IMPROVING THE EFFICACY OF NETWORK SECURITY BASED ON DIMENSIONALITY REDUCTION TECHNIQUES

Hoang Thi Phuong
University of Economics - Technology for Industries, Hanoi, Vietnam

ARTICLE INFORMATION

Journal: *Vinh University Journal of Science Natural Science, Engineering and Technology*
p-ISSN*: 3030-4563*
e-ISSN: *3030-4180*

Volume: 53
Issue: 2A
****Correspondence:*** *htphuong@uneti.edu.vn*

Received: *03 January 2024*
Accepted: *28 January 2024*
Published: *20 June 2024*

Citation:
Hoang Thi Phuong (2024). Improving the efficacy of network security based on dimensionality reduction techniques. ***Vinh Uni. J. Sci.*** *Vol. 53 (2A), pp. 5-14 doi: 10.56824/vujs.2024a001*

ABSTRACT

This paper focuses on proposing a network intrusion detection model applying fundamental machine learning techniques to enhance early detection of network intrusions (rapid detection of attack behaviors) for improved efficiency in preventing network attacks. The system must still ensure technical accuracy in providing high-precision alerts. The research employs several dimensionality reduction techniques to detect abnormal network intrusions caused by Distributed Denial of Service (DDoS) attacks. The proposed model aims to reduce computation time for early attack detection. The results show that the proposed system performs best across all three datasets through the combination of the KNN algorithm and the Feature Importance dimensionality reduction technique. After calculating and returning the number of important features in attack detection using the Importance technique, the performance of the KNN algorithm is enhanced. By retaining only important features, as the dimensionality of the data decreases, the computation speed of KNN increases. Therefore, although the accuracy may slightly decrease, the computation time is significantly reduced. This is acceptable for practical purposes.

Keywords: Network attack; DdoS; machine learning; deep learning; dimensionality reduction.

1. Introduction

The biggest challenge in combating DDoS is the need for early detection of attacks and minimizing the impact as quickly as possible. Research has shown that the inefficiency of detecting and mitigating DDoS attacks is directly related to incorrect system configurations and the time-consuming nature due to the lack of dynamic traffic monitoring tools on the network without human supervision. Traditional intrusion detection methods are mainly divided into anomaly detection and signature-based detection. Anomaly detection primarily relies on expert knowledge and inference methods, with statistical methods and the Bayesian algorithm being representative algorithms used. While these methods generally help detect and counter network attacks reasonably well, given the advancements in technology and the increasing sophistication of attack techniques, they are increasingly challenged to prevent and detect attacks effectively.

Some recent publications related to this research include: the D-FACE method [1-2]; a technique based on the HTTP protocol [3-4]; Multiple-Features-Based Constrained-K-Means technique [5-7]; K-nearest neighbor classifier (KNNC) method [9-11]. These methods have advantages such as being able to detect DDoS attacks but demand a high level of IPS compatibility, limiting their use for general solutions. Additionally, these methods seem impractical in systems that require automated mitigation, especially in production environments that do not support high sampling rates or computationally expensive detection in real-time as the number of concurrent processes increases.

In this study, the authors propose building a system that functions as a sensor deployable anywhere on the network and classifies online traffic using a strategy based on machine learning algorithms. This helps classify random traffic patterns collected on network devices through the transmission protocol. The proposed method is compatible with the Internet infrastructure and does not require software or hardware upgrades. Additionally, user data privacy is ensured at all stages of system operation. The proposed system uses basic machine learning techniques to detect abnormal network intrusions (DDoS) and data dimensionality reduction techniques to eliminate less meaningful features in anomaly detection. The main goal of the proposed system is to reduce computation time for early detection of attacks while ensuring the accuracy of anomaly detection.

2. Research foundation and applications

2.1. Machine learning algorithms

Machine learning algorithms that can be applied for DDoS attack detection or in some intrusion detection systems include [2-5]:

- K-nearest neighbor (KNN): This is one of the simplest supervised learning algorithms that is effective in some cases. During training, the algorithm doesn't learn anything from the training data, which is why it falls into the lazy learning category. All computations will be performed when it is necessary to predict the outcome of new data. K-nearest neighbor can be applied to both types of supervised learning problems: Classification and Regression.

- Random Forests (RF): This is a supervised learning method that can handle both classification and regression problems. Essentially, Random Forests build a collection of decision trees and use a voting method to make decisions about the target variable to be predicted. The number of decision trees in RF is controlled as desired by the user.

