Jurnal Media Teknik Elektro dan Komputer
Vol. 01, No.01 (January - June, 2024)
E-ISSN: XXXX-XXXX

Network Traffic Anomaly Detection Using the Decision Tree Method

[1]Muhammad Iqbal Manalu, [2]Fatur Padilla Hutabarat

[1] Universitas Sumatera Utara; iqbalmanalu12@gmail.com
[2] Universitas Islam Negeri Sumatera Utara; faturhutabarat17@gmail.com

ABSTRACT

With the increasing penetration of computer networks connected to the internet, the risk of network intrusion is also on the rise. Such intrusions attempt to bypass network security mechanisms. One way to detect intrusions is by analyzing network traffic activities. However, manually analyzing all network data can be cumbersome and time-consuming. You can use decision trees to classify network events based on specific attributes. This enables the creation of rules that can detect anomalies in network traffic patterns. We can develop these rules by mapping network events to unique segments within the decision tree. Constructing rules based on the sequence of segments from the decision tree allows for the identification of intrusion indicators, aiding in detecting intrusion attempts in the network. This approach provides a more efficient means for network analysts to identify abnormal network traffic activities without the need for manual inspection of every data point. Thus, the use of decision trees enhances the capability to detect network intrusions, safeguarding computer networks from increasingly complex and diverse cyber threats. This represents a crucial step in fortifying our digital infrastructure's security.

Keywords: Detection, Anomaly, Network Traffic, Decision Tree

Corresponding Author:
Muhammad Iqbal Manalu
Universitas Sumatera Utara; iqbalmanalu12@gmail.com

1. INTRODUCTION

In the increasingly advanced digital era, computer networks have become the main backbone of various business activities, education, government, and daily life (Saputra et al., 2023). As the use of networks increases, a variety of threats have also surfaced, potentially disrupting the performance and security of these networks. One of the most significant threats is the presence of anomalies or abnormalities in network traffic that can indicate cyber attacks, system failures (Najib & Sulistyo, 2020), or configuration problems. Therefore, network traffic anomaly detection is an urgent need to maintain network security and stability. One can interpret security in a computer network system as an effort to shield data and resources from unauthorized access, destruction, and misuse (Hardani & Ramli, 2022),. We must keep a computer network secure, both physically and virtually. One of the main challenges in managing network security is being vulnerable to attacks or acts of system destruction. The frequency of attacks stemming from system vulnerabilities through a network is on the rise.

Buffer overflow is the number one threat and is a security hole from January to October 2007, followed by Denial-of-Service (DOS) (Lindblom, 2023). Cisco typically associates the buffer overflow problem with its Internetwork Operating System (IOS). Routers and switch products use IOS, a multitasking, embedded operating system. There are several techniques commonly used to secure computer network systems, including firewalls, encryption of messages sent, and virtual private networks (VPN). But with the growing understanding of how the system works, the intruder's ability to find system weaknesses to exploit has also grown. Intruders sometimes use patterns that are difficult to track and identify. They often employ multiple stages before breaching the target's security system. Therefore, an additional layer of security is required to identify intruders and their attacks, specifically through anomaly detection in computer network traffic. We are still searching for a practical, effective, and efficient solution to Network Interruption Detection (NID), which is a challenging issue, particularly when handled manually (Lappas & Pelechrinis, 2007).

Decision Tree is a popular algorithm in machine learning due to its ability to perform classification and regression based on a set of easy-to-understand decision rules. The algorithm constructs a decision tree model from training data and uses the resulting rules to categorize new data. The main advantage of this method is its ability to handle large and complex data with relatively quick computation time. This study aims to overcome these problems by developing and evaluating an effective decision tree model for detecting network traffic anomalies and thereby improving the security and performance of computer networks.

2. LITERATURE REVIEW

2.1. Intrusion Detection System

Intrusion is an attempt to bypass a computer system (Sinclair et al., 1999). IDS gathers and tracks operating systems and network data activity, then analyzes this information to shed light on the circumstances surrounding the attack. We classify IDS into two categories based on the analysis of the data.

1. Misuse detection
   The system learns existing and known attack patterns. We learn this pattern by examining all incoming data to identify the type of intrusion. This method is unable to detect new attacks whose patterns are not yet known.
2. Anomaly detection
   Patterns are learned from normal data. Unseen data is checked, and deviations from the learned pattern are sought. This method is unable to identify the type of attack.

