

Enhancing K-Nearest Neighbor Algorithm on Network Traffic Classification using Recursive Feature Elimination

Andika Elok Amalia^{1*}, Alwin Bahari², Luhur Bayuaji³

¹Computer Science Department, School of Computer Science,
Bina Nusantara University, Jakarta, Indonesia 11480

²Fakultas Teknologi Informasi,
Universitas Budi Luhur, Jakarta, Indonesia 12260

³Faculty of Data Science and Information Technology,
INTI International University, Nilai - Negeri Sembilan, Malaysia 71800

andika.elok@binus.ac.id; alwinbahari.165@gmail.com;
luhur.bayuaji@newinti.edu.my

*Correspondence: andika.elok@binus.ac.id

Abstract – Network traffic classification is a category in network traffic monitoring analysis that has a very important function in network management analysis. By identifying the type of traffic flowing through a network, administrators can allocate bandwidth more effectively, maintain service quality, and detect abnormal activity, which makes it a fundamental part of network monitoring. One of the ways to do it is by machine learning both unsupervised and supervised. This study implemented supervised machine learning, namely K-Nearest Neighbor to network traffic dataset with two different schemas, with and without feature selection. These two schemas are compared directly so that any difference in results can be attributed to the feature selection process. It aims to clarify the impact of Recursive Feature Elimination to network traffic classification using K-Nearest Neighbor Algorithm. It classifies data with features to 17 target classes. Previous studies state that the use of feature selection in some techniques can improve its accuracy. Recursive Feature Elimination is chosen and result that the accuracy is increased from 0.66 to 0.76. Another thing that this study shows is that there is no pattern where the more data points from a class will result in better classification performance in that class. In other words, the number of samples available for a class does not guarantee that the class will be classified more accurately. This research clarifies that feature selection can improve classification

performance. Selecting the most relevant features also reduces the amount of irrelevant information the algorithm must process, which makes the classification more efficient and reliable.

Keywords: classification; network; traffic; k-nearest; feature selection

I. INTRODUCTION

Network Traffic Monitoring Analysis (NTMA) is still an interesting topic to explore because it has a crucial role in network management. Based on Nokia Global network traffic reports through its Bell Labs Consulting arm, it is already projected three scenarios of increased network traffic (moderate, aggressive and disruptive). All scenarios show a significant increase for global wide-area network (WAN) traffic in overall traffic, even in 2033 it is predicted global WAN traffic reaches 3280–6641 Exabytes (EB)/month with Compound Annual Growth Rate of 18%-27% (Nokia, 2024). This fact further strengthens the importance of network management. There are several categories of NTMA applications such as traffic prediction, traffic classification, fault management and network security. In this research we will focus on network traffic classification that identifying services producing traffic, because it is a fundamental step in network management and monitoring.

Previous studies already implemented some approaches performing it. This includes using machine learning methods both supervised and unsupervised to discover features from traffic and then train model with historical data to group into services based on feature value or unsupervised that cluster traffic without training data (Alessandro D'Alconzo et al., 2020). Researchers have widely discussed the application of supervised machine learning algorithms in classifying network traffic. Many algorithms deployed are Support Vector Machines (SVM), Naive Bayes (NB), C4.5, C5.0, Random Forest, K-Nearest Neighbor (KNN), Decision Tree (DT), Gradient Boost (GB) and AdaBoost. Comparing the accuracy each other between algorithms in different cases, searching lowest time requirement, determining effectiveness of the supervised machine learning algorithms to classify network traffic at an early stage and detecting untrained versions are aims in their researches (Ahmad Azab et al., 2024).

C4.5 gave the highest performance compared to Artificial Neural Network (ANN) and SVM in classifying traffic to several different application types such as DNS, FTP, TELNET, P2P, WWW, IM, and MAIL. Online internet traffic data from these seven application types was captured, then feature extraction was performed using the NetMate tool. The accuracy of C4.5 reached 97.57%. It also outperforms SVM, BayesNet and NB in terms of accuracy when classifying traffic into five specific applications (Muhammad Shafiq and Xiangzhan Yu, 2018; Mohsin Mahmood et al., 2023).