- AdaBoost: This algorithm involves using very short decision trees (one level), called decision stumps, as weak learners continuously added to the ensemble. Each subsequent model tries to correct the predictions made by the model before it in the sequence. The resulting outcome is the best possible result.

- Support Vector Machine (SVM): This is a supervised learning method in pattern recognition models. It works well not only with linearly separable data but also with nonlinearly separable data.

2.2. Several dimensionality reduction techniques

Data dimensionality reduction includes the following common techniques [6-10].

2.2.1. Principal component analysis (PCA)

PCA is a statistical algorithm that uses orthogonal transformation to convert a dataset from a high-dimensional space to a new space with fewer dimensions, optimizing the representation of data variability. The transformation yields the following advantages for the data:

- Reduces the number of dimensions in the data space when it has a large number of dimensions.
- Constructs new coordinate axes capable of representing data equivalently and ensuring the variability of data on each new axis.
- Creates conditions for hidden data relationships to be discovered in the new space, which may be challenging to detect in the original space as these relationships may not be clearly evident.
- Ensures that the coordinate axes in the new space are always mutually orthogonal, even though the axes in the original space may not be orthogonal.

2.2.2. Feature Importance technique

Feature Importance refers to techniques that assign scores (importance indices) to input features based on their utility in predicting a target value. The importance indices are valuable and can be utilized in various situations within a predictive modeling problem, such as:

- Gaining a deeper understanding of the data.
- Enhancing the understanding of a model.
- Reducing the number of input features: This can be achieved by using the importance indices to select features for removal (low importance) or features to retain (high importance). This represents a type of feature selection and can streamline the modeling problem, speeding up the modeling process (removing features is known as dimensionality reduction) and, in some cases, improving the model's performance.

2.2.3. Univariate selection technique

This technique examines each feature individually to determine the strength of the relationship between the feature and the response variable. These methods are straightforward to implement and quite effective for gaining a better understanding of the data, identifying features with strong relationships to the response variable. Subsequently, a desired number of features can be retained as input features for a predictive model.

2.3. Problem statement and applications

According to [11], the proposed model for detecting online DDoS attacks is a hybrid system, described as Figure 1.

With the reference model in Figure 1, the authors focus on changing feature selection techniques for a network flow and employing various machine learning algorithms to create a different training system. They evaluate real-time execution time and the accuracy of DDoS attack detection among the tested models.

The proposed Intrusion Detection System (IDS) network is a hybrid system characterized by:

- Signature dataset (SDS): intrusion detection based on signatures.

- Employing recursive feature elimination with Cross Validation technique for selecting important features, followed by training with the Random Forest algorithm.

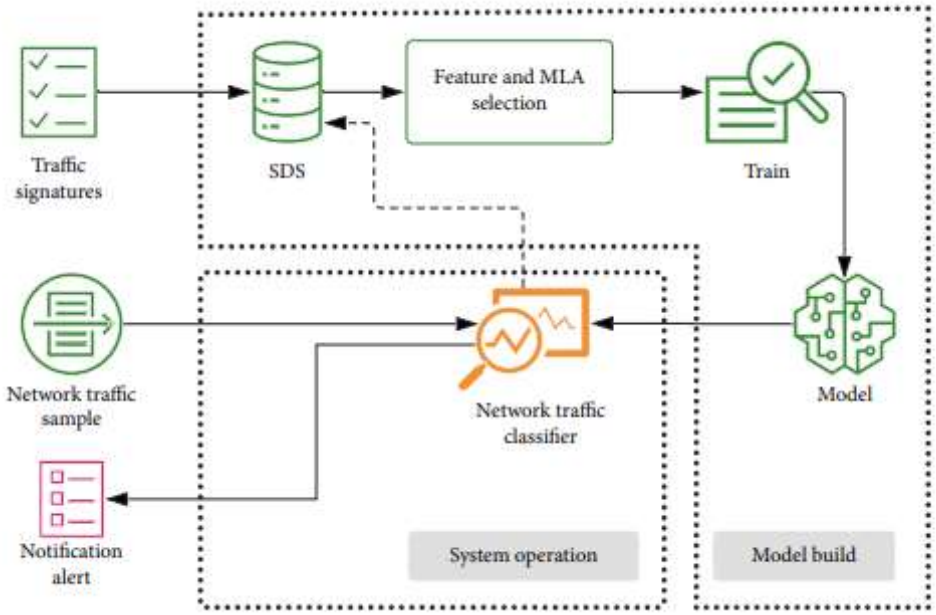

Figure 1: *The reference IDS system model*

The detailed proposed model includes blocks and functions as follows:

