Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13625
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Performance Analysis of Random Forest Algorithm for Network Anomaly Detection using Feature Selection

Triya Agustina[1)*], Masrizal[2)], Irmayanti[3)]
[1,2,3)]Universitas Labuhanbatu, Indonesia
[1)]triyaagustina182@gmail.com, [2)]masrizal120405@gmail.com, [3)]irmayantiritonga2@gmail.com

Submitted : April 5, 2024 | **Accepted** : April 17, 2024 | **Published** : April 21, 2024

Abstract: As the volume and complexity of computer network traffic continue to increase, network administrators face a growing challenge in monitoring and discovering unusual activity. To keep the network safe and functioning, detecting anomalies is essential. Machine learning-based anomaly detection techniques have become increasingly popular in recent years. This is due to the fact that conventional anomaly detection methods make it difficult to detect unknown and complex attacks. This research aims to conduct a performance analysis of two feature selection methods using the random forest algorithm using the UNSW-NB15 dataset to determine which model is most effective in detecting network traffic anomalies. The models evaluated were random forest with the filter method and random forest with the wrapper method. A number of metrics used for model performance assessment are accuracy, F1-score, receiver operating characteristic curve, and precision-recall. Dataset collection, data pre-processing, feature selection, model construction, and evaluation are the main components of the research methodology. The research results show that the Random Forest approach with the Filter method has an accuracy of 0.8950, F1-score of 0.8333, ROC score of 0.8928, and a precision-recall value of 0.8347. Meanwhile, the approach using the Wrapper method obtained an accuracy of 0.9151, F1-score of 0.8510, ROC score of 0.9136, and a precision-recall value of 0.8637. This shows that the performance of Random Forest with the Wrapper method is superior in all assessment metrics. Random Forest with the Wrapper Method is the right choice of model for detecting network traffic anomalies because of its stable performance and ability to handle complex patterns.

Keywords: Anomaly Detection; Feature Selection; Machine Learning; Random Forest; UNSW-NB15.

INTRODUCTION

Anomaly detection in network traffic is an important part of maintaining network security. Every anomaly can have a big impact on many things, such as national security, personal data storage, social welfare, and economic problems (Jr., Rodrigues, Carvalho, Al-Muhtadi, & Jr., 2019). Anomalies usually occur and are associated with sudden behavior and unknown distribution. Moreover, anomalies in low-dimensional space may show obvious abnormal signs, but in high-dimensional space, they remain hidden and difficult to detect (Pang, Shen, Cao, & Hengel, 2020). For modern data security, addressing network anomalies has become critical to detecting unknown cyberattacks effectively (Roshan & Zafar, 2021).

Anomaly detection is a key approach to identifying potential security threats and operational problems in computer networks. Anomaly detection can be used to discover experience attacks and unusual user behavior (Fariadi & Islami, 2022). Anomaly detection is better than abuse detection because it can detect unknown attacks, so only normal examples are needed to train the model (Nixon, Sedky, & Hassan, 2020). The goal of anomaly detection is to determine all such events in a data-driven manner (Chalapathy & Chawla, 2019).

Traditional anomaly detection techniques, which employ rule-based algorithms, have limitations in discovering diverse and sophisticated anomalies. On the other hand, anomaly detection in network traffic using machine learning algorithms has shown encouraging results. Machine learning-based anomaly detection has become increasingly popular (Tan, Sama, Wijaya, & Aboagye, 2023). This is due to the fact that the advantages of machine learning in detecting anomalies are its ability to recognize complex patterns, easily adapt to changes, be applied to large amounts of data without affecting performance, self-learning ability, and ability to detect

* Corresponding Author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

previously unknown anomalies (Nassif, Talib, Nasir, & Dakalbab, 2021). Machine learning also has the flexibility to detect anomalies according to different network needs and characteristics. Another thing that machine learning has is its advantage in providing a fast response to early detection of network anomalies (Wang et al., 2021).

Several machine learning algorithms have been applied by previous researchers to detect computer network anomalies. The research results show that the accuracy of the Random Forest algorithm in classifying anomalies and detecting attacks on computer networks is better than that of other machine learning algorithms (F. A. Khan & Gumaei, 2019) (Hooshmand & Doreswamy, 2019). Research by (Almomani et al., 2021), has conducted a comparative evaluation of ten machine learning classification algorithms, and the results show that random forest outperforms other classification algorithms in terms of accuracy. The random forest algorithm also has the advantage of carrying out a cross-validation test with the most accurate results (Alsahli, Almasri, Al-Akhras, Al-Issa, & Alawairdhi, 2021).