Another comparison of algorithms for network traffic classification, which grouped network patterns into two classes: "heavy traffic" and "light traffic," showed that DT had a slightly higher accuracy than KNN, at 94.71%, while KNN had a 94.28% accuracy. The lowest of the three was SVM, with an accuracy of 91.52%. In other hand, traffic classification in IoT network that implemented machine learning to classify TON_IOT dataset and a real-world traffic dataset by the ISCX network point that KNN algorithm achieves highest accuracy (Zahraa Abbas Hasan Tayyeh et al., 2024; Arqam Awais and Zeshan Iqbal, 2022; Farzana Alam et al., 2021). KNN also demonstrates a very high accuracy compared to

another machine learning techniques, reaching 99.5% in classifying network security attack (Muskan Agarwal et al., 2024) and outperforms NB in Intrusion Detection System (Aditya Dwi Afifaturahman and Firmansyah Maulana, 2021).

Research is developing not only on model evaluation but also on how to improve its performance, one of which is by applying feature selection. It is used to select the most relevant features from a dataset while remove those that are irrelevant. In theory, there are 3 categories of approaches to feature selection techniques, namely filter, wrapper and embedded methods. The approach that often yields higher accuracy is wrapper category, that also has some techniques, those are forward selection, backward elimination, and recursive feature elimination (RFE) (Cheng, 2024). RFE applied before the classification process to group nano materials using random forest showed that the best performance was achieved by the reduced RF model based on RFE where a balanced accuracy of 0.82 was obtained (Bahl et al., 2019). Another research also implemented RFE and supervised machine learning to build early prediction models of chronic heart disease. Many ML techniques such as KNN, SVM, DT, NB and LR, result that RFE is effective at selecting those features in a training dataset in predicting chronic heart disease, but it is recommended to test the RFE on different datasets to validate the findings of the study (Napa et al., 2024).

Based on background above, KNN will be implemented in this study to classify network traffic because in some cases before it shows high accuracy, besides we also use RFE to see how it impacts its performance. The expectation is that RFE will improve the accuracy of KNN.

II. METHODS

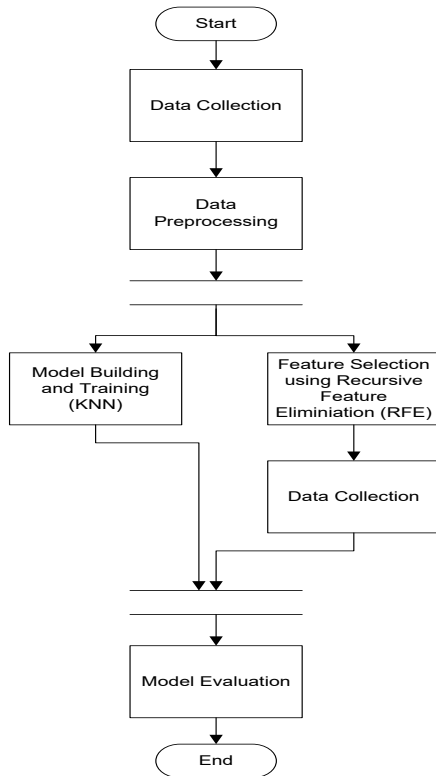


Figure 1 Research Flowchart

Figure 1 show the research steps that can be described as follow.