2.2. Attacks and Disruptions to Network Traffic

We classify attack simulations based on the attacker's actions and goals. Each type of attack falls into one of four main categories (Siriporn & Benjawan, 2008):

1. Denial of Service (DoS) Attacks
   DoS attacks aim to limit or deny provider services to users, computers, or networks (Maulana et al., 2023). Common tactics are to overload the target system (examples: Apache, Smurf, Neptune, Ping of Death, back, mailbomb, etc.).
2. Probing or surveillance attacks
   Probing and surveillance attacks aim to collect information from a computer system or network system (Imam et al., 2019). Generally, this category includes port scans or the sweeping of IP addresses. (examples: saint, portsweep, mscan, nmap, etc.).
3. User-to-Root (U2R) Attacks

User-to-root attacks aim to gain root or super-user access to a particular computer or system where the attacker previously had user access (Ginting et al., 2018). This is an attempt by a non-privileged user to gain administrative privileges. (examples: perl, xterm, etc.)

4. Remote-to-local (R2L) attacks
   A remote-to-local attack occurs when a user sends packets to a computer via the internet, to which they do not have access, with the intention of exposing computer vulnerabilities and exploiting the privileges that local users have on the computer (Setya Wijaya, 2012). Examples of such attacks include xclock, dictionary, guest_password, phf, sendmail, and xsnoop.

2.3. *Classification*

Classification is the process of finding a model or function that explains or distinguishes concepts or classes of data, with the aim of being able to estimate the class of an object whose label is unknown (Marlina & Bakri, 2021). The model itself can take the form of an "if-then" rule, a decision tree, a mathematical formula, or a neural network. Decision trees are one of the most popular classification methods because they are simple for humans to interpret. Each branch outlines the necessary conditions, while the ends of the tree indicate the data class. The most well-known decision tree algorithm is C4.5, but newer algorithms like Rain Forest (Meera, 2010) can handle large-scale data that main memory cannot hold. Other classification methods are Bayesian, neural network, genetic algorithm, fuzzy, case-based reasoning, and k-nearest neighbor. Typically, two stages divide the classification process: learning and testing (Sinaga et al., 2022). In the learning stage, (Riadi et al., 2019) feed some known data to form an estimated model. Then, in the test stage, they test the formed model with additional data to assess its accuracy. If the accuracy is sufficient, we can use this model to predict unknown data classes.

2.4. *Applying Data Mining to Intrusion Detection Systems*

Network-based IDS can apply data mining techniques to protect military subnetworks. Each military subnetwork is a probe that filters and logs network traffic into a central database. We analyze archived data using a rule set to identify intrusive patterns. Simple patterns, such as observing excessive activity or connections from IP addresses with intrusive habits, are identified. An example of this type of intrusion is a low-level, long-running attack that exhibits intrusive habits for hours, days, or weeks, and originates from various networks. We can apply data mining to this problem to develop human pattern recognition.