To detect anomalies accurately, machine learning requires high-quality data (Devia & Soewito, 2023). High-quality data poses problems for learning models. First, it will be more difficult for the learning model to have optimal performance because the problem model must be made more complex as more features are used. Second, due to the large number of feature configurations, even though we have limited data, overfitting can occur with this data. Third, the amount of memory and time required to process high-quality data makes it computationally difficult to process (Ariyoga, 2022). Therefore, it is necessary to select relevant features from the dataset so that the performance and efficiency of machine learning work have a significant impact. The application of feature selection can improve the performance of the algorithm in carrying out classification (Riadi, Utami, & Yaqin, 2023). Feature selection is also useful in terms of simplifying training data (Wardhani & Lhaksmana, 2022). Feature selection in machine learning involves removing less relevant features, thus simplifying the model, reducing overfitting, and increasing computational efficiency (I. A. Khan, Birkhofer, Kunz, Lukas, & Ploshikhin, 2023).

To detect anomalies on a computer network, a dataset is needed, which is a collection of network traffic activity logs. Many previous researchers used the KDD CUP 99 dataset (Tan et al., 2023). Even though the KDD CUP 99 dataset has several weaknesses, namely, data redundancy problems, an unbalanced number of attacks, and a mismatch between the number of attacks and regular traffic (Sahli, 2022). Classifiers trained on the KDDCup99 dataset show a bias towards redundancy (Sapre, Ahmadi, & Islam, 2019). The KDD CUP 99 dataset is an old dataset that contains repeated records, so it is less accurate to use in testing and provide unfair classification results (Kocher & Kumar, 2020). Therefore, in this research, further development was carried out using a different dataset, namely, the UNSW-NB15 dataset. The UNSW-NB15 data set is newer than NSL KDD 99 or KDD 99, CAIDA, Kyoto 2006+, and ISCX. Modern network traffic for both normal and anomalous events, such as current low-footprint attacks, is included. It is more suitable for reliable evaluation of network anomaly detection systems because it is available in a clean format and has no data redundancy (Disha & Waheed, 2022)

Based on the research described previously, it is proven that selecting a feature selection model can improve classification performance in machine learning, but there has been no research that specifically compares the feature selection method applied to the random forest algorithm using the UNSW-NB15 dataset in detecting anomalies in computer networks. Therefore, this research aims to compare the feature selection model of the filter method with the wrapper method applied to the random forest algorithm. Then performance measurements will be carried out to find out the best feature selection method by comparing evaluation matrices such as accuracy, F1-score, receiver operating characteristic (ROC) curve, and precision-recall to find out which method is better.

LITERATURE REVIEW

An important concept in machine learning that has a big influence on model performance is feature selection (Huljanah, Rustam, Utama, & Siswantining, 2019). A pre-processing technique known as “feature selection” helps in choosing the right attributes because real-world data examples are usually large. This technique helps reduce data dimensionality by removing irrelevant and redundant features (Arora & Kaur, 2020). For data analysis, machine learning, and data mining, feature selection is becoming increasingly important, especially for large-dimensional datasets. Irrelevant and redundant features should be filtered out by selecting an appropriate subset of features to avoid overfitting and overcome the curse of dimensionality (Bommert, Sun, Bischl, Rahnenführer, & Lang, 2020). Therefore, feature selection is very important as it facilitates a deeper understanding of the data by learning only relevant features and improves the predictive ability, speed, and accuracy of the classifier.

Three types of feature selection are generally known: the filter method, the wrapper method, and the embedded method (Fei et al., 2022). The filter method works by using a statistical learning approach as a measurement to evaluate attributes independently of the learning or classification algorithm. The filter method directly removes features from the original feature set and selects important features quickly; however, it is difficult to reduce redundancy for some highly correlated features (Fei et al., 2022). For classification datasets with numeric features, all filter methods are applicable; some methods also work for categorical features. Most filter methods calculate a score for each feature and then select the feature with the highest score. However, there are some methods that

* Corresponding Author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13625
e-ISSN : 2541-2019
p-ISSN : 2541-044X

select features iteratively, which means the feature with the highest score is selected at each iteration, but the scores of these iterations are not comparable (Bommert et al., 2020).

In the wrapper method, feature selection is carried out by considering the classification algorithm. The classification algorithm is used repeatedly, each time with a different subset, and the quality of the subset is assessed each time so that the best subset is selected (Arora & Kaur, 2020). The wrapper method takes a subset of the set of each feature. A supervised learning model, such as a classification, is fitted for each subset. The subsets are then evaluated with performance measures calculated based on the resulting models, such as classification accuracy (Bommert et al., 2020). The wrapper method selects features simultaneously with the learning process and generally provides better results than the filter method (Fei et al., 2022). Several studies have confirmed that machine learning algorithms, such as Random Forest and Support Vector Machine, can easily and efficiently achieve feature assessment and selection (Zhu et al., 2019), but the required features must be selected based on some rules.

