

## Two-steps feature selection for detection variant distributed denial of services attack in cloud environment

Kurniabudi<sup>1</sup>, Eko Arip Winanto<sup>1,2</sup>, Sharipuddin<sup>3</sup>

<sup>1</sup>Department of Computer Engineering, Faculty of Computers Science, Universitas Dinamika Bangsa, Jambi, Indonesia

<sup>2</sup>Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia

<sup>3</sup>Department of Informatic, Faculty of Computers Science, Universitas Dinamika Bangsa, Jambi, Indonesia

### Article Info

#### Article history:

Received Aug 13, 2024

Revised Jun 17, 2025

Accepted Jul 10, 2025

#### Keywords:

Attack detection

Classification

DDoS

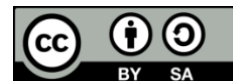
Feature selection

Machine learning

### ABSTRACT

The prevalence of cloud computing among organizations poses a significant problem in ensuring security. Specifically, distributed denial of services (DDoS) attacks targeting cloud computing networks can lead to financial losses for consumers of cloud computing services. This assault has the potential to render cloud services inaccessible. The detection system serves as a remedy to prevent more substantial losses. This research aims to enhance the efficacy of the system detection model by integrating feature selection with three machine learning algorithms: decision tree (DT), random forest (RF), and naïve Bayes (NB). Therefore, our study suggests combining two phases of feature selection into the DDoS attack detection procedure. The first phase uses the information gain (IG) feature selection technique approach, and the second phase uses the principal component analysis (PCA) feature extraction approach. The technique is referred to as two-step feature selection. The test findings indicate that the implementation of two-step feature selection can enhance the performance of the DT and RF detection models by around 9%.

This is an open access article under the [CC BY-SA](#) license.



### Corresponding Author:

Kurniabudi

Department of Computer Engineering, Faculty of Computers Science, Universitas Dinamika Bangsa

Jendral Sudirman Street, Thehok-Jambi, Indonesia

Email: kbudiz@yahoo.com

## 1. INTRODUCTION

Cloud computing has revolutionized information technology and changed the business model for providing IT services. This technology allows users to access various IT resources, such as servers, storage, and applications, through a well-managed and scalable network. As adoption becomes more widespread, many organizations leverage cloud infrastructure for their data management [1]. However, behind the various advantages provided by cloud computing, it brings significant security challenges. Research by Sharma and Singh [2] shows that distributed denial of services (DDoS) attacks are still a major threat in cloud environments. These attacks aim to make cloud services inaccessible to legitimate users by flooding servers and networks with fake traffic and disrupting the regular operation of websites, applications, application programming interfaces (APIs), and other services [3], [4]. The impact is significant on service availability in cloud environments. Therefore, more effective detection methods are needed to counter these attacks.

The suggested methods have not been tested against many DDoS attack variants, despite the fact that a lot of research has been done to detect DDoS attacks in cloud computing. The traffic moving through the cloud environment is neither uniform nor particularly varied, much like the internet. To accurately identify different forms of DDoS attacks, employing a method that can efficiently filter and extract relevant

information (features) is necessary. In order to reduce the many types of DDoS attacks in a cloud computing environment, an intrusion detection system (IDS) that is both dependable and efficient is necessary. This research aims to recognize different types of DDoS assaults inside a cloud computing environment by creating a robust model. Previous research, including Omer *et al.* [5] and He *et al.* [6], have employed diverse algorithms such as logistic regression, support vector machine, decision tree (DT), naïve Bayes (NB), random forest (RF), KMeans, and Gaussian expectation-maximization to identify DDoS attacks, encompassing flooding, spoofing, and brute-force attacks. The test findings indicated a precision level of 99.7% with a negligible false positive rate (FPR) of under 0.07%. A different research [7] uses least square support vector machine (LS-SVM) to identify transmission control protocol (TCP) flood assaults with a precision rate of 97%. Study of Chen *et al.* [8] utilized extreme gradient boosting (XGBoost) to detect internet control message protocol (ICMP) flooding, TCP flooding, TCP-synchronize (SYN) flooding, user datagram protocol (UDP) flooding, and Smurf attacks, achieving a precision rate of 98.5%. In the study of Wani *et al.* [9], researchers presented a combination of the hidden Markov model and RF to detect DDoS attacks. This approach achieved an accuracy of 97.34% and a precision value of 95.45%. A study of Kushwah and Ali [10] introduced a voting extreme learning machine (V-ELM) to detect DDoS attacks. The effectiveness of this method was evaluated using two datasets, namely NSL-KDD and ISCX, resulting in accuracies of 99.18% and 92.11%, respectively.

Furthermore, machine learning techniques are extensively employed for the detection of DDoS attacks [11]–[13]. Nevertheless, certain studies also utilize feature selection to enhance detection. An illustrative study of Bagyalakshmi and Samundeeswari [14] using feature selection techniques using learning vector quantization (LVQ) and principal component analysis (PCA), which are then used for Bayesian classifier, support vector classifier, and tree-based classifier. The test findings indicate that the LVQ and DT algorithms achieved an accuracy of 98.74%, while the PCA and DT algorithms achieved an accuracy of 98.60%. Researchers have integrated feature selection with deep learning in their work. For instance, a study of SaiSindhuTheja and Shyam [15] has suggested combining oppositional crow search algorithm (OCSA) feature selection with a recurrent neural network (RNN) to detect DDoS attacks. The proposed method was tested using the CICIDS2017 dataset and achieved an accuracy of 94.12%. The study by Agarwal *et al.* [16] presents the use of the whale optimization algorithm (WOA) for feature selection in combination with a deep neural network (DNN). The algorithm was evaluated using the KDD-CUP97 dataset and achieved an accuracy of 95.35%.

This research investigates the impact of using a hybrid feature selection method. While previous studies have explored the role of feature selection in enhancing the performance of DDoS detection systems, they have not explicitly examined its effects on DDoS detection systems within cloud computing environments through hybrid feature selection methods. This study proposes a detection model formed by combining feature selection, feature extraction, and machine learning techniques. Feature selection techniques are used to reduce irrelevant features. This study utilizes information gain (IG) combined with PCA to produce optimal features that can identify variations of DDoS attacks. This method is called two-step feature selection. The machine learning methods applied in this study include NB, C4.5, and RF. This work focuses on the improvement mechanism by proposing a two-step feature selection to counter DDoS attacks on cloud computing networks. In addition, this study provides several contributions: i) address the critical need for a robust detection system model that can effectively identify different types of DDoS assaults on cloud computing networks; ii) utilizing a two-stage feature selection approach, this paper will introduce an optimized detection method for mitigating DDoS attacks in specific cloud computing networks; iii) evaluating and assessing the influence of feature selection and extraction on the efficacy of the DDoS attack detection model utilizing a classification methodology; iv) develop a hybrid feature selection methodology utilizing a two-step process that incorporates feature selection through IG and feature extraction via PCA to identify the most pertinent features for identifying diverse versions of DDoS attacks; v) comprehensive evaluation using the CICIoT2023 dataset and multiple classifiers (DT, RF, and NB); and vi) improved detection performance, achieving up to 99% accuracy with DT and RF classifiers.

