International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 13 Issue 6, June, 2024, Impact Factor: 8.285

Hybrid Shield: DDoS Attack Detection with Random Forest and KNN Fusion

Prof. Anjali S. More[1], Aditya Murke[2], Aachal Harinkhede[3], Sakshi Mahajan[4], Bhagyashree Patil[5]

[1]Assistant Professor, Department of Computer Engineering, SRTTC FOE, MH, India,
[2,3,4,5] Student, Department of Computer Engineering, SRTTC FOE, MH, India,
(Savitribai Phule Pune University)

ABSTRACT

In the rapidly evolving landscape of cybersecurity, Distributed Denial of Service (DDoS) attacks pose a significant threat to network stability and data integrity. This paper presents a novel approach to DDoS attack detection by hybridizing the Random Forest (RF) and K-Nearest Neighbors (KNN) algorithms, aiming to leverage their complementary strengths for enhanced detection accuracy and robustness. The proposed hybrid model integrates the high-dimensional data handling capability and feature importance evaluation of Random Forest with the simplicity and effectiveness of KNN in local decision-making. This synergistic combination is designed to improve the precision and recall rates of DDoS attack detection systems. Our methodology involves pre-processing network traffic data to extract relevant features, followed by the application of the hybrid RF-KNN model. The Random Forest algorithm is employed to perform initial classification and feature importance analysis, reducing the feature space and enhancing the efficiency of the subsequent KNN algorithm. The KNN algorithm then refines the classification by evaluating the local neighborhood of the data points, ensuring a more accurate detection of DDoS attacks. Extensive experiments are conducted on benchmark datasets to evaluate the performance of the proposed model. The results demonstrate that the hybrid RF-KNN model outperforms traditional individual algorithms, achieving higher detection rates, reduced false positives, and improved computational efficiency. This research contributes to the field of network security by providing an effective and scalable solution for real-time DDoS attack detection, highlighting the potential of hybrid machine learning approaches in cybersecurity applications.

Keywords: DDoS attack detection, Random Forest, K-Nearest Neighbors, hybrid model, network security, machine learning.

INTRODUCTION

In the rapidly evolving landscape of digital technology, ensuring the security and integrity of network systems has become a paramount concern. Among the myriad of cyber threats, Distributed Denial of Service (DDoS) attacks stand out due to their sheer capacity to disrupt services, cause significant financial loss, and erode consumer trust. As these attacks grow in sophistication and frequency, traditional detection mechanisms often struggle to keep pace. This research paper addresses this critical issue by proposing an innovative solution that leverages the combined strengths of two powerful machine learning algorithms: Random Forest and k-Nearest Neighbors (KNN).

By hybridizing these algorithms, we aim to create a more robust and accurate detection system capable of identifying DDoS attacks with higher precision and speed. The Random Forest algorithm, known for its high accuracy and ability to handle large datasets with numerous features, complements the KNN algorithm's simplicity and effectiveness in pattern recognition. Together, they form a synergistic model that enhances the detection capabilities beyond what either could achieve alone. This paper will delve into the specifics of the hybrid model, exploring the methodology behind its development, the dataset utilized for training and testing, and the performance metrics that demonstrate its efficacy. By integrating these advanced machine learning techniques, our goal is to contribute a significant advancement in the ongoing battle against cyber threats, providing a reliable tool for network administrators and cybersecurity professionals to safeguard their systems against the ever-present menace of DDoS attacks

OBJECTIVE

The primary objective of this research is to investigate the effectiveness of ensemble learning techniques in improving the performance of intrusion detection systems based on network traffic analysis. Specifically, we focus on the integration of KNN and Random Forest classifiers within a voting ensemble framework. By combining the strengths of

both algorithms, we aim to create a more resilient and accurate IDS capable of effectively identifying and mitigating diverse cyber threats.