METHOD

This section presents the methodology used to detect random forest algorithm-based anomalies in network traffic by applying feature selection. The rigorous research methodology makes it easy to create a trustworthy and reliable anomaly detection system. The processes carried out are dataset collection, data pre-processing, feature selection, model construction, and model evaluation (Doreswamy, Hooshmand, & Gad, 2020) (Devia & Soewito, 2023) as shown in Figure 1. Python is a programming language that is used because of its adaptability, simplicity, and broad support for machine learning, data manipulation, and visualization packages. Jupyter Notebook is an open source web tool that allows creating and sharing documents with equations, text, live code, and visuals.

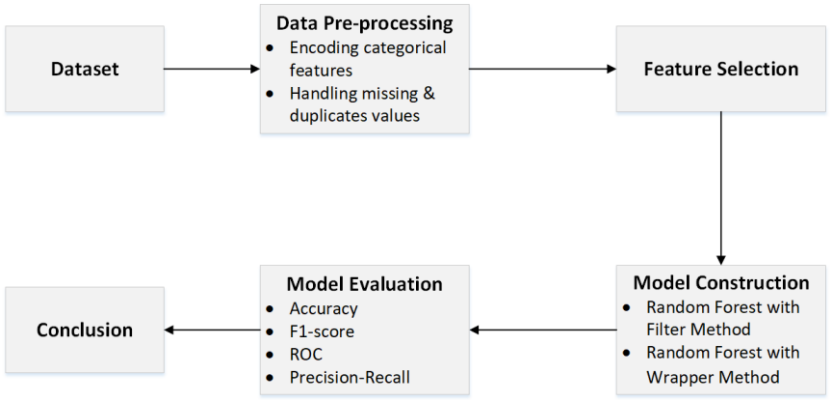


Fig 1. The Research Method Flow

Dataset

This research uses the UNSW-NB15 dataset, which is publicly accessible (UNSW, 2021). The UNSW-NB15 dataset was produced by the Cyber Range Lab of the Australian Center for Cyber Security at the University of New South Wales (UNSW) (Sarhan, Layeghy, Moustafa, & Portmann, 2021). These datasets were selected based on the variety of network traffic data they contain, which includes both regular traffic and various types of attacks. The dataset consists of nine types of modern cyber attacks known as analysis, backdoors, DoS, exploits, fuzzers, generic, reconnaissance, shellcode, and worms. This also includes ordinary packets called normal, which are captured with the Tcpdump tool (Moualla, Khorzom, & Jafar, 2021). Of all the available datasets, the datasets used in this research are only two separate CSV files called "UNSW_NB15_training-set" and "UNSW_NB15_testing-set". Table 1 shows the sample for each class and the percentage.

Table 1. Sample of UNSW-NB15 Dataset

Class type	Training samples	Training samples (%)	Testing samples	Testing samples (%)
Normal	56000	31.94	37000	44.94
Analysis	2000	1.14	677	0.82
Backdoors	1746	1.00	583	0.71
DoS	12264	6.99	4089	4.97
Exploits	33393	19.05	11132	13.52
Fuzzers	18184	10.37	6062	7.36

* Corresponding Author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13625
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Generic	40000	22.81	18871	22.92
Reconnaissance	10491	5.98	3496	4.25
Shellcode	1133	0.65	378	0.46
Worms	130	0.07	44	0.05
Total	**175341**	**100**	**82332**	**100**

Table 1 shows the distribution of the UNSW-NB15 dataset. This dataset has nine types of attacks, namely: analysis, backdoors, DoS, exploits, fuzzers, generic, reconnaissance, shellcode, and worms, as well as normal data. The total amount of data is 257,673 records. This dataset is divided into training and testing data sets, each with a total of 175,341 records for training data and 82,332 records for testing data consisting of attack and normal types. For the training data set, 31.94% of this dataset is normal type data; the remaining 68.06% is attack data. Meanwhile, for the test data set, 44.94% is normal data, and 55.06% is attack data.

Data Pre-processing

An important stage in the anomaly detection process is preprocessing the dataset for the random forest algorithm. The goal of preprocessing is to clean and convert raw data into a format that can be used successfully by the selected algorithm. This process consists of several steps, and each is discussed in more detail below:

- **Handling missing and duplicates values:** to impute missing values in the data set, mode imputation for categorical characteristics and mean imputation for numerical features were used. This strategy, which is used frequently, ensures that the data set is comprehensive and suitable for machine learning algorithms. This method is used to prepare the UNSW-NB15 dataset for preprocessing into a format compatible with the Random Forest algorithm. The data preprocessing stage is critical to machine learning tasks, and careful examination of each stage can help ensure the success of an anomaly detection system.
- **Encoding categorical features:** One-hot coding is used to encode categorical features of a dataset. This method helps random forest algorithms understand how to transform categorical data, which allows them to understand the correlation between categories.