Flow.ID	Source.IP	Source.Po	Destination	Destination	Protocol	Timestamp	Flow.Durs	Total.Fwd	Total.Back	Total.Leng	Total.Leng	Fwd.Packi	Fwd.Packi	Fwd.Packi	Fwd.Packi	Bwd.Packi	Bwd.Packi	Bwd.Packi
172.19.1.4	172.19.1.4	52422	10.200.7.7	3128	6/26/04/201	45523	22	55	132	110414	6	6	6	6	0	4380	1187	2007.527
172.19.1.4	10.200.7.7	3128	172.19.1.4	52422	6/26/04/201	1	2	0	12	0	6	6	6	6	0	0	0	0
10.200.7.2	50.31.185.	80	10.200.7.2	38848	6/26/04/201	1	3	0	674	0	337	0	124.6667	194.567	0	0	0	0
10.200.7.2	50.31.185.	80	10.200.7.2	38848	6/26/04/201	217	1	3	0	0	0	0	0	0	0	0	0	0
192.168.7.192	168.7.	55961	10.200.7.7	3128	6/26/04/201	78068	5	0	1076	0	529	6	215.2	286.4589	0	0	0	0
172.19.1.5	10.200.7.6	3128	172.19.1.5	50004	6/26/04/201	105069	138	0	313554	0	5840	6	2305.544	1220.822	0	0	0	0
192.168.7.192	168.7.	55963	10.200.7.7	3128	6/26/04/201	104443	5	0	1076	0	529	6	215.2	286.4589	0	0	0	0
192.168.11	192.168.11	51848	10.200.7.6	3128	6/26/04/201	11002	3	12	232	3664	226	0	77.33333	128.7841	3628	0	305.3333	
10.200.7.2	68.67.178.	443	10.200.7.2	57300	6/26/04/201	108503	10	6	6904	1302	1448	0	890.4	660.8942	1302	0	217	
192.168.7.192	168.7.	55977	10.200.7.7	3128	6/26/04/201	118415	7	0	2210	0	1096	0	315.7143	533.0434	0	0	0	0
192.168.11	10.200.7.4	3128	192.168.11	57740	6/26/04/201	205118	32	4	6494	3118	1460	6	202.9375	464.0353	1460	51	779.5	
192.168.11	10.200.7.4	3128	192.168.11	57740	6/26/04/201	3	3	0	9991	0	2920	6	1998.2	1267.837	0	0	0	0
192.168.11	10.200.7.4	3128	192.168.11	57740	6/26/04/201	131	3	0	5611	0	4145	6	1870.333	2099.788	0	0	0	0
192.168.11	10.200.7.4	3128	192.168.11	57740	6/26/04/201	8	3	0	5611	0	4180	6	1870.333	2257.279	0	0	0	0
172.19.1.4	10.200.7.6	3128	172.19.1.4	50227	6/26/04/201	108338	123	0	192822	0	5840	0	1567.659	958.8652	0	0	0	0
212.124.1	212.124.1	443	10.200.7.1	44447	6/26/04/201	202096	4	6	174	3130	137	0	63.5	61.76258	1460	0	521.6667	
192.168.11	10.200.7.4	3128	192.168.11	57741	6/26/04/201	202151	31	4	1861	3118	992	6	60.03226	178.4919	1460	51	779.5	
192.168.11	10.200.7.4	3128	192.168.11	57741	6/26/04/201	1	2	0	998	0	992	6	499	697.2073	0	0	0	0
192.168.11	10.200.7.4	3128	192.168.11	57741	6/26/04/201	1	2	0	998	0	992	6	499	697.2073	0	0	0	0

Figure 2 Sample Dataset

ProtocolName is chosen as target class out of 87 attributes with total 47 classes and leave the rest as features.

2.2 Data Preprocessing

In this step, we balanced the class distribution of the imbalanced dataset to optimize the classification results in the model.

2.1 Data Collection

The data used in this study is network traffic flows (Rojas, 2018), which can be downloaded from www.kaggle.com. This dataset is the result of data collection through a project carried out by a team from the network division of the University of Del Cauca, Colombia. They conducted packet captures at different hours, during morning and afternoon, over six days (April 26, 27, 28 and May 9, 11 and 15) of 2017. The dataset consists of 87 attributes and 3.577.296 records (network flows), as shown in Figure 2.

The technique used in this study to balance the class distribution is undersampling. The constraints of this research problem, only classes with more than or equal to 1000 data points will be selected. It filters only 17 target classes from 47 that can be seen in Tabel 1.

Table 1. List of Target Classes

No	Class	Amount of Data Points
1.	GOOGLE	80,888
2.	HTTP	75,256
3.	HTTP_PROXY	46,856
4.	SSL	30,004
5.	HTTP_CONNECT	22,078
6.	YOUTUBE	13,678
7.	MICROSOFT	4,534
8.	AMAZON	4,487
9.	WINDOWS_UPDATE	2,989
10.	GMAIL	2,930
11.	SKYPE	2,528
12.	DROPBOX	2,463
13.	YAHOO	2,076
14.	FACEBOOK	1,736
15.	CLOUDFLARE	1,564
16.	TWITTER	1,432
17.	MSN	1,135

We also remove data points that contain empty or null values in the data points and delete columns that contain zero values “0” and remove unimportant columns so that the data attribute column side has 63 attributes left. we next convert the target class values from categorical to numerical (integer) data type from 0-16.

2.3 Feature Selection

This method will remove less relevant features and re-evaluating the model's performance at each step. The process starts with all features and iteratively eliminates those with the smallest contribution to the model, as determined by metrics like feature coefficients or importance scores (Cheng, 2024). It performs ranking then iterates until number of features fit with the specified parameter that we set to 20 features.

2.4 Model Building and Training

We will build and train the KNN model with two different dataset (without and with feature selection). KNN is one of the simplest and most basic machine learning techniques. It works by grouping all samples that have similar characteristics from a dataset. It starts by selecting k , the number of nearest neighbors to choose then calculate the distance between the new specimen and all points in training

dataset. We split the dataset into training data and test data with a ratio of 80:20. The calculation of data similarity uses Euclidean Distance formula below.

$$d(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^2 \right)^{1/2}$$

It then sorts all points in the training set from nearest to farthest from the new point, and selects the closest k points (Kien Trang and An Hoang Nguyen, 2021; Halder, et al.; 2024).