The remainder of this work is structured as follows. Section 2 describes the experimental setting, detailing each phase of the research, the methodology employed, the data acquisition, and the evaluation metric. Section 3 describes the experimental outcomes and research findings. Ultimately, section 4 summarizes the principal findings of this article, as well as prospective directions and opportunities for future research.

## 2. METHOD

This study employs a two-stage feature selection methodology to identify DDoS assaults. The methods used for detection are DT, RFs, and NB. This section outlines the sequential procedures

required to carry out this study successfully. The components encompassed in this are the dataset, experimental setting, feature selection approach, two-stage feature selection, classification algorithm, and experimental instruments.

## 2.1. Experiment setup

This project aims to tackle security concerns in cloud computing, particularly DDoS attacks. DDoS attacks exhibit several variations; however, few studies have addressed this topic. This paper proposes a two-step technique to enhance the accuracy of the IDS in identifying diverse forms of DDoS attacks. This study thoroughly analyzes several procedures, including feature selection from DDoS attack datasets, feature extraction to derive pertinent features, dataset allocation for training and testing, and the development of an IDS utilizing RF, DT, and NB techniques. The experimental setup, the central component of this research, is separated into four parts, each of which will be further discussed in the following sections.

- Filtering DDoS attack from CICIoT23 dataset, where several DDoS attacks and regular traffic exist.
  - Next, the DDoS dataset is subjected to feature selection for the detection process using IG, PCA, and two-step feature selection (hybrid IG-PCA).
  - Third, comparison and analysis of testing accuracy using RF, DT, and NB methods for each feature selection method.
  - Finally, validation result of model's data split, 5-cross validation and 10-cross validation.
- The experimental stages in this study are shown in Figure 1.

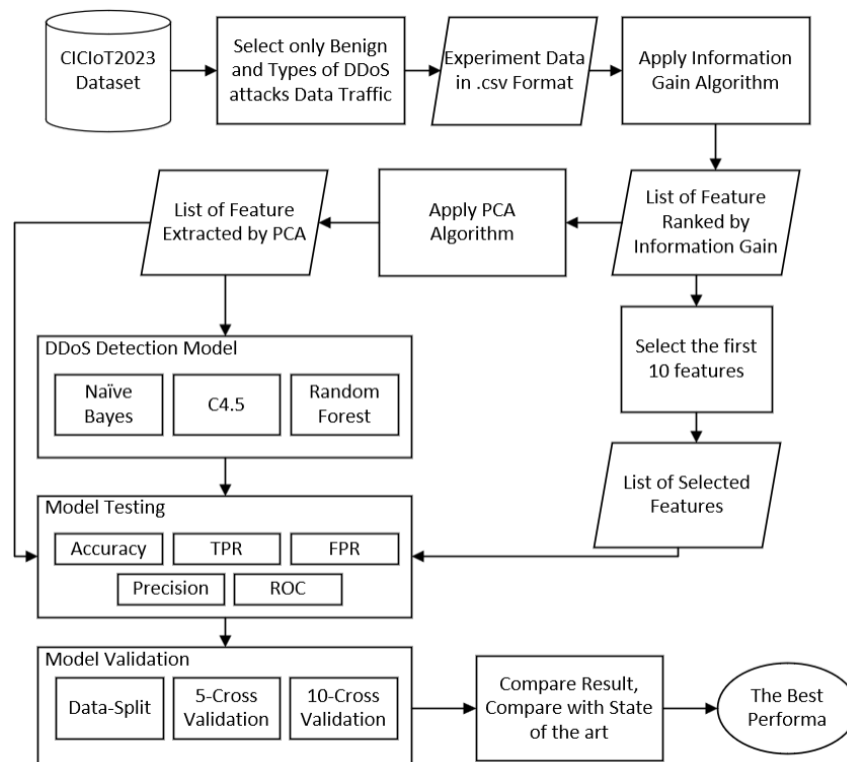


Figure 1. Research experiment

## 2.2. DDoS dataset

This study utilized the CICIoT2023 dataset developed by the University of New Brunswick, Canada [17]. This collection contains traffic associated with security data from internet of things devices and cloud computing. The data in this dataset includes various variables from TCP/IP consisting of 47 features. In addition, this dataset can also cover several attack scenarios, but this study only focuses on DDoS attacks. This study did not utilize all the available datasets owing to resource constraints. Table 1 presents the quantity and classification of DDoS attacks employed in this study.

Table 1. Number of DDoS attacks

Types of attack	Amounts
DDoS-ICMP_Flood	74579
DDoS-UDP_Flood	55800
DDoS-TCP_Flood	46377
DDoS-PSHACK_Flood	42288
DDoS-SYN_Flood	42136
DDoS-RSTFINFlood	41586
DDoS-SynonymousIP_Flood	37354
BenignTraffic	11423
DDoS-ICMP_Fragmentation	4589
DDoS-ACK_Fragmentation	2992
DDoS-UDP_Fragmentation	2956
DDoS-HTTP_Flood	331
DDoS-SlowLoris	243

### 2.3. Information gain

The commonly used strategy for selecting dataset features is IG, which acts as a filter. This method seeks to mitigate interference resulting from extraneous features by employing a straightforward attribute ranking technique, followed by identifying features that possess the highest information content inside a specific class. Feature entropy evaluation is employed to discern exceptional characteristics [18]. The IG algorithm was chosen as a feature selection technique because IG is effective in reducing feature dimensions by selecting the most relevant and informative attributes, thereby significantly increasing the accuracy of the classification model. In addition, IG is easy to calculate and is often used in various data processing applications, helping to eliminate irrelevant features that can interfere with the performance of the mode [19]. IG determines feature ranking, which considers weight values and minimum weights. In this study, the original set of 47 features was reduced to a final set of 10 by filtering. Moreover, the chosen characteristics will be utilized to detect DDoS attacks in cloud computing. The IG calculation can be expressed mathematically using (1).

$$IG(X_i, Y) = H(Y) - H(Y | X_i) \quad (1)$$

Where  $H(Y)$  is the entropy of the target  $Y$ , and  $H(Y | X_i)$  is the conditional entropy of  $Y$  given  $X_i$ .