- *Feature Selection Block:* After receiving the data, this functional block focuses on using three different feature selection techniques: PCA, Feature Importance (using Extra tree and SelectFromModel from the scikit-learn library), and Univariate Selection (using SelectKBest with the chi-squared algorithm). Specifically:

+ *PCA:* recalculates the relationships between features and reduces the dimensionality of the data to the desired number. To find the appropriate number of dimensions, multiple experiments may be required.

+ *Feature Importance and Univariate Selection:* although they assess features differently, both techniques assign a "score" to each feature and then retain features with a score above a set threshold.

- *Machine learning block:* after reducing the data dimensionality with the aforementioned techniques, we obtain a new dataset with fewer dimensions than the original data. This dataset is then trained with each machine learning algorithm (KNN, AdaBoost, Random Forest, SVM) to classify attack traffic and normal traffic.

The proposed system offers several advantages as follows:

- Combining dimensionality reduction techniques accelerates processing speed to be suitable, ensuring accuracy when detecting abnormal traffic.

- When reducing the input data dimensionality for algorithms, the computation time of the algorithms can decrease, enhancing the ability to detect abnormalities early and improving the performance of DDoS attack prevention.

- With feature selection techniques like Feature Importance and Univariate Selection, after finding suitable features to retain, other redundant features can be eliminated. This helps monitor network traffic more effectively, only requiring observation and rule-setting for necessary features when new traffic enters the system. There's no need to monitor all features of the traffic, thereby minimizing the size of the input data for the intrusion detection system.

3. Implementation of evaluation on simulated network dataset

3.1. Dataset

The training dataset was generated according to the network model in Figure 2. VLAN 5, 6, 7 and 8 are used as victim machines. VLAN 100 is reserved for users of an academic unit. VLAN 10 is used as an attack server, monitored in VLAN 1. All networks have regular access to the Internet. The attack plan creates an attack every 30 minutes, totaling 48 attack events within 24 hours, starting from 00:00:00 and ending at 23:59:00. All attacks are carried out by the attack server (belonging to VLAN 10), and during that time, it does not transmit legitimate access traffic to the victims. The attack tools are parameterized to create sneaky low-volume, medium-volume, or light mode, and massive high-volume attacks. The initial dataset consists of 73 features for each record and is labeled as 'normal' and 'attack' explicitly.

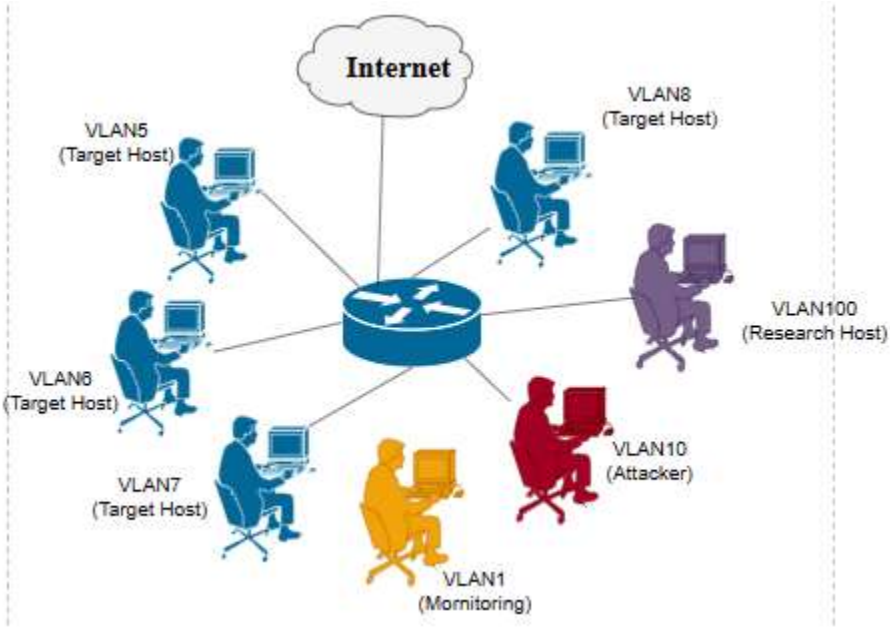

Figure 2: *Network structure for building the training dataset*

This dataset is constructed quite closely to a network environment operating in practice. Therefore, using this dataset for testing with the proposed model will help evaluate the system's effectiveness. However, this dataset is not large enough, consisting of 45,500 records (including 22,412 attacks and 23,088 normal records). This serves as a foundation for reference and building a larger dataset to further develop the system in the future [12-16].