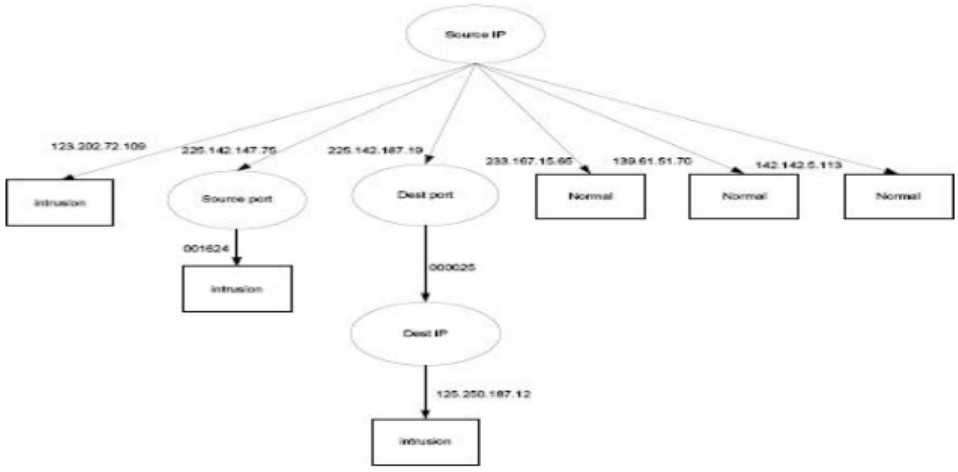

Figure 1. Decision Tree Traffic Network for IDS

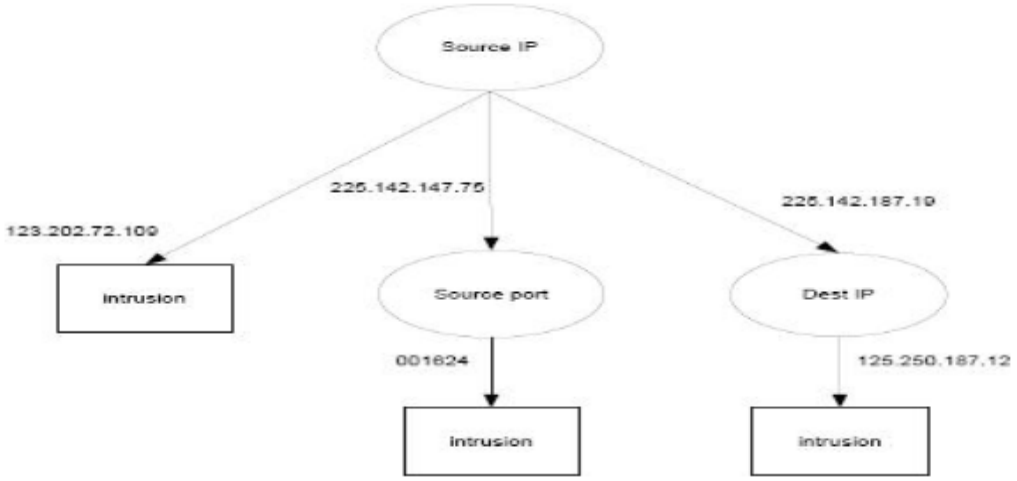

Figure 2. Pruning Decision Tree Traffic Network for IDS

3. METHODS

This study employs an experimental research method, specifically testing the accuracy of the decision tree algorithm in analyzing network traffic. Algorithm testing is carried out using data from the 1999 KDD Cup, which is network traffic data from traffic monitoring from the IDS Snort software, which is processed and classified into several types of intrusions defined in class labels such as normal, back, neptune, imap, pod, satan, smurf, and several other attacks, so that the total type of intrusion is 23 types.

3.1. Application of Decision Tree for Network Traffic Anomaly Detection

The following describes the main procedure of the decision tree algorithm, which is used to solve the problem of detecting traffic anomalies on a network using previously prepared training data.

Input : *dataset* D
Output : *decision tree* T
Procedure :

1. Initialize all weights in *D*, Wi = 1/n, where n is the total number of samples.
2. Calculate the probability *P(Cj)* for each class *Cj*.

$$in\ D.\ P\ (C_j) = \frac{\sum_{c_i} W_i}{\sum_{i=1}^{n} W_i} \quad (1)$$

Description:

a. Identification of Weight *(Wi)*
   Collection of weight (*Wi*), in each row (*i*) in class *(Cj)*
b. Calculating the total weight of class *(Cj)*
   $\sum_{i=1}^{n} W_i$ is used in the process of accumulating all weights (*Wi*) in each row (*i*) in class (*Cj*)
c. Calculating the total weight for class *Cj*
   $\sum_{c_i} W_i$is used in the process of adding up all weights Wi for each row (*i*) in each class
d. Calculating the probability *P*(*Cj*)

Equation (1) is used in calculating the probability P(*Cj*) by dividing the total weight for class *Cj* by the total weight values in all classes.

3. Calculate the conditional probability *P(Aij|Cj)* for each attribute value in *D*.

$$P\ (A_{ij}\ |\ C_j) = \frac{P(A)_{ij}}{\sum_{c_i} W_i} \quad (2)$$

4. Calculate the posterior probability for each example in *D*.

$$P(e_i|C_j) = P\ (C_j) \prod P\ \ A_{ij}|C_j \quad (3)$$

5. Update the sample weights in D with the maximum likelihood (ML) posterior probability P(Cj|ei); Wi= PML(Cj|ei)
6. Find the attribute to perform splitting with the highest information from the gain using the weight update, Wi in D.
7. T = Create a root node and label it with the attribute for splitting.
8. For each branch of T, D represents the data created by applying splitting to D. Then, repeat steps 1 to 7 until each created part or leaf node has the same class.
9. When the decision tree shape is complete, the algorithm ends.