### 2.4. Principal component analysis

In the field of machine learning, PCA is a dimensionality reduction technique that is employed to simplify a dataset while keeping critical information. PCA functions by discerning patterns within the data and categorizing associated variables into uncorrelated principal components [20], [21]. This method can be utilized for feature extraction (generating new features) or feature selection (choosing a subset of the original features), contingent upon the analytical requirements [22]. This study will reduce the dimensionality with PCA from 47 to 10 features. Machine learning will then use these features in the classification training process. PCA has five stages of data standardization: covariance metrics, eigenvectors and eigenvalues, principal components, and data transformation. Mathematically, data standardization is in (2), and data transformation is in (3).

$$Z = \frac{X - \mu}{\sigma} \quad (2)$$

$$X_{new} = ZV_{selected} \quad (3)$$

Where  $Z$  is the data and  $V_{selected}$  is the matrix of selected eigenvectors.

### 2.5. Two-step feature selection (Hybrid IG-PCA)

This study primarily aims to present a two-step methodology for feature selection. This methodology is referred to as two-step feature selection, employing both hybrid feature selection along with feature extraction techniques. This approach aims to optimize the detection system on the cloud computing network in the feature selection process. The first step is to select features using IG and divide them into ten features. Then, the results of IG serve as input for the PCA method into eight features. The combination of the two methods can be formulated as follows:

- i) Feature selection: select a subset of  $X_{IG}$  features based on the IG value  $IG(X_i, Y)$  as shown in (4).

$$X_{IG} = \{X_i | IG(X_i, Y) > threshold\} \quad (4)$$

- ii) Dimensionality reduction: apply PCA to the selected features to obtain data with lower dimensions XPCA as shown in (5).

$$XPCA = PCA(XIG) \quad (5)$$

This process produces final XPCA data with relevant features and lower dimensionality than IG.

## 2.6. Classification algorithm

The suggested detection system for identifying DDoS assaults on cloud computing employs a classification algorithm. Three classification techniques will be employed: RF, DT, and NB. This proposed approach aims to find the best method to detect DDoS attacks on cloud computing networks. In addition, it seeks an optimization method for the detection system with a feature selection process. This study proposes three feature selection schemes: IG, PCA, and two-step feature selection. The following are details about the detection methods used in this study.

- i) DT is a supervised machine learning algorithm that employs a tree structure for classification or regression. It begins at the root node, which signifies the primary features, and recursively partitions the data at decision nodes (internal nodes) according to specific criteria until it arrives at the leaf nodes, which yield the predicted outcomes. This technique identifies the most effective feature partition to provide a homogeneous data subset, hence promoting transparent and comprehensible decision-making [23], [24]. Its advantages are ease of interpretation and visualization of results.
- ii) RF is an ensemble machine learning algorithm that generates a final prediction based on average (for regression) or majority voting (for classification) after aggregating predictions from numerous randomly built DTs using bootstrap sampling techniques and random feature selection at each node [25]. RF enhances accuracy, mitigates overfitting, and yields a more stable and dependable model by amalgamating numerous uncorrelated trees, in contrast to a solitary DT [26].
- iii) NB is a probabilistic-based supervised learning algorithm that uses Bayes' theorem, assuming that features are independent (independence assumption) [27]. This algorithm calculates the posterior probability of each class by analyzing the distribution of input data. It subsequently identifies the class with the greatest probability as the definitive outcome [28]. NB is known for its simplicity, computational efficiency, and good performance on large datasets and text classification.

## 2.6. Experiment testing

The testing in this study is conducted across three scenarios. First, testing is performed using a split dataset to build the model. Second, testing is conducted with 5-fold cross-validation. Finally, testing with 10-fold cross-validation is used to construct the detection system model.

## 2.7. Analysis tools

This research was conducted within a cloud computing environment, making use of platforms such as Kaggle to obtain and manage datasets, perform feature selection, and run the overall detection system in a scalable manner. In addition to cloud resources, various computational tools were integrated to ensure efficiency and reproducibility throughout the experiments. The scikit-learn library played a central role, serving as the primary framework for implementing both feature selection techniques and detection algorithms during the computation process. By combining cloud-based resources and machine learning libraries, the study was able to streamline data processing, enhance detection accuracy, and support flexible experimentation in different scenarios.

## 2.8. Evaluation

The performance of the detection system was assessed in this study utilizing a number of criteria, including accuracy, precision, true positive rates (TPR), FPR, and receiver operating characteristic (ROC)s. The equation is used to formulate this measurement are shown in (6) to (9).

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$TPR = \frac{TP}{TP+FN} \quad (8)$$

$$FPR = \frac{FP}{FP+TP} \quad (9)$$

### 3. RESULTS AND DISCUSSION

The research findings and a comprehensive discussion are presented in this section. The results are illustrated with figures and tables. The discussion is divided into several sub-sections to facilitate comprehension, including the results of the IG, PCA, and two-step feature selection.

#### 3.1. Result of information gain

The IG method is employed in this section to identify and filter relevant attributes for the detection process. The results of feature selection are outlined in this section. Table 2 presents the outcomes of feature selection analyses conducted using the IG approach. The computation produces the weight value for each feature. A ranking is conducted for each feature weight to determine those with the highest significance, which will subsequently be employed as detection features. This study identified the top 10 features according to their weight values, which would be employed in the detection process.