Feature Selection

The feature selection techniques used are correlation analysis and recursive feature reduction, which are used to find the most relevant features that correlate with each other. This procedure aims to make machine learning models using random forest algorithms more efficient and effective by reducing data dimensions.

Model Construction

To assess the effectiveness of the created model, the preprocessed dataset is divided into training, validation, and testing datasets. The training set is used to train the model. The validation set serves to select and tune hyperparameters, and the test set assesses the final performance of the model. The dataset is divided into training datasets (70%), validation data (15%), and testing data (15%). After that, a random forest model was built by applying two feature selection methods, namely, the filter method and the wrapper method.

- **Filter Method:** This method aims to narrow the feature space to the most important features. This can improve the efficiency of the random forest algorithm and reduce the possibility of overfitting. The filter technique used is correlation analysis. The random forest algorithm is then trained and evaluated using the selected characteristics as input. This process uses the chi2 and SelectKBest feature selection methods imported from the Sklearn library in Python. It creates an instance of SelectKBest with the chi2 scoring algorithm to select the 25 best features. Next, the input data is matched and transformed correctly.
- **Wrapper Method:** This method aims to select the best features that can improve the performance of the random forest algorithm. The ideal number of features for the random forest algorithm is selected in this implementation using the recursive feature elimination technique. The Random Forest algorithm is trained and evaluated using selected characteristics as input, after which the RFE wrapper approach is used to select the ideal number of features for the algorithm, which can improve performance and reduce the dimensionality of the dataset. In this process, the recursive feature elimination algorithm is used to select 25 main features from the data set. The data were fitted and transformed using a random forest classifier and recursive feature elimination.

Model Evaluation

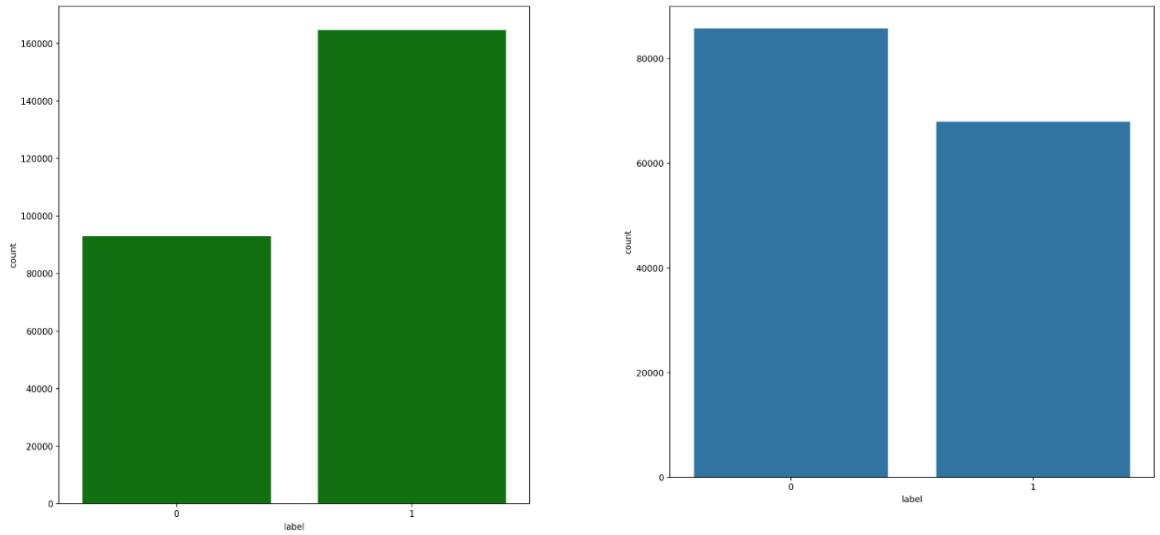
The final phase of this research uses various evaluation metric criteria to measure the effectiveness of the model created. These include accuracy, F1-score, ROC curve, and precision-recall. This metric provides a comprehensive assessment of the model's ability to discover anomalous network traffic. Qualitative analysis is also carried out to identify the advantages and disadvantages of the model so that it can be assessed which model is better.

* Corresponding Author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13625
e-ISSN : 2541-2019
p-ISSN : 2541-044X

RESULT

This section discusses the performance results of the Random Forest algorithm, which is used to detect anomalies based on the UNSW-NB15 dataset using a filter and wrapper feature selection method approach. The main components of the methodology are data collection, pre-processing, feature selection, model construction, and evaluation. During the data collection stage, a large amount of network traffic data, including both legitimate and malicious traffic, must be collected. Fixed data pre-processing issues, including category coding, normalization, class imbalance, and category coding, to allow raw data to be formatted so machine learning algorithms can use it. Feature selection techniques are used to find the most relevant and useful network anomaly detection features. Therefore, overfitting is reduced, and model performance is improved. Various machine learning methods, such as deep and conventional learning, are used to generate these models.