CONTRIBUTION

This paper makes several contributions to the field of cybersecurity and intrusion detection:

1. We propose a novel approach to enhancing intrusion detection through the use of ensemble learning techniques, specifically combining KNN and Random Forest classifiers.
2. We demonstrate the efficacy of the proposed ensemble model in accurately detecting various types of network intrusions, including DoS attacks, malware infections, and port scans.
3. We conduct a comprehensive evaluation of the ensemble model using real-world network traffic data, comparing its performance against individual classifiers and traditional signature-based IDS approaches.
4. We provide insights into the strengths and limitations of ensemble learning for intrusion detection and discuss potential avenues for future research and development in this area.

METHODOLOGY

1. Data Loading: The code starts by loading the dataset from the specified file path using Pandas. The dataset is assumed to be in CSV format.
2. Identifying Textual Columns: It identifies columns with object dtype (which usually indicates textual or categorical data).
3. Processing Textual Columns: LabelEncoder from scikit-learn is used to convert textual data into numerical format.
4. Converting Target Column: The target column (assumed to be the first textual column) is converted to categorical type.
5. Converting Numeric Columns and Handling Missing Values: Numeric columns are converted to numeric data type and missing values are filled with the mean of the respective columns.
6. Normalization: MinMaxScaler from scikit-learn is used to normalize numeric features to a range between 0 and 1.
7. Feature Selection: Domain-specific features are selected based on domain knowledge or feature importance techniques.
8. Data Splitting: The dataset is split into training and testing sets using train_test_split from scikit-learn.
9. Handling Class Imbalance: SMOTE (Synthetic Minority Over-sampling Technique) is applied to balance the training set.
10. Saving Data: Balanced training set, balanced training labels, test set, and test labels are saved to specified file paths.
11. Creating Classifiers: KNN and Random Forest classifiers are created.
12. Creating Voting Classifier: A Voting Classifier is created to combine the predictions from KNN and Random Forest classifiers.
13. Training the Voting Classifier: The Voting Classifier is trained on the balanced training data.
14. Making Predictions: Predictions are made using the trained Voting Classifier on the test data.
15. Evaluating Performance: Confusion matrix and classification report are generated to evaluate the performance of the Voting Classifier.

KNN

K-Nearest Neighbors (KNN) is a non-parametric, lazy learning algorithm used for classification and regression. It operates by finding the K training samples closest in distance to a new sample and makes predictions based on the majority label (for classification) or the average (for regression) of these neighbors.

KEY CHARACTERISTICS

- Non-parametric: KNN does not assume a fixed form for the underlying data distribution, making it flexible and easy to use.
- Lazy Learning: It doesn't build a model during training but instead memorizes the training dataset, making predictions slower but training faster
- Distance Metrics: Commonly used distance metrics include Euclidean, Manhattan, and Minkowski distances, which are pivotal in determining the "nearness" of points.

RANDOM FOREST CLASSIFICATION

RF algorithm [8-10] uses the committee (ensemble) of the decision trees. The high quality classification in the RF algorithm is achieved by combining a large number of simple classifiers (decision trees). The final classification result

is obtained on the basis of the responses aggregation of many trees. The RF algorithm (in comparison with one decision tree) reduce the problem of overfiting and improve the classification quality. The RF algorithm combines the ideas of the bagging, the bootstrap aggregating and the random subspace method. In the RF algorithm, as in the bagging, the training of classifiers occurs independently at the different subsets of the training set that solves the problem of building the same trees on the same dataset. As a result, the class of any object will be equal to the class for which the majority of the trees voted, on the assumption that one tree has one vote. The defining the values of the parameters is important in the case of the RF algorithm implementing. The main parameters of the RF algorithm are the following: the number of trees in the forest, the number of characteristics to consider when looking for the best split, the maximum depth of the tree, the splitting criterion.

HYBRIDIZATION OF KNN AND RANDOM FOREST

a) Motivation for Hybridization

Combining Random Forest and KNN leverages the strengths of both algorithms, aiming to enhance the accuracy and robustness of DDoS attack detection.

Random Forest (RF): A powerful ensemble learning method based on decision trees that provides high accuracy, handles large datasets well, and is resilient to over fitting.

KNN: Complements RF by providing a simple and effective way to capture local patterns in the data.

ADVANTAGES OF HYBRID MODEL.