Table 2. The performance of IG

No	Number of features	Name of features	Weight	No	Number of features	Name of features	Weight
1	39	IAT	2.110198	24	9	rst_flag_number	0.348559
2	1	Header_Length	1.210035	25	10	psh_flag_number	0.345456
3	38	Tot size	1.129592	26	37	Std	0.325441
4	41	Magnitue	1.118270	27	42	Radius	0.322530
5	34	Min	1.116962	28	43	Covariance	0.317749
6	36	AVG	1.110629	29	44	Variance	0.268860
7	33	Tot sum	1.101664	30	3	Duration	0.169837
8	35	Max	1.061710	31	40	Number	0.150890
9	2	Protocol Type	1.026949	32	45	Weight	0.148681
10	15	syn_count	0.658792	33	20	HTTPS	0.066578
11	26	TCP	0.658677	34	19	HTTP	0.032225
12	4	Rate	0.577531	35	31	IPv	0.015481
13	5	Srate	0.577285	36	32	LLC	0.012103
14	0	flow_duration	0.572258	37	24	SSH	0.002293
15	18	rst_count	0.547138	38	13	cwr_flag_number	0.001538
16	30	ICMP	0.533837	39	29	ARP	0.000983
17	8	syn_flag_number	0.524269	40	6	Drate	0.000846
18	17	urg_count	0.501401	41	22	Telnet	0.000243
19	27	UDP	0.424965	42	28	DHCP	0.000155
20	16	fin_count	0.393286	43	12	ece_flag_number	0.000000
21	14	ack_count	0.386183	44	23	SMTP	0.000000
22	11	ack_flag_number	0.379289	45	21	DNS	0.000000
23	7	fin_flag_number	0.358502	46	25	IRC	0.000000

#### 3.2. Result of principal component analysis

PCA is a feature reduction technique that transforms existing features into new feature representations. Table 3 is an instance of use PCA to decrease the total amount of features from 47 to 10. The new feature transforms the PCA findings into a ranking distinct from the original data values. The value is typically normalized to a range of -1 to 1.

Table 3. The performance of PCA

Value PCA	Row1	Row2	Row3	Row4	Row5
PCA0	8675.561	-152738	-91943.8	120524.6	76580.89
PCA1	214703.8	-45951.4	-46090.3	-46399.4	-46347.1
PCA2	185975.4	15616.08	15599.35	15322.46	15397.55
PCA3	-12003.9	-11986.8	-12196.8	-12200.9	-12117.9
PAC4	-8565.2	251.7574	384.5478	264.1639	262.6733
PCA5	78.792	10.08704	3.635533	12.25255	12.24731
PCA6	141.9468	-1.47545	-5.98376	0.179532	-0.14366
PCA7	25.50638	0.135234	1.243965	0.297605	1.526457
PCA8	64.84385	4.058592	-4.96672	4.492461	4.835289
PCA9	86.19377	-3.56293	1.823169	-2.8915	-2.83645

#### 3.3. Result of two-step feature selection

This section discusses the results of the proposed feature selection method. They are named two-step feature selection. The process is to do two feature selection processes. First, the feature is done

using IG. Then, the feature selection results with IG are used as input for the PCA method. The process flow of the 47 features is selected into ten features, then extracted using PCA into eight features, as in Table 4.

Table 4. The performance of tow-step feature selection

Value PCA	Row1	Row2	Row3	Row4	Row5
PCA0	8131.762	-152680	-91885.5	120583.3	76639.5
PCA1	-38307.9	-38261.8	-38324.9	-38262	-38295.8
PCA2	9447.373	-306.421	-439.521	-320.543	-318.748
PCA3	-419.262	-2.66787	7.760183	-4.66937	-4.73017
PCA4	-148.211	3.106256	7.680614	1.901873	2.096895
PCA5	91.9793	1.496252	-3.67944	2.103574	2.043199
PCA6	-9.01527	-1.68345	3.461927	-1.30394	-1.33527
PCA7	5.819091	0.77438	5.699311	0.73147	0.714889

### 3.4. Result of attack detection

The next stage is to conduct the detection process to obtain a reliable detection system model for detecting DDoS attacks in cloud computing. The model testing used in this study is DT, RF, and NB. Each model is presented with three feature selection methods, namely IG, PCA, and two-step feature selection. Then, the evaluation parameters used are accuracy, precision, TPR, FPR and ROC. Additionally, model validation was conducted using data splitting, 5-fold cross-validation, and 10-fold cross-validation.

Figure 2 is one of the results of testing the RF model using two-step feature selection (IG-PCA). Where this is the value of the confusion matrix or the number of successes of each type of DDoS attack successfully detected, there are still some detection errors, but they can be tolerated. Unsuccessful testing was obtained using the NB method, where many detection errors occurred. Figure 3 is an example of one of the tests with the NB method. The results show detection errors that are almost all detected as DDoD-TCP\_Flood attacks.