Dataset before Pre-processing Dataset after Pre-processing

Fig 2. UNSW-NB15 Dataset Distribution

Figure 2 shows the distribution of the number of UNSW-NB15 datasets before and after data preprocessing. The number 0 is data that is categorized as normal, while the number 1 is data that is categorized as an attack. Before data preprocessing was carried out, the total amount of data was 257,673 records, which were divided into 93,000 data records in the normal category and 164,673 data records in the attack category. After data preprocessing was carried out, which included the process of removing duplicate data, the amount of data became 153,669 records, which were divided into 85,720 data records in the normal category and 67,949 data records in the attack category. From this process, there was a data reduction of 59.63%, namely 104,004 records.

To reduce the size of the dataset, the correlation between each feature and the target variable is evaluated. Features showing high correlation were excluded from the list. This strategy helps the random forest algorithm perform better and prevents overfitting. The results are shown in Figure 3.

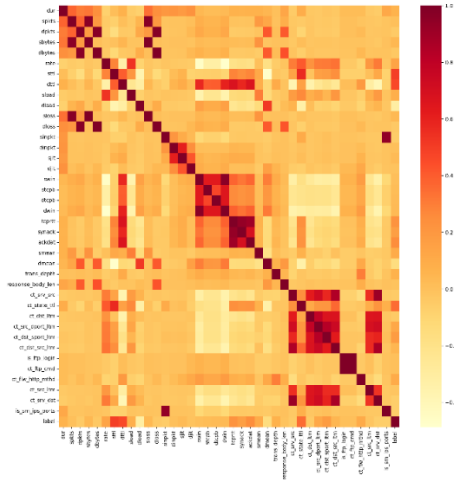


Fig 3. Visualization of Feature Correlation Values

* Corresponding Author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13625
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In Figure 3, you can see the correlation between each feature in the dataset. A lighter color shows that two correlated features have a high correlation value (closer to 1), while a darker color shows that two correlated features have a low correlation value (closer to 0).

Various random forest algorithm models for detecting network traffic anomalies have different levels of success. Random Forest with the Filter method and Random Forest with the Wrapper method are the models that are compared. Comparison of evaluation metrics such as accuracy, F1-score, ROC, and precision-recall can help in determining the most suitable model for network traffic anomaly detection.

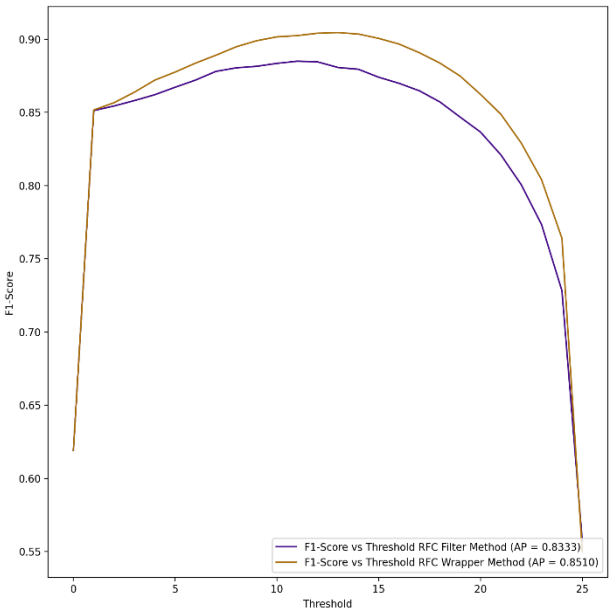


Fig 4. F1-score comparison

The F1-score comparison in Figure 4, which combines precision and recall, provides a reasonable assessment of the model's accuracy. A higher F1-score indicates better overall performance; the Random Forest with Wrapper Method model received the highest F1-score, 0.8510. In this comparison, which shows its possibility to detect anomalies in network traffic better.

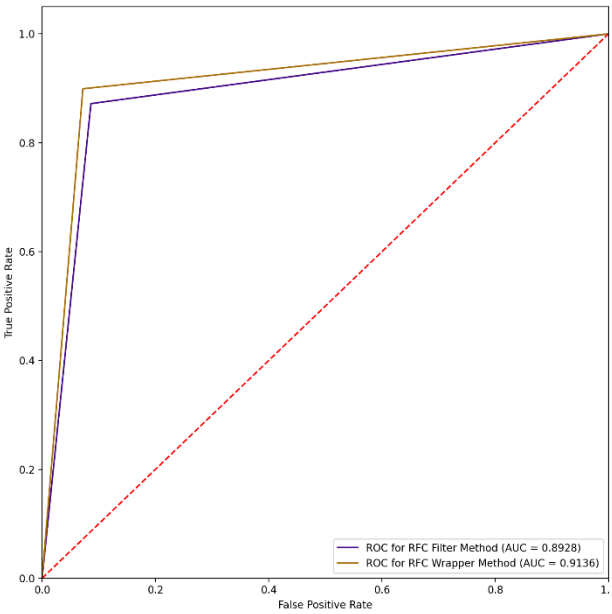


Fig 5. ROC comparison

The Receiver Operating Characteristic (ROC) curve in Figure 5 shows how well the model can differentiate regular from irregular network traffic. The model's performance increases with the ROC value. Random Forest