Improved Accuracy: Combining RF's global pattern recognition with KNN's local pattern recognition can lead to more accurate detection of DDoS attacks.

Robustness: The hybrid approach is more robust to noise and outliers in the data, as RF can handle noisy features while KNN focuses on local data points.

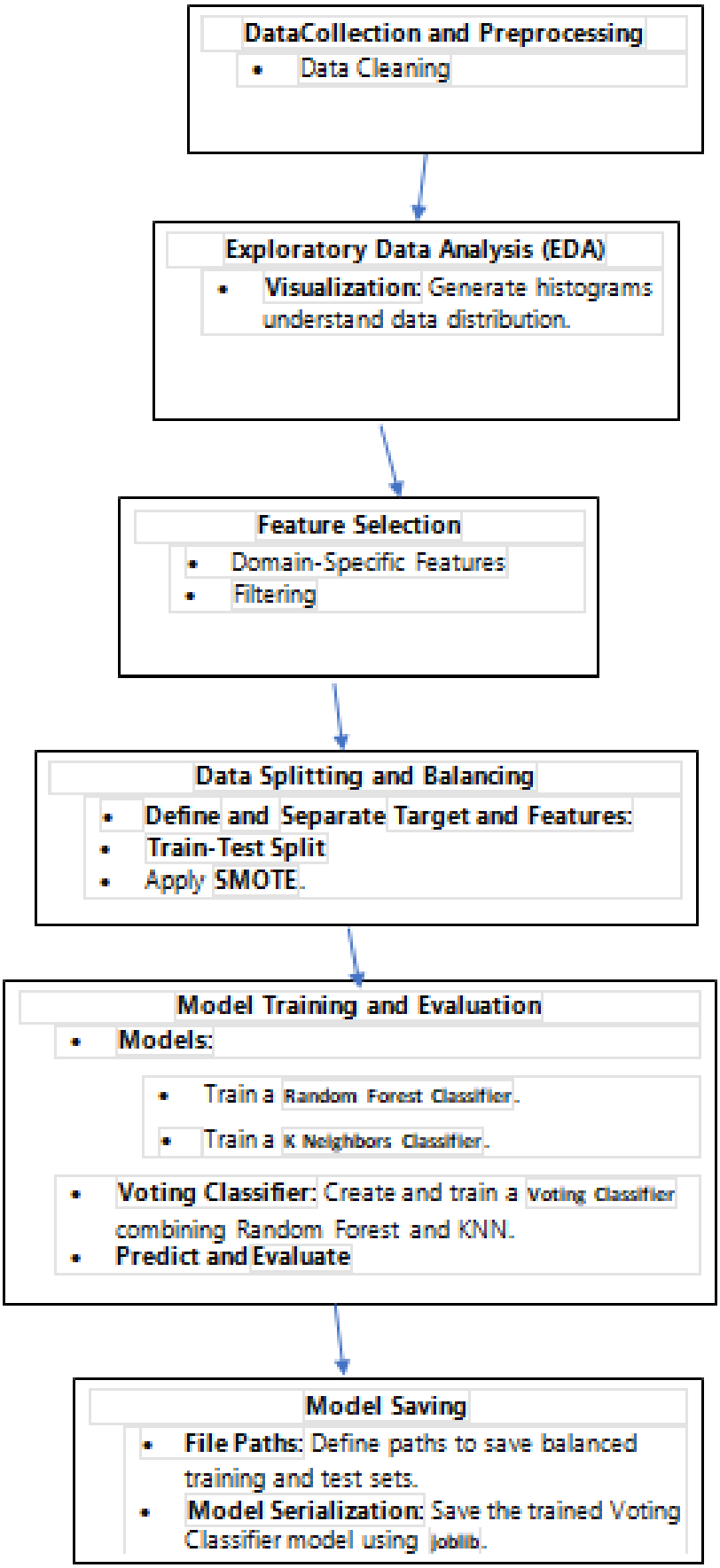
Scalability: While KNN's lazy learning can be computationally intensive, RF can preprocess and filter features to make KNN more scalable to large datasets.