3.2. Evaluation of the results

The results obtained are the average values over 15 training and testing iterations.

3.2.1. After performing dimensionality reduction using PCA

Table 1: *Simulated dataset with dimensionality reduction using PCA*

Algorithm	Initial Accuracy (%)	Accuracy after Dimensionality Reduction (%)	Initial Real-time Processing (ms)	Real-time Processing after Dimensionality Reduction (ms)
KNN	*95.88*	*97.22*	*7113.66*	*1528.69*
AdaBoost	*95.70*	*97.21*	*658.15*	*871.15*
Random Forest	*95.66*	*97.17*	*5091.79*	*7331.68*
SVM	*95.73*	*97.15*	*858.23*	*871.45*

The computation time after dimensionality reduction includes the processing time with the PCA technique. Due to the nature of this technique, which recalculates the relationships between features to reduce the dimensionality from a higher-dimensional space to a lower-dimensional one, the system needs to adjust the data dimensionality for each network traffic flow, followed by an analysis of whether the flow is valid or an attack. As observed in Table 1, except for the Random Forest and AdaBoost algorithms, which experienced a significant increase in execution time, the execution time for the remaining algorithms decreased considerably. This is because PCA transformed the dataset into a new one, altering the structure of the newly constructed trees compared to the original trees. In general, accuracy tends to decrease after dimensionality reduction. However, this reduction is acceptable considering the execution time. It is noted that the KNN algorithm is well-suited for this training dataset because, after being trained on the reduced dataset, it exhibits significantly faster execution time while maintaining relatively high accuracy.

3.2.2. After performing dimensionality reduction using Feature Importance

Dimensionality reduction was performed with the Feature Importance technique using Extra Tree to calculate the importance of each feature then employing the SelectFromModel algorithm to select features that meet user-defined conditions. The dimensionality reduction resulted in the elimination of 53 features, leaving only 20 features in use. The remaining features after applying Feature Importance: *'tcp_dataofs_median', 'tcp_dataofs_mean', 'tcp_flags_mean', 'ip_proto', 'ip_ttl_cv', 'tcp_flags_rte', 'ip_len_std', 'ip_ttl_std', 'tcp_flags_median', 'ip_len_entropy', 'sport_entropy', 'tcp_seq_mean', 'tcp_dataofs_rte', 'ip_len_cv', 'ip_ttl_cvq', 'tcp_ack_entropy', 'tcp_flags_cv', 'tcp_seq_entropy', 'tcp_ack_cvq', 'ip_len_mean'.*

According to the observations in Table 2, the obtained results are very promising. Although the accuracy of each model decreases slightly, the execution time of the model decreases significantly. Furthermore, when using the Feature Importance technique, we obtain results with only 20 features used. Therefore, when managing the IDS system, administrators only need to set rules to select the correct 20 features for an incoming data stream, reducing the time for data sampling and accelerating the processing speed for the

entire system. The dimensionality reduction method helps the intrusion detection model, applying basic machine learning algorithms, achieve the research objectives. The shorter the execution time, the earlier the detection of attacks, ensuring the accuracy of attack classification.

Table 2: *Simulated dataset with dimensionality reduction using Feature Importance*

Algorithm	Initial Accuracy (%)	Accuracy after Dimensionality Reduction (%)	Initial Real-time Processing (ms)	Real-time Processing after Dimensionality Reduction (ms)
KNN	*81.51*	*81.54*	*6927.16*	*798.21*
AdaBoost	*81.15*	*81.29*	*498.16*	*117.68*
Random Forest	*81.21*	*81.17*	*3896.57*	*1678.73*
SVM	*81.32*	*81.63*	*998.79*	*401.21*