* Corresponding Author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13625
e-ISSN : 2541-2019
p-ISSN : 2541-044X

model with the wrapper method, with the best ROC score of 0.9136. In this comparison, it shows extraordinary discriminatory abilities.

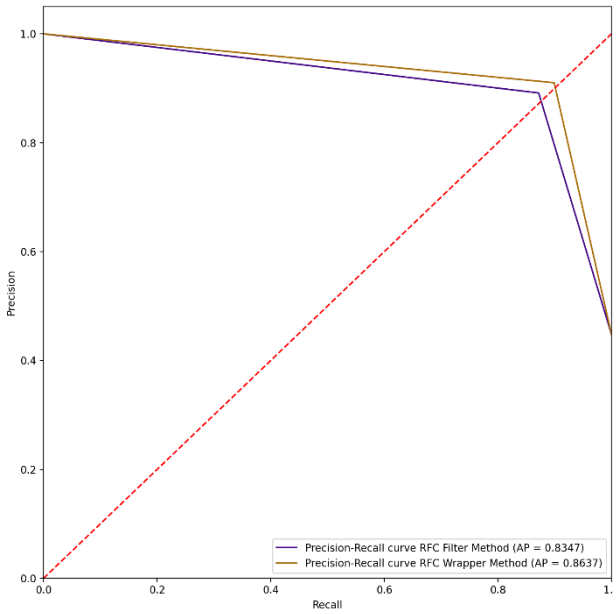


Fig 6. Precision-Recall comparison

The precision-recall graph in Figure 6 shows the difference between sensitivity (recall) and precision (positive predictive value) at different probability thresholds. Having a model with higher recall and precision values will be better. In this comparison, the Random Forest with Wrapper Method model shows a high precision-recall value of 0.8637, indicating the ability to correctly find anomalies while reducing false positives. For comparison, several evaluation metrics were used, such as accuracy, F1-score, ROC (receiver operating characteristic), and precision-recall (PR), the output of which is shown in Table 2.

Table 2. Model Comparison

Model	Accuracy	F1-Score	ROC	PR
Random Forest with Filter Method	0.8950	0.8333	0.8928	0.8347
Random Forest with Wrapper Method	0.9151	0.8510	0.9136	0.8637

Overall, the accuracy of the model is measured. In the comparison in Table 2, the Random Forest with Filter model has an accuracy value of 0.8950, while the Random Forest with Wrapper model has an accuracy value of 0.9151.

DISCUSSIONS

This research has produced a machine learning model with a random forest algorithm through feature selection to detect anomalies in computer network traffic. During the data collection stage, various network traffic data, including normal and malicious traffic, must be collected to make the model flexible. Fixed data pre-processing issues, including missing values, category coding, normalization, and class imbalance, to convert the raw data into a format usable by the random forest algorithm. Feature selection techniques are used to find the most relevant and useful network anomaly detection features. Therefore, overfitting is reduced, and model performance is improved. The filter and wrapper method, which is applied to the random forest algorithm, is used to produce this model. These models are created, improved, and evaluated using various evaluation metrics and cross-validation techniques to ensure that they are robust and applicable in real-world situations. The rigorous research methodology makes it easy to create a trustworthy and reliable anomaly detection system. They also have the ability to identify malicious network data under various conditions. A thorough evaluation method shows the advantages and disadvantages of each model so that the most suitable model for network security needs can be selected. However, there are several limitations to this research, including:

- Dataset limitations: This research uses a freely accessible dataset, namely UNSW-NB15. The quality of the representation and the quality of the dataset determine how well the results can be generalized to other network traffic data sets.

* Corresponding Author

- Algorithmics limitations: This study only applies the random forest machine learning algorithm, so it may not be comprehensive or demonstrate all possible methods for finding network traffic anomalies.
- Time constraints: The time constraints available for this research project may preclude a thorough investigation of multiple subjects or the creation of more complex models.
- Computational resourcesAvailable computer resources can influence the optimization process and selection of machine learning algorithms; this can limit the complexity and effectiveness of the models created.

Despite these limitations, the goal of this research project is to provide in-depth knowledge and useful techniques for using machine learning random forest algorithms to find anomalies in network traffic. The results will be used to guide research and development in this area and help network managers protect their systems from cyber threats.