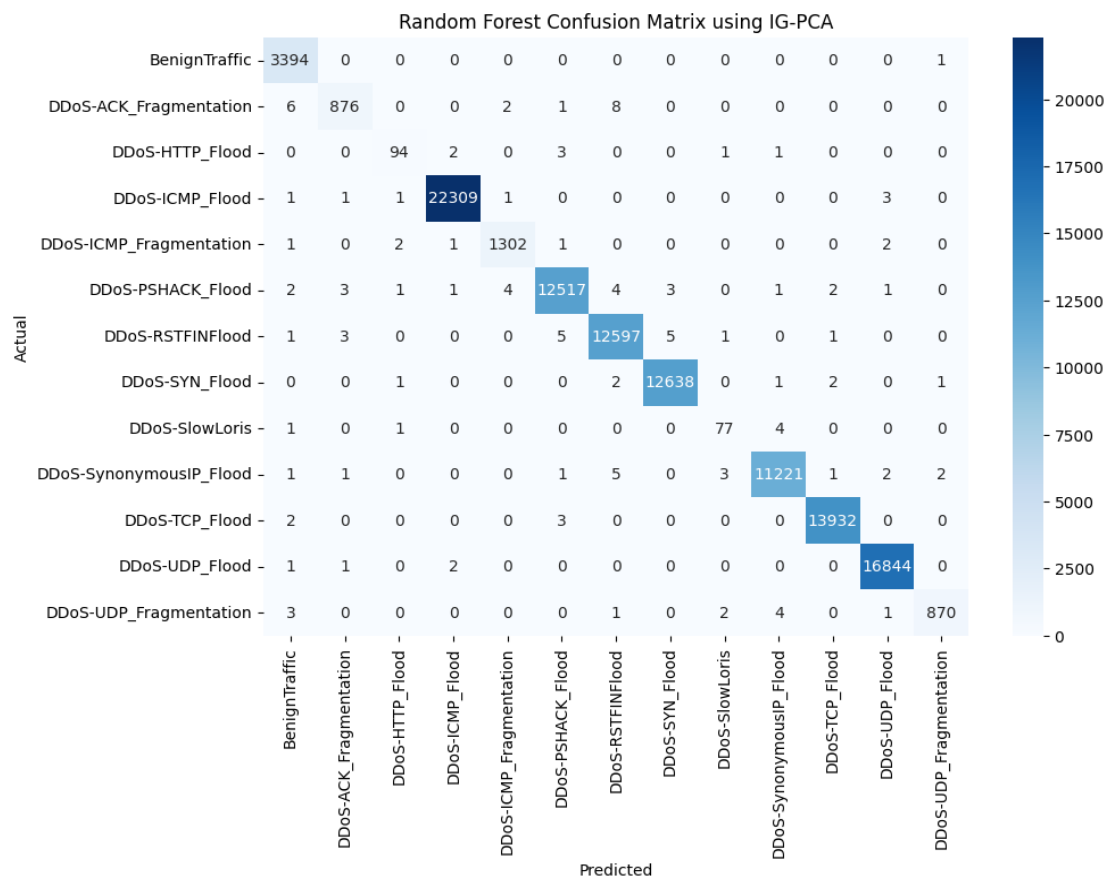


Figure 2. Result of RF confusion matrix

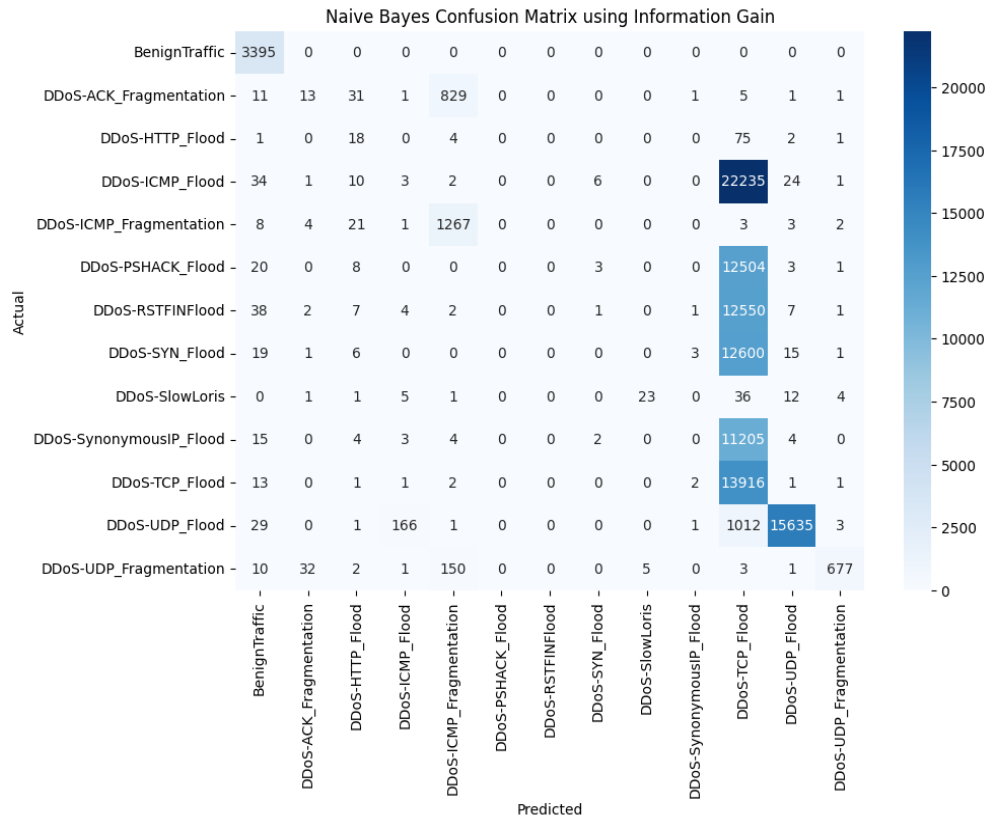


Figure 3. Result of NB confusion matrix

The next step is to calculate the performance of each test model using various feature selection methods. Table 5 shows the results of DDoS attack detection on cloud computing networks using the RF algorithm and the two-step feature selection method. The metrics used for evaluation include accuracy, precision, TPR, FPR, and ROC. These results show very good performance for various types of DDoS attacks, with a success rate of 99%. However, for the DDoS-SlowLoris attack, the results are less satisfactory with an accuracy of around 85%.

Table 5. Result of detection DDoS attack using IG-PCA (two-step feature selection)

Types of attack	RF-IG-PCA					
	Accuracy	Precision	Recall	TPR	FPR	ROC
BenignTraffic	0.994	1.000	0.999	1.000	0	1.000
DDoS-ACK_Fragmentation	0.990	0.996	1.000	0.982	0	0.991
DDoS-HTTP_Flood	0.878	1.000	1.000	0.95	0	0.975
DDoS-ICMP_Flood	0.999	1.000	1.000	1.000	0	1.000
DDoS-ICMP_Fragmentation	0.989	0.997	0.997	0.996	0	0.998
DDoS-PSHACK_Flood	0.997	0.999	0.999	0.998	0	0.999
DDoS-RSTFINFlood	0.998	0.999	0.999	0.999	0	0.999
DDoS-SYN_Flood	0.998	0.999	0.999	0.999	0	1.000
DDoS-SlowLoris	0.855	1.000	1.000	0.94	0	0.97
DDoS-SynonymousIP_Flood	0.997	1.000	0.999	0.999	0	0.999
DDoS-TCP_Flood	0.999	0.999	0.999	1.000	0	1.000
DDoS-UDP_Flood	0.999	0.999	1.000	1.000	0	1.000
DDoS-UDP_Fragmentation	0.983	1.000	1.000	0.993	0	0.997

Then, the average results of each model test and feature selection that has been done are calculated in Table 6. This table compares each model with each feature selection method. The best results were obtained from the RF method for all feature selection methods in split data. The two best results were also obtained from the DT method with all feature selection methods. However, the opposite result occurred in the NB method, which failed to detect DDoS attacks on cloud computing networks, which only reached



30 to 40%. The results of the TPR and FPR parameters for DT and RF indicate that these models are capable of effectively detecting DDoS attacks.

Table 6. Result of validation model detection DDoS attack in data split

Model	Feature selection	TPR	FPR	Precision	ROC	Accuracy
DT	IG	0.999	0	0.999	0.999	0.999
	PCA	0.983	0	0.987	0.991	0.998
	Two-step feature selection	0.999	0	0.999	0.999	0.999
RF	IG	0.993	0	0.990	0.996	0.999
	PCA	0.98	0	0.992	0.989	0.998
	Two-step feature selection	0.988	0	0.986	0.994	0.998
NB	IG	0.394	0.059	0.375	0.667	0.321
	PCA	0.403	0.059	0.401	0.671	0.322
	Two-step feature selection	0.398	0.059	0.349	0.669	0.320

In Tables 7 and 8, the results of the DDoS detection system model validation testing using 5-fold cross-validation and 10-fold cross-validation are presented. The results show that each measurement parameter exhibits satisfactory values. Notably, a significant improvement is observed in the NB model, where the accuracy reached only 30% in the data split model. In the cross-validation tests, the accuracy increased to 89%. This indicates that the NB model requires the use of cross-validation for its training process.