Calculations from the Rapidminer application tools yield the following decision tree results:

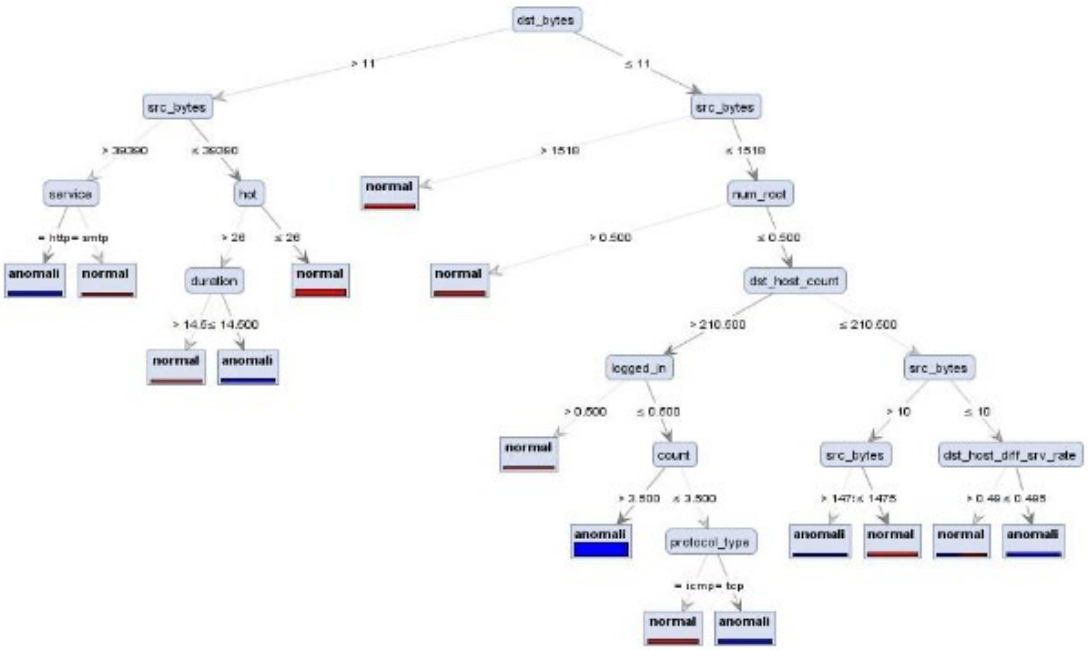

Figure 4. Decision Tree Formed From Calculations Using the Rapidminer Tool

4. RESULTS AND DISCUSSION

The tests conducted reveal that out of 9750 sample data records, 14 have incorrect predictions. This shows the accuracy level of the decision tree algorithm is 99.86%. In an effort to improve network security and reliability, we have implemented a decision tree algorithm to detect network traffic anomalies. Network management relies heavily on anomaly detection, as these anomalies can signal cyber attacks or technical problems that require immediate attention. Figure 5 displays the results.

accuracy: 99.86%

	true anomali	true normal	class precision
pred. anomali	7810	4	99.95%
pred. normal	10	1926	99.48%
class recall	99.87%	99.79%	

Figure 5. Accuracy Test Results

We conducted accuracy testing to assess the model's performance in distinguishing between normal and suspected anomalous traffic. The test results show that the decision tree algorithm is able to achieve a very high level of accuracy in detecting network traffic anomalies, which is 99.86%. This level of accuracy shows that the model has the ability to distinguish between normal and anomalous traffic with a very low error rate.

precision: 99.48% (positive class: normal)

	true anomali	true normal	class precision
pred. anomali	7810	4	99.95%
pred. normal	10	1926	99.48%
class recall	99.87%	99.79%	

Figure 6. Precision Test Results

Implementing a decision tree algorithm to detect network traffic anomalies can improve network reliability and security. The significance of anomaly detection lies in its ability to pinpoint potential network threats or disruptions that require immediate attention. Precision is a key model performance metric.

Precision measures how accurate the model is in identifying anomalous instances from all instances predicted as anomalies. High precision indicates that the model rarely gives false positives, meaning that most anomaly predictions are truly anomalies. The test results show that the decision tree algorithm achieves a precision level of 99.48% in detecting network traffic anomalies. This means that of all instances predicted as anomalies by the model, 99.48% of them are truly anomalies, while only 0.52% are false positive predictions.