CONCLUSION

Several random forest algorithm feature selection models for detecting network traffic anomalies have been tested. These models are random forests with the filter method and random forests with the wrapper method. Accuracy, F1-score, ROC, and precision-recall are scoring evaluation metrics used to measure performance. Based on the research results, the Random Forest model with the Wrapper method is superior in performance for all evaluation metrics. A good level of accuracy is shown by obtaining a model accuracy value of 0.9151. The Random Forest model with the Wrapper Method received the highest F1-score of 0.8510, indicating its capacity to find irregularities in network traffic. Its strong discrimination ability is demonstrated by its highest ROC score of 0.9136. Additionally, it shows a precision-recall value of 0.8637, demonstrating the ability to correctly locate anomalies while avoiding false positives. These results show that the Random Forest model with the Wrapper Method is an attractive choice for finding network traffic anomalies. Due to its stable performance and ability to handle complex patterns, it is suitable for detecting anomalous behavior and improving network security. Overall, the model comparisons performed in this study provide a useful picture of how well various random forest algorithm models function to detect anomalies in network traffic. These findings can help in model selection and lay the foundation for further studies and advances in network security and anomaly detection.

REFERENCES

Almomani, O., Almaiah, M. A., Alsaaidah, A., Smadi, S., Mohammad, A. H., & Althunibat, A. (2021). Machine Learning Classifiers for Network Intrusion Detection System: Comparative Study. *2021 International Conference on Information Technology (ICIT)*, 440–445. https://doi.org/10.1109/ICIT52682.2021.9491770

Alsahli, M. S., Almasri, M. M., Al-Akhras, M., Al-Issa, A. I., & Alawairdhi, M. (2021). Evaluation of Machine Learning Algorithms for Intrusion Detection System in WSN. *International Journal of Advanced Computer Science and Applications*, *12*(5), 617–626. https://doi.org/10.14569/IJACSA.2021.0120574

Ariyoga, D. (2022). *Perbandingan Metode Seleksi Fitur Filter, Wrapper, Dan Embedded Pada Klasifikasi Data NIRS Mangga Menggunakan Random Forest Dan Support Vector Machine (SVM)* (Universitas Islam Indonesia). Universitas Islam Indonesia. Retrieved from https://dspace.uii.ac.id/handle/123456789/38955

Arora, N., & Kaur, P. D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing*, *86*, 105936. https://doi.org/https://doi.org/10.1016/j.asoc.2019.105936

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, *143*, 106839. https://doi.org/https://doi.org/10.1016/j.csda.2019.106839

Chalapathy, R., & Chawla, S. (2019). *Deep Learning for Anomaly Detection: A Survey*. 1–50. Retrieved from http://arxiv.org/abs/1901.03407

Devia, A., & Soewito, B. (2023). Analisis Perbandingan Metode Seleksi Fitur untuk Mendeteksi Anomali pada Dataset CIC-IDS-2018. *JTeksis: Jurnal Teknologi Dan Sistem Informasi Bisnis*, *5*(4), 572. https://doi.org/10.47233/jteksis.v5i4.1069

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

Doreswamy, Hooshmand, M. K., & Gad, I. (2020). Feature selection approach using ensemble learning for network anomaly detection. *CAAI Transactions on Intelligence Technology*, *5*(4), 283–293. https://doi.org/10.1049/trit.2020.0073

Fariadi, & Islami, M. R. R. (2022). Deteksi Dini Serangan Pada Website Menggunakan Metode Anomali Based. *JIKO (Jurnal Informatika Dan Komputer)*, *5*(3), 224–229. https://doi.org/10.33387/jiko

Fei, H., Fan, Z., Wang, C., Zhang, N., Wang, T., Chen, R., & Bai, T. (2022). Cotton Classification Method at the County Scale Based on Multi-Features and Random Forest Feature Selection Algorithm and Classifier. *Remote Sensing*, *14*(4). https://doi.org/10.3390/rs14040829

* Corresponding Author

Hooshmand, M. K., & Doreswamy. (2019). Machine Learning Based Network Anomaly Detection. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(4), 542–548. https://doi.org/10.35940/ijrte.d7271.118419

Huljanah, M., Rustam, Z., Utama, S., & Siswantining, T. (2019). Feature Selection using Random Forest Classifier for Predicting Prostate Cancer. *IOP Conference Series: Materials Science and Engineering*, *546*(5). https://doi.org/10.1088/1757-899X/546/5/052031

Jr., G. F., Rodrigues, J. J. P. C., Carvalho, L. F., Al-Muhtadi, J. F., & Jr., M. L. P. (2019). A comprehensive survey on network anomaly detection. *Telecommunication Systems*, *70*(3), 447–489. https://doi.org/10.1007/s11235-018-0475-8

Khan, F. A., & Gumaei, A. (2019). A Comparative Study of Machine Learning Classifiers for Network Intrusion Detection. In X. Sun, Z. Pan, & E. Bertino (Eds.), *ICAIS 2019: Artificial Intelligence and Security* (pp. 75–86). Cham: Springer International Publishing.

