Instance Selection with Naïve Bayes to Improve DDoS Attack Classification Accuracy Using Random Forest

Aditya Putra Ramdani[1,*] Achmad Solichan[2] Muhammad Zainudin Al Amin[3] Nova Christina Sari[4] Basirudin Ansor[5] Mulil Khaira[6]
[1,2,3,4,5,6]Universitas Muhammadiyah Semarang, Semarang, Central Java 50273, Indonesia
adityaputraramdani@unimus.ac.id

Abstract. DDoS Attack is one of the threats in a series of network systems. Attacks on a network in one unit of time can subsequently occur in a very large number of attacks. Previous research has been done to avoid DDoS attack through classification process and one of which is based on Random Forest method. The large number of attacks requires classification. In previous research, Random Forest was one way to classify DDoS attacks. The classification used is using the Random Forest algorithm. The Random Forest classification model produces an accuracy of 98.02%. This research is a preprocessing step involving Naïve Bayes instance selection which is compared with Adaboost instance selection which is expected to remove noise data due to the relatively large amount of data. With large quantities, it is hoped that this preprocessing step can get maximum results. The research also involved the Naïve Bayes and ZeroR classification methods, where the best results were using Naïve Bayes instance selection with the random forest classification method with an accuracy of 100%.

Keywords: Classification, DDoS, Random Forest, Naïve Bayes.

1. Introduction

Currently technology have entered the 4.0 era where digitalization has penetrated all aspects of human life. As technology develops, today's technology has many risks that can cause the failure of these technological facilities. This can be influenced by several factors. Such as economic, social, political, and so on [1][2].

DDoS attacks are one of the threats that exist in a computer network system today. A DDoS attack is an attack that targets the performance of a server within a computer network with the aim of damaging the performance of the server. In an attack there are several things that are the object of the attack. Attacks can be carried out from the hardware resource side to the worst case, paralyzing the system so that the system cannot function as it should. DDoS attacks that occur in one unit of time can number hundreds or even thousands of attacks. For this reason, it is necessary to have a classification process for attack data which is used to determine anticipatory steps for further attacks in the future [3].

I. Yustar Afif and R. Nindyo Sumarno (eds.), *Proceedings of the 2nd Lawang Sewu International Symposium on Engineering and Applied Sciences (LEWIS-EAS 2023)*, Advances in Engineering Research 234,
https://doi.org/10.2991/978-94-6463-480-8_19

DDoS attacks are a vicious threat that can shut down a network system. Delays in detecting DDoS attacks can cause system paralysis in the network or service process and there is still attack identification that has low accuracy. In research put forward by several researchers, one of whom is Alkasassbeh, et al produced an accuracy level of 98.02% in identifying DDoS attacks using the Random Forest method (M. Alkassbeh, 2016). Previous research conducted by Alkasassbeh, et al mentioned that DDoS attacks can be detected using Data Mining. The researcher performed several methods to classify the data including: MLP, Random Forest, and Naive Bayes. From some of these methods, it was found that the MLP classification method has the highest level of accuracy compared to others of 98.63% [4][5][6]. In previous research, the dataset used was created by the researcher himself based on personal experience. Which in the end the dataset becomes a public dataset. From the dataset it was found that MLP classification is the best result with the highest accuracy value.

Data mining is broadly classified as descriptive, concentrating more on data depiction, grouping data into categories as well as summarizing data. Predictive data mining analyzes past data and produces conclusions or so-called trends for future predictions. Predictive data mining is rooted in the statistical development process of classic models. Random forest is one of the Ensemble Supervised Machine Learning techniques that has appeared recently. Random forest produces many decision trees, including random sampling of data for bootstrap samples as in bagging and random selection of input features to produce individual base decision trees. Random Forest has a very good potential to become a popular technique for classifying in the future because of its performance which is said to be comparable to the bagged and improved ensemble technique [7].

The activity of classifying a data cannot be separated from the activity of data preprocessing. The preprocessing is very influential on the classification process as well as an initial step in data mining. The part of data mining that goes into preprocessing is Instance Selection which is one of the relatively effective ways to process large data. Instance selection here has the use of being able to reduce the data without leaving the essence of the existing data.