Table 7. Result of validation model detection DDoS attack in 5-cross validation

Model	Feature selection	TPR	FPR	Precision	ROC	Accuracy
DT	IG	0.999	0	0.999	0.999	0.999
	PCA	0.990	0	0.991	0.995	0.999
	Two-step feature selection	0.998	0	0.997	0.999	0.999
RF	IG	0.993	0	0.989	0.996	0.999
	PCA	0.984	0	0.985	0.992	0.999
	Two-step feature selection	0.988	0	0.987	0.994	0.999
NB	IG	0.396	0.06	0.383	0.668	0.893
	PCA	0.406	0.059	0.411	0.673	0.894
	Two-step feature selection	0.398	0.06	0.366	0.669	0.893

Table 8. Result of validation model detection DDoS attack in 10-cross validation

Model	Feature selection	TPR	FPR	Precision	ROC	Accuracy
DT	IG	0.999	0	0.999	0.999	0.999
	PCA	0.991	0	0.991	0.995	0.999
	Two-step feature selection	0.998	0	0.998	0.999	0.999
RF	IG	0.993	0	0.991	0.996	0.999
	PCA	0.988	0	0.988	0.993	0.999
	Two-step feature selection	0.99	0	0.988	0.994	0.999
NB	IG	0.396	0.06	0.39	0.668	0.893
	PCA	0.406	0.059	0.41	0.673	0.894
	Two-step feature selection	0.399	0.06	0.367	0.669	0.893

### 3.5. Discussions

This section delineates the results of the conducted experiments. This research aims to develop a detection system model that can identify the various forms of DDoS attacks that occur on cloud computing networks. The proposed detection methods are DT, RF, and NB; there are feature selection methods, namely IG, PCA, and two-step feature selection. The test results are superior to other methods, namely RF, DT, and NB. This happens because RF has superior characteristics regarding its number of calculations. RF is a collection of several DT methods so that it can recognize DDoS attacks better than other methods.

The DT method can recognize DDoS attacks well, with a percentage reaching 99%. The opposite occurs in the NB method, which can be concluded to fail to recognize DDoS attacks on cloud computing networks. This result is likely due to the statistical characteristics of the NB method, which is less suitable for DDoS attack detection models. Figure 4 shows a comparison of three evaluation parameters, namely: i) TPR, ii) FPR, iii) precision, iv) ROC, and v) accuracy. Of all these parameters, only NB shows less than satisfactory performance.

In Figure 4(a), the comparison of the proportion of true positives detected by the model is presented. The results indicate that the RF model outperforms the NB model. Then, in Figure 4(b), the FPR parameter shows the proportion of negatives incorrectly detected as positives by the model. The

results demonstrate satisfactory performance across all models. Precision is depicted in Figure 4(c), which quantifies the ratio of genuine positive predictions to the total number of positive predictions. The precision findings provide enhanced model evaluation for the DT and NB models. Next, in Figure 4(d), the ROC values are displayed, illustrating the relationship between TPR and FPR, where a ROC value above 0.5 is considered successful. In this study, the ROC performance is very satisfactory, except for the NB model. Figure 4(e) illustrates the accuracy, indicating highly satisfactory performance, with accuracy enhancement observed throughout model validation using cross-validation. Consequently, it can be inferred that the implementation of cross-validation can improve the efficacy of the detection system when integrated with feature selection. From the test results, it can be concluded that the feature selection method successfully improves the performance of the DDoS attack variation detection system using machine learning. In addition, the two-step feature selection method is also successful in improving and detecting DDoS attacks on cloud computing networks. Then, the detection model with RF and DT can recognize several types of DDoS attacks better than the NB detection model.