3.2.3. After performing dimensionality reduction using Univariate Selection

Performing dimensionality reduction with the Univariate Selection technique using the chi-squared algorithm to calculate the chi-squared value for each feature in the dataset and sorting them in descending order. Then, set the parameter for the features to be retained for SelectKBest, selecting features from high to low according to the calculated chi-squared score until sufficient. Implementing dimensionality reduction using this method removes 53 features, leaving only 20 features to be used. The remaining features after using Univariate Selection are: *'ip_ttl_cv', 'ip_len_cv', 'ip_len_cvq', 'ip_ttl_cvq', 'tcp_ack_rte', 'tcp_seq_cvq', 'tcp_seq_rte', 'tcp_dataofs_median', 'tcp_dataofs_mean', 'tcp_window_median', 'dport_cv', 'tcp_window_mean', 'tcp_flags_mean', 'tcp_flags_median', 'tcp_ack_cvq', 'tcp_seq_mean', 'tcp_seq_median', 'tcp_seq_cv', 'ip_ttl_std', 'ip_len_std'.*

Table 3: *Simulated dataset with dimensionality reduction using Univariate Selection*

Algorithm	Initial Accuracy (%)	Accuracy after Dimensionality Reduction (%)	Initial Real-time Processing (ms)	Real-time Processing after Dimensionality Reduction (ms)
KNN	*85.75*	*85.12*	*5872.18*	*1388.56*
AdaBoost	*85.68*	*85.31*	*399.17*	*153.27*
Random Forest	*85.79*	*85.43*	*3999.79*	*2245.72*
SVM	*85.81*	*85.19*	*992.28*	*289.71*

Based on the observations in Table 3, we can see that the results obtained are quite promising. The accuracy of the model decreases slightly, but the model execution time decreases significantly. Additionally, when using the Feature Importance technique, the result retains only 20 features. Therefore, when managing the IDS system, administrators only need to set rules to capture the exact 20 features for an incoming data stream, reducing the data sampling time and speeding up processing for the entire system.

4. Conclusion

With a dataset containing only two labeled categories, normal and attack, the simulated dataset shows that the proposed models all provide good results. Timely network intrusion detection is guaranteed (rapid classification of attacks and normal traffic) while maintaining a relatively high system accuracy. The system provides relatively accurate classification of normal traffic but exhibits low accuracy when specifically classifying individual attack types, leading to false alerts. The proposed system is suitable for labeled datasets to classify between attack traffic and normal traffic. Two models using KNN and Random Forest combined with dimensionality reduction techniques both yield good results in terms of both accuracy and execution time. The obtained results indicate that the proposed system performs best on all three datasets when combining the KNN algorithm with the Feature Importance dimensionality reduction technique. After calculating and returning the number of important features in attack detection using the Importance technique, the performance of the KNN algorithm is improved. By retaining only the important features, as the dimensionality of the data decreases, the computation speed of KNN increases. Therefore, although the accuracy may decrease slightly, the computation time decreases significantly, which is acceptable.

REFERENCES

[1] S. A. Dheyab, "Efficient Machine Learning Model for DDoS Detection," *Acta Informatica Pragensia*, vol. 11, issue 3, pp. 348-360, 2022. DOI: 10.18267/j.aip.199

[2] S. A. Abbas and M. S. Almhanna, "Distributed Denial of Service Attacks Detection System by Machine Learning Based on Dimensionality Reduction," *Journal of Physics: Conference Series*, 1804(1), 2021. DOI: 10.1088/1742-6596/1804/1/012136

[3] A. A. Abdulrahman and M. K. Ibrahem, "Evaluation of DDoS Attacks Detection in a CICIDS2017 Dataset Based on Classification Algorithms," *Iraqi Journal of Information and Communications Technology*, 1(3), 49-55, 2018. DOI: 10.31987/ijict.1.3.40

[4] Alduailij, "Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method," *Symmetry*, 14(6), 1095, 2022. DOI: 10.3390/sym14061095

[5] Y. Zhen, "A systematic literature review of methods and datasets for anomaly-based network intrusion detection," *Journal of Computers & Security*, vol. 116, issue C, pp. 1-10, 2022. DOI: 10.1016/j.cose.2022.102675

[6] E. Alhajjar, "Adversarial machine learning in Network Intrusion Detection Systems," *Expert Systems with Applications*, vol. 186, pp. 1-10, 2021. DOI: 10.1016/j.eswa.2021.115782

[7] Y. Alharbi and S. Kautish, "Denial-of-Service Attack Detection over IPv6 Network Based on KNN Algorithm," *Wireless Communications and Mobile Computing*, 2021, Article ID 8000869, 2021. DOI: 10.1155/2021/8000869