Research by N. G. Pedrajas and A. D. Haro-Garcia in that research used instance selection in preprocessing in order to eliminate useless data that could later cause noise in the data [8]. By proposing instances of selection boosting in the research. In this research, the classification used is the Random Forest classification to process DDoS attack data that had previously been done using the Instance Selection feature using Naïve Bayes [9][10]. Instance Selection is a computational activity where there is input data that is an accumulation of the entire data while the output data is data that has been processed from the input data that consists of several parts until the purpose of using Data Mining is achieved and functions as if all the data is used.[11]

The dataset used in this research uses the same dataset as previous research. The total dataset of DDoS attacks entering the network is 2,000,000 data. The attack data includes a very large amount in a unit of time. Due to the large number of available datasets, it is necessary to perform additional classification to eliminate the classification results that are "false". Based on the description above, it can be concluded that the accuracy of DDoS attack classification can still be improved by using the Random Forest method combined with Instance Selection Naive Bayes, so that better results can be obtained to achieve a higher level of accuracy than previous research.

Based on the display from the background and the identification of specific problems, this research used the instance selection method and the Naive Bayes Algorithm to produce a high level of accuracy by using the Random Forest method selected based on references from several of these studies.

2. Materials and Methods

The development of technology at this time makes the popularity of Cloud Computing systems increase rapidly. However, the widespread use of Cloud Computing raises a security problem. one of the main threats to the security of a cloud computing is a DDoS attack. The importance of providing security mechanisms to prevent DDoS attacks can maintain the performance of a cloud computing system [10]. Based on the research that has been done, information about the development of DDoS attacks and methods to deal with them are studied and analyzed, their advantages and disadvantages. This analysis is done with the aim of finding a new method to develop the existing research, so as to obtain improved performance in detecting DDoS attacks

Data collection at this stage used is public data, data collection is done by taking a dataset that has been selected, which is data from a dataset created by previous researchers with the dataset option "Classification DDoS Attacks" with a total of 2,160,668 datasets with a distribution of label classes original normal as many as 1,935,959, UDP-Flood as many as 201,334 Smurfs as many as 12,590, SIDDOS as many as 6,665 and HTTP-Flood as many as 4,110 involving features as many as 27 features by loading all the information about each network packet. In this research, an experiment was used. The experiments carried out in this research are by using the Weka 3.8.6 tool and Microsoft Excel 2016. The software in this research uses a device with the Windows 10 operating system. The hardware that helps this research is an Intel® Core™ i3 3.33GHz Processor, RAM 8.00 GB with 64-bit Operating System.

2.1. Model Proposal

The Random Forest classification model has been proposed in this research, by previously doing preprocessing in the form of instance selection in order to get the best accuracy by discarding less important data that can cause a lack of classification results. This

preprocessing step itself aims to get the best accuracy by trying to improve the accuracy of the existing dataset which will then be classified using the Random Forest classification method which is then compared with other classifications.

2.2. Preprocessing Step Using Instance Selection

Before doing the classification step in this research, a preprocessing step is done in the form of instance selection. The stage of preprocessing this time is to involve Instance Selection using Naïve Bayes on the DDoS attack dataset to determine the level of accuracy in DDoS attacks in order to obtain the maximum accuracy value. This preprocessing step consists of the following stages:

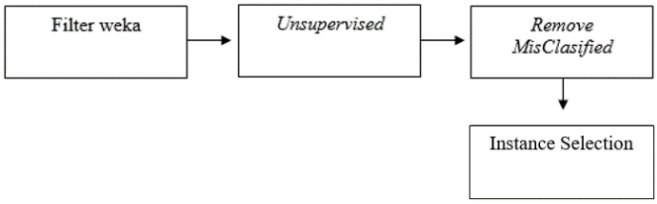

Fig 1. Preprocessing steps

The stage in this preprocessing is to select a filter and then Remove MisClassified by entering a Naïve Bayes Instance which is then compared with the instance selection of adaboost. This research is done by testing the data that has been obtained with the proposed method and evaluation. This research was conducted based on several previous research sources, which later combining the problem with previous research sources resulted in solving the problem, which is using the Random Forest classification method which will then also be compared with Naïve Bayes.

The proposed method on the DDoS attack dataset is the random bayes method for the initial classification process as training data processing. It is useful to determine the weight on the variation of the training data.