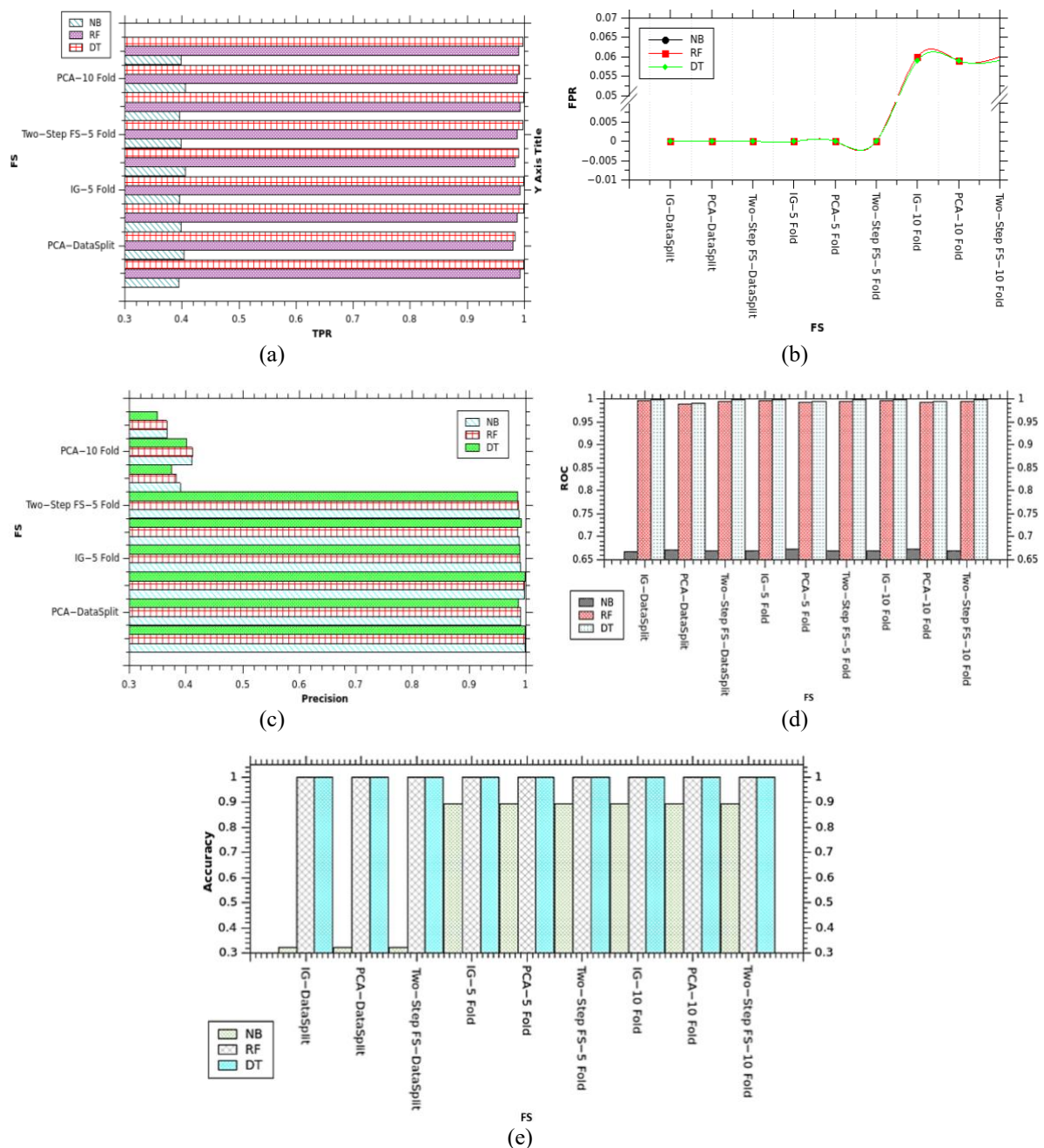


Figure 4. Performance of detection DDoS attack (a) TPR, (b) FPR, (c) precision, (d) ROC, and (e) accuracy

### 3.5. Comparison of DDoS attack detection

The results of this study indicate an increase in the DDoS attack detection system using feature selection in the DT, RF, and NB detection models. This is in addition to an increase from previous studies. Table 9 shows the results of a comparison of the proposed method with several methods from previous studies. The results show an increase in accuracy of around 2% to 5% from the method proposed in this study.

Table 9. Result of comparison detection DDoS attack with previous study

Author	Method	Accuracy
Bagyalakshmi and Samundeeswari [14]	LVQ-DT, PCA-DT	0.984, 0.986
SaiSindhuTheja and Shyam [15]	OCSA-RNN	0.941
Agarwal <i>et al.</i> [16]	WOA-DNN	0.953
Proposed method	IG-RF	0.999
Proposed method	IG-DT	0.999
Proposed method	PCA-RF	0.998
Proposed method	PCA-DT	0.998
Proposed method	IG-PCA-RF	0.999
Proposed method	IG-PCA-DT	0.998

## 4. CONCLUSION

The objective of this research is to identify variations of DDoS attacks on networks utilized for cloud computing. The proposed methodology aims to enhance the performance of detection models by utilizing DT, RF, and NB algorithms, together with feature selection techniques. The suggested feature selection approach will consist of IG, PCA, and a two-step feature selection. The results were also validated using data splitting and cross-validation. The findings from the tests showed that the IG-RF and IG-DT methods managed to achieve an accuracy level of 99%, as did the two-step method which also achieved an accuracy of 99%. The results suggest that the implementation of IG and two-step feature selection effectively enhances the performance of the DDoS attack detection system on cloud computing networks. Nevertheless, the outcomes obtained using the NB method are suboptimal, with a mere 30% success rate. However, in the tests with 5-fold and 10-fold cross-validation, the NB model showed an improvement in accuracy, reaching 89%. The study is limited by the additional time required for the two-step feature selection approach, as it involves an extra stage compared to single-step selection. While this method enhances detection accuracy, it may increase computational demands, making it less suitable for real-time systems where speed is critical. Our study demonstrates that two-step feature selection is more resilient than single feature selection. Future studies may investigate the effect of feature selection in deep learning and explore feasible methods for producing detection methods for DDoS in cloud computing.

## ACKNOWLEDGEMENTS

We would like to thank LPPM Universitas Dinamika Bangsa for facilitating this research activity.

## FUNDING INFORMATION

We would like to thank the Indonesian Ministry of Research, Technology, and Higher Education for funding and supporting this research through the Regular Fundamental Research Grant (contract number 057/LL10/PG.AK/2024).

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Kurniabudi	✓	✓		✓	✓	✓		✓	✓	✓		✓		
Eko Arip Winanto		✓	✓	✓	✓	✓		✓	✓	✓	✓			
Sharipuddin	✓			✓			✓			✓		✓	✓	✓

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**xperiment

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

*Two-steps feature selection for detection variant distributed denial of services attack ... (Kurniabudi)*

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

The datasets used in this study are publicly available and have been referred in the manuscript.