2.5 Model Evaluation

The evaluation is carried out using a confusion matrix with the main parameter that will be referred to being accuracy to see how accurate the model is in classifying correctly.

III. RESULTS AND DISCUSSION

The results of class classification on the protocol name attribute of the original data (63 features) using the KNN algorithm are presented in Table 2. Based on the calculation formula for the evaluation metrics, the highest precision, recall, and f1-score values are in the Microsoft class with values of 0.962162, 0.851675, and 0.903553, respectively. The lowest value is in the YouTube class with values of 0.4, 0.379147, and 0.389294, respectively. Meanwhile, the overall accuracy is 0.66.

Table 2. Evaluation Result using Original Data

Class Name	Precision	Recall	F1-score
GOOGLE	0.78866	0.809524	0.798956
HTTP	0.878049	0.923077	0.9
HTTP_PROXY	0.726027	0.823834	0.771845
SSL	0.705426	0.846512	0.769556
HTTP_CONNECT	0.409091	0.507692	0.453089
YOUTUBE	0.4	0.379147	0.389294
MICROSOFT	0.962162	0.851675	0.903553
AMAZON	0.47549	0.530055	0.501292
WINDOWS_UPDATE	0.538071	0.519608	0.528678
GMAIL	0.906977	0.753623	0.823219
SKYPE	0.671717	0.777778	0.720867
DROPBOX	0.52	0.522613	0.521303
YAHOO	0.724138	0.588785	0.649485

FACEBOOK	0.492147	0.451923	0.471178
CLOUDFLARE	0.869159	0.920792	0.894231
TWITTER	0.81768	0.721951	0.766839
MSN	0.487952	0.405	0.442623
accuracy	0.665294	0.665294	0.665294
macro avg	0.668985	0.666682	0.665059
weighted avg	0.670093	0.665294	0.664887

Table 3. Evaluation Result using Data after RFE with 20 features

Class Name	Precision	Recall	F1-score
GOOGLE	0.941176	0.955224	0.948148
HTTP	0.939815	0.948598	0.944186
HTTP_PROXY	0.678414	0.781726	0.726415
SSL	0.841584	0.85	0.845771
HTTP_CONNECT	0.52193	0.574879	0.547126
YOUTUBE	0.56	0.538462	0.54902
MICROSOFT	1	0.967033	0.98324
AMAZON	0.551111	0.645833	0.594724
WINDOWS_UPDATE	0.67284	0.539604	0.598901
GMAIL	0.948454	0.906404	0.926952
SKYPE	0.83871	0.842593	0.840647
DROPBOX	0.654378	0.666667	0.660465
YAHOO	0.894118	0.839779	0.866097
FACEBOOK	0.681564	0.61	0.643799
CLOUDFLARE	0.908163	0.951872	0.929504
TWITTER	0.822581	0.809524	0.816
MSN	0.572139	0.552885	0.562347
accuracy	0.760882	0.760882	0.760882
macro avg	0.766293	0.763593	0.763726
weighted avg	0.76354	0.760882	0.761006

Table 3 shows the results of the class classification on the protocol name attribute of the RFE data (20 features). The highest precision, recall, and f1-score values is also in Microsoft classes with values of 1, 0.967033, and 0.98324, respectively. The lowest values are in the MSN class with values of 0.56, 0.538462, and 0.54902, respectively. Meanwhile, the overall accuracy is 0.76.

Based on the results, it can be seen that the class with the lowest evaluation score in both datasets, both the original and the RFE-implemented dataset, is MSN, as it has the fewest data points. However, the class with the highest performance is MICROSOFT, even though it does not have the highest number of data points. There are still six classes with more data points than MICROSOFT, all of which have more than 10,000 data points. When viewed across all classes, there is no

pattern that the more data available, the better the performance.

Accuracy increased after RFE was implemented on the processed data, increasing from 0.66 to 0.76. This indicates that RFE's effect on the dataset we used can maximize the performance of the KNN classification process.

IV. CONCLUSION

This study can clarify the influence of the implementation of feature selection, namely RFE, on the network traffic classification process using KNN. Previous study stated that RFE shows good influence on ML classification models, but it is recommended to test in another cases and datasets to strengthen the statement. In line with that, the comparison of accuracy values in this research ensure that the implementation of RFE can improve the performance of KNN.