VOTING CLASSIFIER RESULTS

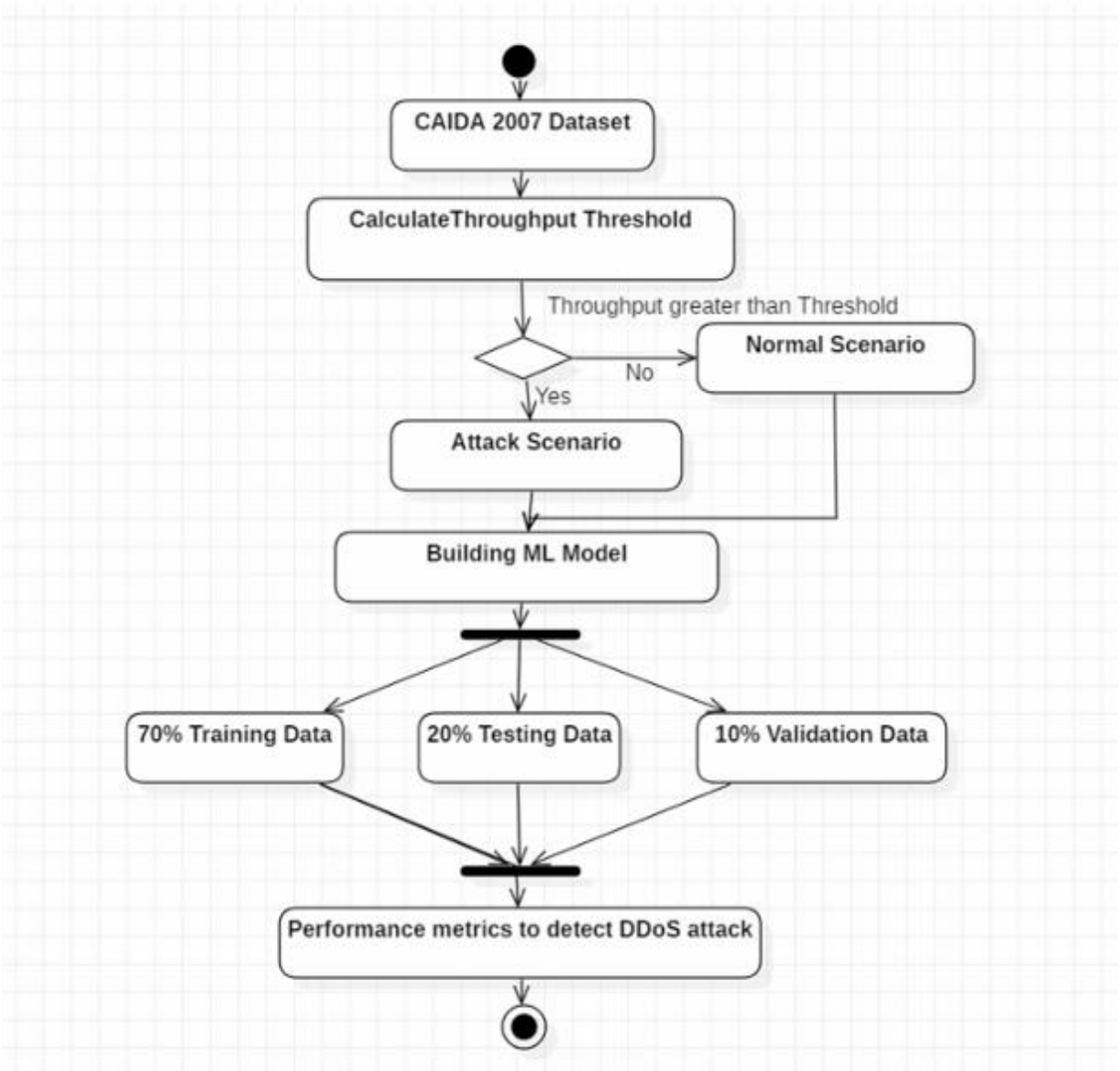
Voting Classifier Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	13566
1	0.52	0.39	0.44	512
2	0.26	0.51	0.34	205
3	0.41	0.78	0.53	865
4	1.00	0.99	1.00	16997
5	0.11	0.17	0.14	83
6	0.62	0.62	0.62	385
7	0.37	0.09	0.14	1458
8	0.41	0.41	0.41	263
9	0.57	0.24	0.34	1188
10	0.30	0.69	0.41	87
11	0.34	0.26	0.29	93
12	0.96	0.99	0.98	6907
13	1.00	0.99	1.00	13857
14	0.60	0.88	0.71	2546
15	0.90	0.57	0.70	1248
16	0.07	0.14	0.09	7
17	0.16	0.62	0.25	8
accuracy			0.93	60275
macro avg	0.53	0.57	0.52	60275
weighted avg	0.93	0.93	0.92	60275