[8] Arowolo and O. Olugbara, "Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier," *Journal of Big Data*, 8(1), 1-14, 2021. DOI: 10.1186/s40537-021-00415-z

[9] Y. Liao and R. V. Vemuri, "Use of K-Nearest Neighbor classifier for intrusion detection," *Computers & Security*, 21(5):439-448, 2002. DOI: 10.1016/S0167-4048(02)00514-X

[10] M. Aamir and S. M. A. Zaidi, "Clustering based semi-supervised machine learning for DDoS attack classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, issue 4, pp. 436-446, 2021. DOI: 10.1016/j.jksuci.2019.02.003

[11] F. S. D. L. Filho, A. M. B. Junior, G. V. Solar and L. F. Silveira, "Smart Detection: An Online Approach for DoS/DDoS Attack Detection Using Machine Learning," *Security and Communication Networks*, vol. 2019, pp. 1-15, 2019. DOI: 10.1155/2019/1574749

[12] Adnan Helmi Azizan, Salama A. Mostafa, Aida Mustapha , Cik Feresa Mohd Foozy, Mohd Helmy Abd Wahab , Mazin Abed Mohammed and Bashar Ahmad Khalaf, "A Machine Learning Approach for Improving the Performance of Network Intrusion Detection Systems," *Annals of Emerging Technologies in Computing*, 5(5), 201-208, 2021. DOI: 10.33166/AETiC.2021.05.025

[13] R. A. Disha and S. Waheed, "Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique," *Cybersecurity*, 5(1), 2022. DOI: 10.1186/s42400-021-00103-8

[14] L. H. Hiep, "Improve network security system in Vietnam using reverse method," *TNU Journal of Science and Technology*, vol. 225, no. 09, pp. 125-133, 2020.

[15] L. H. Hiep, "Study to applying Blockchain technology for preventing of spam email," *TNU - Journal of Science and Technology*, vol. 208, no. 15, pp. 161-167, 2019.

[16] L. H. Hiep, "Study to analyse, compare and evaluate the performance of Next General Firewalls: case of Palo Alto and Fortigate Firewall," *Vinh University Journal of Science (VUJS)*, vol 51, no. 2A, pp. 24-36, 2022. DOI: 10.56824/vujs.2022nt08

TÓM TẮT

NÂNG CAO HIỆU QUẢ AN NINH MẠNG DỰA TRÊN MỘT SỐ KỸ THUẬT GIẢM CHIỀU DỮ LIỆU

Hoàng Thị Phượng
Trường Đại học Kinh tế - Kỹ thuật Công nghiệp, Hà Nội, Việt Nam
Ngày nhận bài 03/01/2024, ngày nhận đăng 28/01/2024

Bài báo này tập trung nghiên cứu đề xuất mô hình phát hiện xâm nhập mạng áp dụng các kỹ thuật học máy cơ bản với mục đích tăng cường việc phát hiện xâm nhập mạng sớm (mô hình phát hiện hành vi tấn công nhanh chóng) để tăng hiệu suất cho việc ngăn chặn các cuộc tấn công mạng. Hệ thống đó vẫn phải đảm bảo về mặt kỹ thuật là đưa ra được những thông báo có tính chính xác cao. Nghiên cứu sử dụng một số kỹ thuật giảm chiều dữ liệu để phát hiện xâm nhập bất thường mạng do bị tấn công kiểu Distributed Denial of Service (DDoS). Mô hình đề xuất này với mục đích nhằm làm giảm thời gian tính toán giúp phát hiện sớm tấn công. Kết quả cho thấy hệ thống đề xuất đạt kết quả tốt nhất trên cả ba bộ dữ liệu là việc kết hợp giữa giải thuật KNN và kỹ thuật giảm chiều dữ liệu Feature Importance. Sau khi tính toán và trả về số lượng đặc trưng quan trọng trong việc phát hiện tấn công bởi kỹ thuật Importance thì hiệu năng của giải thuật KNN được cải thiện. Vì chỉ giữ lại các đặc trưng quan trọng, số chiều dữ liệu càng giảm thì khả năng tính toán của KNN càng nhanh. Vì vậy, tuy độ chính xác có giảm nhẹ nhưng thời gian tính toán thì giảm đi rất nhiều. Điều này là có thể chấp nhận được.

Từ khóa: Tấn công mạng; DdoS; học máy; học sâu; giảm chiều dữ liệu.