In other hand, this research still needs further development to confirm the relation between amount of data points with class performance in classification process, because no pattern was found, the more data there is in a class, the more accurate the model is in classifying. Future research can enhance KNN with feature selection technique other than RFE or implement ensemble learning to KNN to see how far it affect the accuracy.

REFERENCES

- A. Awais and Z. Iqbal. (2022). Network Traffic Classification through Machine Learning methods in IoT Networks. International Conference on IT and Industrial Technologies (ICIT), Chiniot, Pakistan, 2022, pp. 01-06, doi: 10.1109/ICIT56493.2022.9989079
- Afifaturahman, A. D., Maulana, F., & Artikel, S. (2021). Perbandingan Algoritma K-Nearest Neighbour (KNN) dan Naive Bayes pada Intrusion Detection System (IDS) INFORMASI ARTIKEL ABSTRACT. INNOVATION IN RESEARCH OF INFORMATICS, 3(1), 17–25. <http://innovatics.unsil.ac.id>

- Azab, A., Khasawneh, M., Alrabaaee, S., Choo, K. K. R., & Sarsour, M. (2024). Network traffic classification: Techniques, datasets, and challenges. In *Digital Communications and Networks* (Vol. 10, Issue 3, pp. 676–692). KeAi Communications Co. <https://doi.org/10.1016/j.dcan.2022.09.009>
- Bahl, A., Hellack, B., Balas, M., Dinischiotu, A., Wiemann, M., Brinkmann, J., Luch, A., Renard, B. Y., & Haase, A. (2019). Recursive feature elimination in random forest classification supports nanomaterial grouping. *NanoImpact*, 15. <https://doi.org/10.1016/j.impact.2019.100179>
- Cheng, X. (2024). Signals and Systems Ins. In *Comput. Signal Syst* (Vol. 1, Issue 1). <https://soapubs.com/index.php/ICSS>
- D'Alconzo, A., Drago, I., Morichetta, A., Mellia, M., & Casas, P. (2020). A Survey on Big Data for Network Traffic Monitoring and Analysis. <https://doi.org/10.1109/TNSM.2019.2933358>
- F. Alam, R. Kashef and M. Jaseemuddin. (2021). Enhancing The Performance of Network Traffic Classification Methods Using Efficient Feature Selection Models. *IEEE International Systems Conference (SysCon)*, Vancouver, BC, Canada, 2021, pp. 1-6, doi: 10.1109/SysCon48628.2021.9447076.
- Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, 11(1). <https://doi.org/10.1186/s40537-024-00973-y>
- M. Agarwal, K. S. Gill, R. Chauhan, A. Kapruwan and D. Banerjee. (2024). Classification of Network Security Attack using KNN (K-Nearest Neighbour) and Comparison of different Attacks through different Machine Learning Techniques. 3rd International Conference for Innovation in Technology (INOCON), Bangalore, India, pp. 1-7, doi: 10.1109/INOCON60754.2024.10512250.
- Mahmood, M., Ahmad Khalil, S., Jamil Shah, S., & Ahmad, M. (2023). Network Traffic Classification Techniques and Comparative Evaluation of Machine Learning Models. In *International Journal of Innovative Science and Research Technology* (Vol. 8, Issue 6). www.ijisrt.com
- Napa, K. K., Kumar, A. K., Murugan, S., Mahammad, K., Assegie, T. A., & Okokpujie, K. (2024). Early prediction of chronic heart disease with recursive feature elimination and supervised learning techniques. *IAES International Journal of Artificial Intelligence*, 13(1), 730–736. <https://doi.org/10.11591/ijai.v13.i1.pp730-736>
- Shafiq, Muhammad & Yu, Xiangzhan & Wang, Dawei. (2018). Network Traffic Classification Using Machine Learning Algorithms. 621-627. 10.1007/978-3-319-69096-4_87
- Trang, K., & Nguyen, A. H. (2022). A Comparative Study of Machine Learning-based Approach for Network Traffic Classification. *Knowledge Engineering and Data Science*, 4(2), 128. <https://doi.org/10.17977/um018v4i22021p128-137>
- Z. A. Hasan Tayyeh, R. Alubady, M. Molhem and M. Alsamman. (2024). Network Traffic Classification Using Machine Learning. 7th International Conference on Internet Applications, Protocols, and Services (NETAPPS), Kuala Lumpur, Malaysia, 2024, pp. 1-7, doi: 10.1109/NETAPPS63333.2024.1082342