The precision level of 99.48% reflects the ability of the decision tree model to accurately identify network traffic anomalies with very few errors. In this context, high precision is critical because errors in detecting anomalies can lead to false alarms that disrupt network operations and waste resources identifying non-existent threats.

recall: 99.79% (positive class: normal)

	true anomali	true normal	class precision
pred. anomali	7810	4	99.95%
pred. normal	10	1926	99.48%
class recall	99.87%	99.79%	

Figure 7. Recall Test Results

Recall evaluates the model's capacity to distinguish all genuinely anomalous instances from the total number of truly anomalous instances. High recall signifies that the model seldom overlooks instances that warrant classification as anomalies. The test results show that the decision tree algorithm achieves a recall rate of 99.79% in detecting network traffic anomalies. This indicates that the model accurately detects 99.79% of truly anomalous instances, with only 0.21% remaining undetected. The recall rate of 99.79% reflects the ability of the decision tree model to detect almost all network traffic anomalies with very few errors. In this context, high recall is essential because errors in detecting anomalies can result in undetected security threats, which can damage the integrity and performance of the network.

The Rapidminer software calculations yielded prediction accuracy rates above 99% for all test results. This shows that the decision tree algorithm is excellent for applying anomaly detection to network traffic.

5. CONCLUSION

The Decision Tree algorithm demonstrates an accuracy level above 99% in detecting network traffic anomalies, indicating its potential as a solution to network traffic anomaly detection. We can extend the application of the decision tree algorithm in this study by developing online or real-time tools or applications that monitor network traffic and send alerts to users, enabling network administrators to identify disruptions.

REFERENCES

Ginting, S. E. B., Widodo, A. W., & Adikara, P. P. (2018). Voting Based Extreme Learning Machine dalam Klasifikasi Computer Network Intrusion Detection. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(6), 2158–2167.

Hardani, M. S., & Ramli, K. (2022). Perancangan Manajemen Risiko Keamanan Sistem Informasi Manajemen Sumber Daya dan Perangkat Pos dan Informatika (SIMS) Menggunakan Metode NIST 800-30. *JURIKOM (Jurnal Riset Komputer)*, *9*(3), 591–599.

Imam, R. M., Sukarno, P., & Nugroho, M. A. (2019). Deteksi Anomali Jaringan Menggunakan Hybrid Algorithm. *E-Proceeding of Engineering*, *6*(2), 8766–8787.

Lappas, T., & Pelechrinis, K. (2007). Data mining techniques for (network) intrusion detection systems. *Department of Computer Science and Engineering UC Riverside, Riverside CA*, *92521*.

Lindblom, H. (2023). *Nuking Duke Nukem: Reaching the Stack via a Glboal Buffer Overflow in DOS Protected Mode*.

Marlina, D., & Bakri, M. (2021). Penerapan Data Mining Untuk Memprediksi Transaksi Nasabah Dengan Algoritma C4. 5. *Jurnal Teknologi Dan Sistem Informasi*, *2*(1), 23–28.

Maulana, A. B., Hartiana, S. N., & Fardan, F. (2023). Analisis Serangan Denial Of Service (DOS) Pada Jaringan Privat Seluler 5G Stand Alone Berbasis Open Seluler. *EProceedings of Engineering*, *9*(6).

Meera, G. (2010). Adaptive Machine Learning Algorithm (AMLA) Using J48 Classifier. *Advances in Computational Sciences and Technology*, *3*, 291–304.

Najib, W., & Sulistyo, S. (2020). Tinjauan Ancaman dan Solusi Keamanan pada Teknologi Internet of Things. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *9*(4), 375–384.

Riadi, I., Umar, R., & Aini, F. D. (2019). Analisis Perbandingan Detection Traffic Anomaly Dengan Metode Naive Bayes Dan Support Vector Machine (Svm). *ILKOM Jurnal Ilmiah*, *11*(1), 17–24.

Saputra, A. M. A., Kharisma, L. P. I., Rizal, A. A., Burhan, M. I., & Purnawati, N. W. (2023). *TEKNOLOGI*

INFORMASI: Peranan TI dalam berbagai bidang. PT. Sonpedia Publishing Indonesia.

Setya Wijaya, E. (2012). *Deteksi Anomali Trafik Jaringan Dengan Menggunakan Metode Decision Tree*. Universitas Dian Nuswantoro.

Sinaga, S., Sembiring, R. W., & Sumarno, S. (2022). Penerapan Algoritma Naive Bayes untuk Klasifikasi Prediksi Penerimaan Siswa Baru. *Journal of Machine Learning and Data Analytics*, *1*(1), 55–64.

Siriporn, O., & Benjawan, S. (2008). Anomaly Detection and Characterization to Classify Traffic Anomalies. Case Study: TOT Public Company Limited Network. *World Academy of Science, Engineering and Technology*, *48*, 407–415.