recall

Recall, also known as sensitivity or true positive rate, is the proportion of true positives out of the total number of actual positive instances in the dataset. It measures the ability of the classifier to identify all the positive instances [24] [4].

Recall = True Positives / (True Positives + False Negatives)

3.4.4 F1-score

F1-score is the harmonic mean of precision and recall, providing a single metric that balances both precision and recall [25]. It is particularly useful when dealing with imbalanced datasets, as it takes into account both false positives and false negatives [26].

F1-score = 2 * (Precision * Recall) / (Precision + Recall)

An F1-score of 1 indicates perfect precision and recall, while an F1-score of 0 indicates that either precision or recall (or both) are zero.

4 Experiment

4.1 Results

In this experiment, the performance of a custom-designed Artificial Neural Network (ANN) was assessed and compared to two established machine learning models, XGBoost and RandomForest, with respect to their classification accuracy. To ensure a fair comparison, all models were initially trained on a smaller dataset and subsequently tested on a larger dataset containing 8,000 randomly selected rows. The performance metrics for each model are as follows:

	Accuracy	**Precision**	**Recall**	**F1-Score**
Neural Network	0.74	0.74	0.74	0.71
XGBoost	0.83	0.83	0.83	0.83
RandomForest	0.81	0.80	0.81	0.80

Table 01: Experiments Results

5 Discussion

The results demonstrate that both XGBoost and RandomForest models outperformed the custom-designed neural network in terms of accuracy, precision, recall, and F1-scores for classifying darknet traffic. Although deep learning models hold great potential, the neural network did not surpass the performance of the XGBoost and RandomForest models in this specific classification task.

6 Future Work

In future work, the authors plan to enhance the ANN model by adding more nodes and hidden layers and continue experimenting until satisfactory results are achieved compared to the XGBoost classifier. Additionally, the entire dataset will be utilized for a more comprehensive analysis.

7 Conclusion

In conclusion, this study investigated the performance of a custom-designed ANN model in comparison to established machine learning models, XGBoost and RandomForest, for darknet traffic classification. The results showed that the ANN model did not outperform the other two models in this specific task. Further experimentation and improvements to the ANN model are necessary to achieve better classification results.

References

[1] M. Chertoff and T. Simon, "The impact of the dark web on internet governance and cybersecurity," Global Commission on Internet Governance, Paper Series 6, 2015. [Online]. Available: https://www.cigionline.org/static/documents/gcig_paper_no6.pdf

[2] G. Owen and N. Savage, "The Tor dark net /," 2015, Accessed: Feb. 28, 2023. [Online]. Available: https://policycommons.net/artifacts/1223621/the-tor-dark-net/1776697/

[3] D. Moore and T. Rid, "Cryptopolitik and the Darknet," Survival, vol. 58, no. 1, pp. 7–38, Jan. 2016, doi: 10.1080/00396338.2016.1142085.

[4] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016, doi: 10.1109/COMST.2015.2494502.

[5] S. Sridhar and S. Sanagavarapu, "DarkNet Traffic Classification Pipeline with Feature Selection and Conditional GAN-based Class Balancing," in 2021 IEEE 20th International Symposium on Network Computing and Applications (NCA), Nov. 2021, pp. 1–4. doi: 10.1109/NCA53618.2021.9685743.

[6] "Darknet 2020 | Datasets | Research | Canadian Institute for Cybersecurity | UNB." https://www.unb.ca/cic/datasets/darknet2020.html (accessed Feb. 09, 2023).

[7] L. A. Iliadis and T. Kaifas, "Darknet Traffic Classification using Machine Learning Techniques," in 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST), Jul. 2021, pp. 1–4. doi: 10.1109/MOCAST52088.2021.9493386.

[8] J. Lan, X. Liu, B. Li, Y. Li, and T. Geng, "DarknetSec: A novel self-attentive deep learning method for darknet traffic classification and application identification," Computers & Security, vol. 116, p. 102663, May 2022, doi: 10.1016/j.cose.2022.102663.

[9] A. Almomani, "Darknet traffic analysis, and classification system based on modified stacking ensemble learning algorithms," Information Systems and e-Business Management, pp. 1–32, Feb. 2023, doi: 10.1007/s10257-023-00626-2.

[10] K. Demertzis, K. Tsiknas, D. Takezis, C. Skianis, and L. Iliadis, "Darknet Traffic Big-Data Analysis and Network Management for Real-Time Automating of the Malicious Intent Detection Process by a Weight Agnostic Neural Networks Framework," Electronics, vol. 10, no. 7, Art. no. 7, Jan. 2021, doi: 10.3390/electronics10070781.

[11] M. Al-Qatf, Y. Lasheng, M. Al-Habib, and K. Al-Sabahi, "Deep Learning Approach Combining Sparse Autoencoder With SVM for Network Intrusion Detection," IEEE Access, vol. 6, pp. 52843–52856, 2018, doi: 10.1109/ACCESS.2018.2869577.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, Art. no. 7553, May 2015, doi: 10.1038/nature14539.

[13] G. Zhang, B. Eddy Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks:: The state of the art," International Journal of Forecasting, vol. 14, no. 1, pp. 35–62, Mar. 1998, doi: 10.1016/S0169-2070(97)00044-7.

[14] R. Tadeusiewicz, "Neural networks: A comprehensive foundation: by Simon HAYKIN; Macmillan College Publishing, New York, USA; IEEE Press, New York, USA; IEEE Computer Society Press, Los Alamitos, CA, USA; 1994; 696 pp.; $69–95; ISBN: 0-02-352761-7." Pergamon, 1995.

[15] C. M. Bishop and P. of N. C. C. M. Bishop, Neural Networks for Pattern Recognition. Clarendon Press, 1995.

[16] Z. Hu, Z. Zhang, H. Yang, Q. Chen, and D. Zuo, "A deep learning approach for predicting the quality of online health expert question-answering services," Journal of Biomedical Informatics, vol. 71, pp. 241–253, Jul. 2017, doi: 10.1016/j.jbi.2017.06.012.

[17] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." arXiv, Jan. 29, 2017. doi: 10.48550/arXiv.1412.6980.

[18] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[19] A. Liaw and M. Wiener, "Classification and Regression by randomForest," vol. 2, 2002.

[20] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[21] X. He et al., "Practical Lessons from Predicting Clicks on Ads at Facebook," in Proceedings of the Eighth International Workshop on Data Mining for Online Advertising, in ADKDD'14. New York, NY, USA: Association for Computing Machinery, Aug. 2014, pp. 1–9. doi: 10.1145/2648584.2648589.

[22] J. D. Kelleher, B. M. Namee, and A. D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies. MIT Press, 2020.

[23] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[24] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[25] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, p. 6, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[26] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." arXiv, Oct. 10, 2020. doi: 10.48550/arXiv.2010.16061.