## REFERENCES

- [1] A. Sunyaev, "Cloud computing," in *Internet Computing, Principles of Distributed Systems and Emerging Internet-Based Technologies*. Cham, Switzerland: Springer, 2020, pp. 195–236.
- [2] A. Sharma and U. K. Singh, "Investigation of cloud computing security issues & challenges," *Proceedings of the 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021)*, vol. 4, pp. 445–453, 2021, doi: 10.2991/ahis.k.210913.055.
- [3] K. Srinivasan, A. Mubarakali, A. S. Alqahtani, and A. D. Kumar, "A survey on the impact of DDoS attacks in cloud computing: prevention, detection and mitigation techniques," in *Intelligent Communication Technologies and Virtual Mobile Networks*. Berlin, Germany: Springer, 2019, pp. 252–270, doi: 10.1007/978-3-030-28364-3\_24.
- [4] A. R. Kunduru, "The perils and defenses of enterprise cloud computing: a comprehensive review," *Central Asian Journal of Mathematical Theory and Computer Sciences*, vol. 4, no. 9, pp. 29–41, 2023.
- [5] M. A. Omer, A. A. Yazdeen, H. S. Malallah, and L. M. Abdulrahman, "A survey on cloud security: concepts, types, limitations, and challenges," *Journal of Applied Science and Technology Trends*, vol. 3, no. 2, pp. 101–111, 2022, doi: 10.38094/jastt301137.
- [6] Z. He, T. Zhang, and R. B. Lee, "Machine learning based DDoS attack detection from source side in cloud," *4th IEEE International Conference on Cyber Security and Cloud Computing, CSCloud 2017 and 3rd IEEE International Conference of Scalable and Smart Cloud, SSC 2017*, 2017, pp. 114–120, doi: 10.1109/CSCloud.2017.58.
- [7] A. Sahi, D. Lai, Y. Li, and M. Diykh, "An efficient DDoS TCP flood attack detection and prevention system in a cloud environment," *IEEE Access*, vol. 5, pp. 6036–6048, 2017, doi: 10.1109/ACCESS.2017.2688460.
- [8] Z. Chen, F. Jiang, Y. Cheng, X. Gu, W. Liu, and J. Peng, "XGBoost classifier for DDoS attack detection and analysis in SDN-based cloud," *2018 IEEE International Conference on Big Data and Smart Computing, BigComp 2018*, pp. 251–256, 2018, doi: 10.1109/BigComp.2018.00044.
- [9] A. R. Wani, Q. P. Rana, and N. Pandey, "Machine learning solutions for analysis and detection of DDoS attacks in cloud computing environment," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 3, pp. 2205–2209, 2020, doi: 10.35940/ijeat.b3402.029320.
- [10] G. S. Kushwah and S. T. Ali, "Distributed denial of service attacks detection in cloud computing using extreme learning machine," *International Journal of Communication Networks and Distributed Systems*, vol. 23, no. 3, pp. 328–351, 2019, doi: 10.1504/ijcnds.2019.10022365.
- [11] G. S. Kushwah and V. Ranga, "Voting extreme learning machine based distributed denial of service attack detection in cloud computing," *Journal of Information Security and Applications*, vol. 53, 2020, doi: 10.1016/j.jisa.2020.102532.
- [12] S. R. Mugunthan, "Soft computing based autonomous low rate DDOS attack detection and security for cloud computing," *Journal of Soft Computing Paradigm*, vol. 2019, no. 2, pp. 80–90, 2019, doi: 10.36548/jscp.2019.2.003.
- [13] G. S. Kushwah and V. Ranga, "Optimized extreme learning machine for detecting DDoS attacks in cloud computing," *Computers and Security*, vol. 105, 2021, doi: 10.1016/j.cose.2021.102260.
- [14] C. Bagyalakshmi and E. S. Samundeeswari, "DDoS attack classification on cloud environment using machine learning techniques with different feature selection methods," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 5, pp. 7301–7308, 2020, doi: 10.30534/ijatcse/2020/60952020.
- [15] R. SaiSindhuTheja and G. K. Shyam, "An efficient metaheuristic algorithm based feature selection and recurrent neural network for DoS attack detection in cloud computing environment," *Applied Soft Computing*, vol. 100, 2021, doi: 10.1016/j.asoc.2020.106997.
- [16] A. Agarwal, M. Khari, and R. Singh, "Detection of DDOS attack using deep learning model in cloud storage application," *Wireless Personal Communications*, vol. 127, no. 1, pp. 419–439, 2022, doi: 10.1007/s11277-021-08271-z.
- [17] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "CICIoT2023: a real-time dataset and benchmark for large-scale attacks in IoT environment," *Sensors*, vol. 23, no. 13, 2023, doi: 10.3390/s23135941.
- [18] R. A. Ghazy, E. S. M. El-Rabaie, M. I. Dessouky, N. A. El-Fishawy, and F. E. A. El-Samie, "Feature selection ranking and subset-based techniques with different classifiers for intrusion detection," *Wireless Personal Communications*, vol. 111, no. 1, pp. 375–393, 2020, doi: 10.1007/s11277-019-06864-3.
- [19] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Multi-stage optimized machine learning framework for network intrusion detection," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1803–1816, 2021, doi: 10.1109/TNSM.2020.3014929.
- [20] G. Serpen and E. Aghaei, "Host-based misuse intrusion detection using PCA feature extraction and kNN classification algorithms," *Intelligent Data Analysis*, vol. 22, no. 5, pp. 1101–1114, 2018, doi: 10.3233/IDA-173493.
- [21] Y. C. Huang and C. C. Hou, "Using feature engineering and principal component analysis for monitoring spindle speed change based on Kullback–Leibler divergence with a Gaussian mixture model," *Sensors*, vol. 23, no. 13, 2023, doi: 10.3390/s23136174.
- [22] J. Hu, X. Yang, and L. Yang, "A novel diagnosis scheme against collusive false data injection attack," *Sensors*, vol. 23, no. 13, 2023, doi: 10.3390/s23135943.
- [23] M. H. L. Louk and B. A. Tama, "Dual-IDS: a bagging-based gradient boosting decision tree model for network anomaly intrusion detection system," *Expert Systems with Applications*, vol. 213, 2023, doi: 10.1016/j.eswa.2022.119030.
- [24] M. A. Ferrag, L. Maglaras, A. Ahmim, M. Derdour, and H. Janicke, "RDTIDS: rules and decision tree-based intrusion detection system for internet-of-things networks," *Future Internet*, vol. 12, no. 3, 2020, doi: 10.3390/fi12030044.
- [25] S. Waskle, L. Parashar, and U. Singh, "Intrusion detection system using PCA with random forest approach," *Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020*, pp. 803–808, 2020, doi: 10.1109/ICESC48915.2020.9155656.
- [26] C. Liu, Z. Gu, and J. Wang, "A hybrid intrusion detection system based on scalable k-means+random forest and deep learning," *IEEE Access*, vol. 9, pp. 75729–75740, 2021, doi: 10.1109/ACCESS.2021.3082147.




- [27] J. Gu and S. Lu, "An effective intrusion detection approach using SVM with naïve Bayes feature embedding," *Computers and Security*, vol. 103, 2021, doi: 10.1016/j.cose.2020.102158.
- [28] T. Wisanwanichthan and M. Thammawichai, "A double-layered hybrid approach for network intrusion detection system using combined naive Bayes and SVM," *IEEE Access*, vol. 9, pp. 138432–138450, 2021, doi: 10.1109/ACCESS.2021.3118573.

## BIOGRAPHIES OF AUTHORS






**Kurniabudi**    received a Doctor of Engineering from Universitas Sriwijaya. He is currently a senior lecturer at the Faculty of Computer Science, Universitas Dinamika Bangsa, Indonesia. His research interests include technology adoption, information technology, information security, and network security. He can be contacted at email: kbudiz@yahoo.com.



**Eko Arip Winanto**    received the B.Sc. degree in Computer Science from the University of Sriwijaya, Indonesia, and the M.Phil. degree in Computer Science from Universiti Teknologi Malaysia, Malaysia. He is currently a doctoral student at Universiti Teknologi Malaysia and a lecturer at the Faculty of Computer Science, Universitas Dinamika Bangsa, Indonesia. His research interests include IoT, machine learning, blockchain, and network security. He can be contacted at email: winanto@unama.ac.id.



**Sharipuddin**    received a Doctor of Engineering from Universitas Sriwijaya. He is currently a senior lecturer at the Faculty of Computer Science, Universitas Dinamika Bangsa, Indonesia. His research interests include information technology and information security. He can be contacted at email: sharipuddin@unama.ac.id.