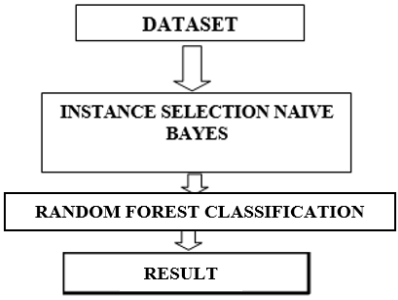

Fig 2. Proposed Method

2.3. Preprocessing Step Using Instance Selection

In the process of finding data accuracy with the Random Forest method in detecting DDoS attacks, this stage is used to test the performance of the proposed model. Testing is done by using the WEKA application as a tool in preprocessing and classification of DDoS attack data. This is done after getting the prediction results from the method. To find the accuracy, the author uses the Confusion Matrix, with the accuracy formula as follows:

$$AC = \frac{TN+TP}{TN+FP+FN+TP} \quad (1)$$

With variables obtained from the Confusion Matrix table, with the following table:

Table 1. Confusion Matrix Table

Actual \ Predicted	Negative	Positive
Negative	TN	FP
Positive	FN	TP

Description:

TP : True Positive

TN : True Negative

FP : False Positive

FN : False Negative

After going through the accuracy search process, the best accuracy will be obtained with the Random Forest method. This research proposes a new model that will be used as a method to improve the performance of DDoS attack classification. The proposed model is to use feature selection as preprocessing. Decision tree C4.5 is chosen as a feature selection method to obtain features that have strong information about the class label.

3. Results

In this research, a classification process was carried out that previously performed preprocessing by performing an Instance Selection filter with Naïve Bayes which will be compared with Instance Selection Adaboost. Where Instance Selection here has the purpose of eliminating data noise in order to obtain the maximum value. The classification algorithm used is Random Forest which will also be compared with naïve bayes classification.

The DDoS attack dataset used in this research is log data consisting of 2,086,528 server attack data with 27 attributes that will be subjected to a classification process that was

previously preprocessed to eliminate data noise until the case data processed in the classification process and ready for modeling and produce maximum value. The large amount of data that has been explained above is worried about the presence of noise due to the presence of inappropriate data, so the process of data preprocessing instance selection needs to be done to obtain attributes and data that are in accordance with the necessary provisions.

3.1. Original Dataset DDos Attack

With a total of 2,160,668 datasets with a normal original label class distribution of 1,935,959, UDP-Flood of 201,334, Smurf of 12,590, SIDDOS of 6,665 and HTTP-Flood of 4,110 involving 27 features by loading all information about each packet network. The purpose of the experiment with the collection of this dataset is to obtain better accuracy by applying several processes in performing the detection of weapons in the determination of the decision of anticipating the next attack.

3.2. Data Preprocessing

In this research, the data used is data that has a relatively large amount. The existence of a large amount of data is worried that there is noise so that it can cause a lack of results or result in less-than-optimal results. Classification results on objects will experience a decrease in accuracy or produce results that are less than optimal. In this preprocessing, the missing value is fixed by performing a cleaning process by providing the Instance Selection function. Instance Selection itself has stages, one of which is by providing a filter on unsupervised detection and then RemoveMisclassified to remove inappropriate data that creates noise on the existing dataset or called the Instance Selection stage.

3.3. Classification and Evaluation

This research uses three classification algorithms to test the selected dataset. The three algorithms are Naive Bayes, Random Forest, and ZeroR. Where each algorithm is given four sub-datasets and then compared with the original dataset. Each algorithm test is given the same, using the cross-validation method, with a fold value of 10.

3.4. Discussion

Naïve Bayes Instance Selection testing results using ZeroR classification. In testing classification on DDoS attacks using datasets from previous research conducted with the ZeroR classification method with several experiments that have the purpose of obtaining the maximum accuracy value using weka tools has the following results:

Table 2. Experimental table using the ZeroR algorithm with Naïve Bayes Instance Selection

Correctly Classified Instances	1895187
Incorrectly Classified Instances	191341
Kappa Statistic	0
Accuracy (%)	**90,82%**

Adaboost Instance Selection testing results using ZeroR Classification. The results of the second experiment conducted by applying the Adaboost Instance Selection method and ZeroR classification with the aim of comparing with the previous method namely Naïve Bayes Instance Selection with ZeroR classification which has an accuracy value of 90.82%. This experiment was carried out using the same tools, namely weka, with the following experimental results:

Table 3. Experimental table using the ZeroR algorithm with Instance Selection Adaboost

Correctly Classified Instances	1895187
Incorrectly Classified Instances	191341
Kappa Statistic	0
Accuracy (%)	**91,43%**

Naïve Bayes Instance Selection Testing Results by using Random Forest Classification. The results of testing with the dataset in this experiment were done by applying the Random Forest classification method with the aim of being the same as the previous classification, which is to compare with the previous method. The results of the experiment are as follows:

Table 4. Experimental table using the Random Forest algorithm with Instance Selection Naïve Bayes

Kappa Statistic	1
Accuracy (%)	**100%**

The results obtained from testing ZeroR, Naïve Bayes, and Random Forest by previously providing a preprocessing step in the form of Instance Selection, then the test results can be concluded that the DDoS Attack dataset using Naïve Bayes Instance Selection gets the highest value or accuracy, which is from the Random Forest classification method with 100% accuracy.