Khan, I. A., Birkhofer, H., Kunz, D., Lukas, D., & Ploshikhin, V. (2023). A Random Forest Classifier for Anomaly Detection in Laser-Powder Bed Fusion Using Optical Monitoring. *Materials*, Vol. 16. https://doi.org/10.3390/ma16196470

Kocher, G., & Kumar, G. (2020). *Performance Analysis of Machine Learning Classifiers for Intrusion Detection using UNSW-NB15 Dataset*. 31–40. https://doi.org/10.5121/csit.2020.102004

Moualla, S., Khorzom, K., & Jafar, A. (2021). Improving the Performance of Machine Learning-Based Network Intrusion Detection Systems on the UNSW-NB15 Dataset. *Computational Intelligence and Neuroscience*, *2021*, 5557577. https://doi.org/10.1155/2021/5557577

Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). Machine Learning for Anomaly Detection: A Systematic Review. *IEEE Access*, *9*, 78658–78700. https://doi.org/10.1109/ACCESS.2021.3083060

Nixon, C., Sedky, M., & Hassan, M. (2020). Autoencoders: A Low Cost Anomaly Detection Method for Computer Network Data Streams. *ACM International Conference Proceeding Series*, 58–62. https://doi.org/10.1145/3416921.3416937

Pang, G., Shen, C., Cao, L., & Hengel, A. Van Den. (2020). Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, *1*(1), 1–36. https://doi.org/10.1145/3439950

Riadi, S., Utami, E., & Yaqin, A. (2023). Comparison of NB and SVM in Sentiment Analysis of Cyberbullying using Feature Selection. *Sinkron*, *8*(4), 2414–2424. https://doi.org/10.33395/sinkron.v8i4.12629

Roshan, K., & Zafar, A. (2021). Utilizing XAI Technique to Improve Autoencoder Based Model for Computer Network Anomaly Detection with Shapley Additive Explanation(SHAP). *International Journal of Computer Networks & Communications (IJCNC)*, *13*(6), 109–128. https://doi.org/10.5121/ijcnc.2021.13607

Sahli, Y. (2022). A comparison of the NSL-KDD dataset and its predecessor the KDD Cup '99 dataset. *International Journal of Scientific Research and Management (IJSRM)*, *10*(04), 832–839. https://doi.org/10.18535/ijsrm/v10i4.ec05

Sapre, S., Ahmadi, P., & Islam, K. (2019). A Robust Comparison of the KDDCup99 and NSL-KDD IoT Network Intrusion Detection Datasets Through Various Machine Learning Algorithms. *Journal of Student-Scientists' Research*, *1*. https://doi.org/10.13021/jssr2019.2681

Sarhan, M., Layeghy, S., Moustafa, N., & Portmann, M. (2021). NetFlow Datasets for Machine Learning-Based Network Intrusion Detection Systems BT - Big Data Technologies and Applications. In Z. Deze, H. Huang, R. Hou, S. Rho, & N. Chilamkurti (Eds.), *International Conference on Big Data Technologies and Applications* (pp. 117–135). Cham: Springer International Publishing.

Tan, T., Sama, H., Wijaya, G., & Aboagye, O. E. (2023). Studi Perbandingan Deteksi Intrusi Jaringan Menggunakan Machine Learning: (Metode SVM dan ANN). *Jurnal Teknologi Dan Informasi (JATI)*, *13*(2). https://doi.org/10.34010/jati.v13i2

UNSW. (2021). The UNSW-NB15 Dataset. Retrieved March 20, 2024, from IXIA PerfectStorm website: https://research.unsw.edu.au/projects/unsw-nb15-dataset

Wang, S., Balarezo, J. F., Kandeepan, S., Al-Hourani, A., Chavez, K. G., & Rubinstein, B. (2021). Machine learning in network anomaly detection: A survey. *IEEE Access*, *9*, 152379–152396. https://doi.org/10.1109/ACCESS.2021.3126834

Wardhani, F. H., & Lhaksmana, K. M. (2022). Predicting Employee Attrition Using Logistic Regression With Feature Selection. *Sinkron*, *7*(4), 2214–2222. https://doi.org/10.33395/sinkron.v7i4.11783

Zhu, J., Pan, Z., Wang, H., Huang, P., Sun, J., Qin, F., & Liu, Z. (2019). An Improved Multi-temporal and Multi-feature Tea Plantation Identification Method Using Sentinel-2 Imagery. *Sensors*, *19*(9). https://doi.org/10.3390/s19092087

* Corresponding Author