SYSTEM ARCHITECTURE



ACTIVITY DIAGRAM FOR MODEL



CONCLUSION

In this research, we presented a novel approach for detecting Distributed Denial of Service (DDoS) attacks by hybridizing the K-Nearest Neighbors (KNN) and Random Forest models. Our aim was to leverage the strengths of both algorithms to enhance detection accuracy and reduce false positives in network security systems.

The hybrid model combines the simplicity and efficiency of KNN with the robustness and high performance of Random Forests. KNN, being a lazy learner, provides quick predictions based on proximity to other data points, making it effective in handling diverse network traffic patterns

REFERENCES

[1]. Vidyaev I G, Ivashutenko A S, Samburskaya M A. Smart Grid Concept As A Modern Technology For The Power Industry Development[C]// 2017:012173.
[2]. Huang H B, Hong L, Chang-Yue Y U, et al. Analysis on Ukraine Power Grid Blackout and Its Enlightenment of ICS in China[J]. Standard Science, 2016.
[3]. JianyeHao, Eunsuk Kang, Jun Sun, Zan Wang, “An Adaptive Markov Strategy for Defending Smart Grid False Data Injection from Malicious Attackers”, IEEE Transactions on Smart Grid. Sept. 2016.
[4]. JiaxuanFei,TaoZhang,YuanyuanMa,Cheng Zhou. A DDoS attack detection method for power grid industrial control system based on BF-DT-CUSUM algorithm[J]. Telecommunications Science.2015 (12).

[5]. Yanan Sun, Xiaohon Guan, Ting Liu, Yang Liu, "A cyber-physical monitoring system for attack detection in smart grid", Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on, Turin, Italy, Dec. 2014.
[6]. Mina Rahbari and Mohammad Ali JabreilJamali, "Efficient Detection of Sybil Attack Based on Cryptography in VANET," IJNSA, Vol.3, No.6, November 2011.
[7]. Mohamed Salah Bouassida, Gilles Guette, Mohamed Shawky, and Bertrand Ducourthial, "Sybil Nodes Detection Based on Received Signal Strength Variations within VANET," International Journal of Network Security, Vol.9, No.1, PP.22- 33, July 2009.
[8]. Yi P, Zhu T, Zhang Q, et al. A denial of service attack in advanced metering infrastructure network[C]// IEEE International Conference on Communications. IEEE, 2015:1029-1034.
[9]. Wang K, Du M, Maharjan S, et al. Strategic Honeypot Game Model for Distributed Denial of Service Attacks in the Smart Grid[J]. IEEE Transactions on Smart Grid, 2017, PP(99):1-1.
[10]. Pooja B, Pai M M M, Pai R M, et al. Mitigation of internal and external DoS attack against signature-based authentication in VANETs[C]// Computer Aided System Engineering.IEEE, 2014:152-157.
[11]. Saxena H, Richariya V. Disturbance detection in the KDD99 dataset using SVM-PSO and Feature Reduction with Information Gain[J].International Journal of Computer Applications, 2014, 98(6):25-29
[12]. Sousa, P. H. F.; Nascimento, N. M. M.; Almeida, J. S.; Rebouças Filho, P. P. and Albuquerque, V. H. C. (2019). Intelligent incipient fault detection in wind turbines based on an industrial IoT environment. Journal of Artificial Intelligence and Systems, 1, 1–19.