4. Conclusion

After conducting research using a dataset in the form of a DDoS attack using a preprocessing step in the form of a comparison of two Instance Selection and several classification methods, the results obtained from the research and testing using ZeroR, Naïve Bayes, and Random Forest classification, then the results of the testing can be concluded that the dataset DDoS attacks that use Instance Selection Naïve Bayes get the highest value or accuracy from the Random Forest classification method with an accuracy of 100%. This research has shown that the performance of modern DDoS attack detection can be improved by using feature selection. Instance selection can be used to remove unnecessary sub-datasets to facilitate the increase in accuracy.

This research also proves that the instance selection feature can increase the speed of DDoS attack detection on all algorithms tested. The highest speed increase occurs in the random forest algorithm that provides the best performance with perfect values compared to other algorithms.

Authors' Contributions

Author 1: Conceptualization, Methodology, & Investigation. **Author 2**: Supervision & Validation. **Author 3**: Writing – review & editing. **Author 4**: Formal analysis. **Author 5**: Writing – original draft. **Author 6**: Visualization.

Conflicts of Interest

The authors declare they have no conflicts of interest.

Acknowledgments

This work was supported by LPPM Universitas Muhammadiyah Semarang.

References

1. P. Bosch-Sijtsema, C. Claeson-Jonsson, M. Johansson, and M. Roupe, "The hype factor of digital technologies in AEC," *Construction Innovation*, vol. 21, no. 4, 2021, doi: 10.1108/CI-01-2020-0002.
2. G. Strupczewski, "Defining cyber risk," *Saf Sci*, vol. 135, 2021, doi: 10.1016/j.ssci.2020.105143.

3. R. Khader and D. Eleyan, “Sustainable Engineering and Innovation Survey of DoS/DDoS attacks in IoT,” vol. 3, no. 1, pp. 23–28, 2021, doi: 10.37868/sei.v3i1.124.

4. K. Garg and R. Chawla, “Detection of DDoS attacks using data mining,” *International Journal of Computing and Business Research*, 2011, Accessed: Feb. 19, 2024. [Online]. Available: https://www.researchgate.net/publication/263696968

5. M. Alkasassbeh, G. Al-Naymat, A. B.A, and M. Almseidin, “Detecting Distributed Denial of Service Attacks Using Data Mining Techniques,” *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 1, 2016, doi: 10.14569/ijacsa.2016.070159.

6. Y. Wei, J. Jang-Jaccard, F. Sabrina, A. Singh, W. Xu, and S. Camtepe, “AE-MLP: A Hybrid Deep Learning Approach for DDoS Detection and Classification,” *IEEE Access*, vol. 9, pp. 146810–146821, 2021, doi: 10.1109/ACCESS.2021.3123791.

7. L. F. Viera Valencia and D. Garcia Giraldo, “DDoS Attack Detection Based on Random Forest,” *Angewandte Chemie International Edition, 6(11), 951–952.*, vol. 2, pp. 4–10, 2019.

8. N. García-Pedrajas and A. De Haro-García, “Boosting instance selection algorithms,” *Knowl Based Syst*, vol. 67, pp. 342–360, Sep. 2014, doi: 10.1016/J.KNOSYS.2014.04.021.

9. Y. Shang, “Prevention and detection of DDOS attack in virtual cloud computing environment using Naive Bayes algorithm of machine learning,” *Measurement: Sensors*, vol. 31, p. 100991, Feb. 2024, doi: 10.1016/J.MEASEN.2023.100991.

10. R. Saxena and S. Dey, “DDoS attack prevention using collaborative approach for cloud computing,” *Cluster Comput*, vol. 23, no. 2, pp. 1329–1344, 2020, doi: 10.1007/s10586-019-02994-2.

11. M. Blachnik and M. Kordos, “Comparison of instance selection and construction methods with various classifiers,” *Applied Sciences (Switzerland)*, vol. 10, no. 11, 2020, doi: 10.3390/app10113933.